Loughborough University

This item was submitted to Loughborough University as a PhD thesis by the author and is made available in the Institutional Repository (https://dspace.lboro.ac.uk/) under the following Creative Commons Licence conditions.

creative commons

COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/

BLL ID No: D50744/84

LOUGHBOROUGH
UNIVERSITY OF TECHNOLOGY
LIBRARY

AUTHOR/FILING TITLE

PATRICK, P J

ACCESSION/COPY NO.

004613/02

VOL. NO. | CLASS MARK

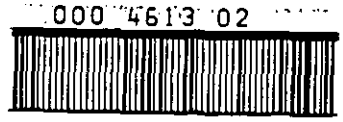
LOAN COPY

-5. JUL. 1985

-4. JUL. 1986

-4. JUL. 1986

000 4613 02



ENHANCEMENT OF BANDLIMITED SPEECH SIGNALS

Volume II

Loughborough University
of Technology
Date May 84
Class
Acc. No. 004613/02

Chapter 5

Speech Bandwidth Enhancement

5.1 Introduction

Having discussed some of the aspects involved in classifying certain speech sounds, we now turn our attention to the objective of speech quality improvement; later we will find that some of this work will draw upon the results just described. In our case the term "speech quality improvement" means 'to endeavour to make speech that has been conveyed by a telephone-bandwidth channel (of 300 to 3400Hz) such that it perceptually appears to have been conveyed by a wider-than-telephone bandlimited channel' i.e. to deceive the listener into assuming that a wider-than-telephone bandwidth was used to convey the signal.

When speech signals are transmitted via the telephone network, they are generally degraded by additive noise, multiplicative noise, impulse and switching noise, cross talk, dispersion. There is also the frequency attenuation distortion introduced by the handset carbon microphone and band-limitations due to pre- and post-line filtering. The filtering process causes the telephone speech signal quality to appear both "tinny" due to its lack of base frequencies and "muffled" due to its lack of high frequencies.

We consider that when speech is bandlimited to the telephone bandwidth, the perception of the high frequencies of voiced speech is almost unaffected while the unvoiced speech can bear the most

degradation. The effect of band limitation may be acceptable for handset speech communication, as in the case of a person-to-person telephone call, however, the degradation is much more apparent in the case of hands-free speech communication for instance in audio-teleconferencing.[refs 11-14] This effect is also apparent during the phone-in programmes on radio and television broadcasting. For audio-teleconferencing, an à priori knowledge of the high quality wideband speech is available. Bandwidth enhancement in this case can be achieved by processing the speech signal in both the transmitter and receiver. In the transmitter, speech is spectrally compressed using a bandwidth compression algorithm which confines the transmitted signal into the range of 300 to 3400 Hz while the receiver employs a corresponding bandwidth expansion process. This will be discussed in the following chapter. For the "phone-in" programme signals, only the bandlimited version of the speech signal is available which suggests a 'receiver-only' type of processing arrangement in order to improve the quality of the received speech signal.

This chapter discusses some of the experimentation performed in order to achieve bandwidth enhancement by receiver-only processing. The receiver attempts to extend the spectrum of the speech signal beyond its 3400Hz cut-off frequency by employing a high frequency regenerator whose output is added to the received speech signal from the telephone bandlimited channel. The input to the high frequency regenerator (HFR) is the telephone bandlimited signal shown in figure 5.1.

All of the processing methods discussed in this chapter will stem from this basic model of figure 5.1.

The first attempt at enhancing the high frequencies involves a simple process applied to the whole of the speech signal, namely the high frequency spectral duplication method. We then examine the subjective effect of synthesising an additional formant to the out-of-band spectrum for voiced speech only. Our investigations will then be directed towards generating higher frequencies for unvoiced sounds and leaving voiced-sounds unprocessed. To generate the higher frequencies a simple noise source is used. Later on this noise source will be substituted by non-linear transfer functions and spectral shaping filters. We will then look at schemes used for replacing the analogue 300-3400Hz filter with a digital encoder and consider how the subjective performance of conventional coders may be improved by simply adding a post-processor to the decoded signal.

Finally, we investigate the possibility of bandwidth enhancement of low frequencies i.e. the re-synthesis of the baseband frequencies in the 0-300Hz range. This, hopefully, should again afford a quality improvement of speech over that obtained from the 300-3400Hz channel alone.

5.2 High Frequency Regeneration (HFR) by Spectral Duplication

High frequency regeneration by spectral translation or spectral folding has been described by Makhoul and Berouti.(ref 70) Their scheme employed digital modulation and filtering to duplicate a

bandlimited spectrum such that the resultant signal possessed a spectral width which was twice or three times greater than the original input signal. The process was used in conjunction with a voice-excited vocoder system, i.e. used to provide the flattened wide-band spectrum for the excitation signal to a vocoder. In our technique, we consider spectral duplication in the frequency domain to provide a spectrally enhanced signal without any additional vocal tract modelling.

Briefly, the method operates as follows; the 300Hz to 3400Hz band limited signal is first discrete Fourier Transformed by the FFT to produce a frequency spectrum for the signal. The in-band spectral coefficients are copied to out-of-band spectral locations as shown in figure 5.2. The new 6.8kHz spectrum is transformed back into the time domain by the inverse discrete Fourier Transform to give the processed output signal. In order to provide some continuity at the boundary between the original spectrum and the duplicated spectrum (i.e. at the 3400Hz frequency ordinate), the value of the signal energy over 0 to 1000Hz was compared to the signal energy between 2400Hz and 3400Hz. The duplicated spectral values were attenuated by the ratio of these two values. This technique will now be explained in more detail.

5.2.1 The Spectral Duplication Process

Consider a block of N speech samples represented by the sequence $\{x_i\}$. The discrete Fourier Transform (DFT) of $\{x_i\}$ is the sequence $\{X_k\}$

whose components are

$$X_k = \frac{1}{N} \sum_{i=0}^{N-1} x_i \exp\left\{-j \frac{2\pi ki}{N}\right\} \qquad (5.1)$$

$$k = 0,1,\ldots,N-1$$

The magnitude of X_k is

$$|X_k| = \left\{Re^2(X_k) + Im^2(X_k)\right\}^{1/2}$$

$$k = 0,1,\ldots,N/2 \qquad (5.2)$$

where Re(.) and Im(.) are the real and imaginary values of (.). Because of the symmetrical nature of $\{X_k\}$, there are (N/2)+ 1 distinct components in each real and imaginary sequence of X_k. The phase of X_k is

$$\theta_k = \frac{180}{\pi} \text{Tan}^{-1} \left\{\frac{Im(X_k)}{Re(X_k)}\right\} \qquad (5.3)$$

Equations (5.2) and (5.3) are the polar representation of X_k. The values of X_k from k= N/2+1 to N are discarded.

Now the frequency resolution (frequency separation) of each DFT component, Δf is equal to F_s/N where F_s is the sampling frequency (in kHz). Therefore, the nearest value of k which corresponds to a

frequency 'f' kHz is:

$$k_f = \text{NINT}(f/\Delta f) \tag{5.4}$$

where NINT(.) corresponds to the nearest integer value of (.).

The average value of the spectral energy within the 0 to 1kHz band is given by:

$$E_L = \frac{1}{k_{1.0}} \sum_{k=1}^{k_{1.0}} |X_k|^2 \tag{5.5}$$

Similarly, the average value of the spectral energy contained within the band of 2.4 to 3.4kHz is given by:

$$E_H = \frac{1}{(k_{3.4} - k_{2.4})} \sum_{k=k_{2.4}}^{k_{3.4}} |X_k|^2 \tag{5.6}$$

The ratio of the upper and lower spectral energies for each block of N samples is then found viz:

$$R = E_H/E_L \tag{5.7}$$

The parameter R now forms the attenuation factor for the block by which the in-band spectral samples are multiplied by to form the duplicated spectral samples, i.e.

$$|X_{k+k_{3.4}}| = R \cdot |X_k| \tag{5.8}$$

$$\text{for } k = k_{0.3}, (k_{0.3}+1) \ldots, k_{3.4}$$

The phase values are duplicated without scaling i.e.

$$\theta_{k+k_{3.4}} = \theta_k \qquad k = k_{0.3},(k_{0.3}+1),\ldots,k_{3.4} \tag{5.9}$$

The new spectrum now occupies a bandwidth of 6.8kHz whereas the spectrum of the input signal occupied a bandwidth of 3.4kHz.

Having completed the spectral duplication, the $\{|X_k|\}$ and $\{\theta_k\}$ sequences are converted into real and imaginary sequences whose components are given by

$$\text{Re}(X_k) = |X_k| \text{ Cos } \theta_k$$

$$\text{and} \quad \text{Im}(X_k) = |X_k| \text{ Sin } \theta_k \qquad \text{for } k = 0,1,\ldots,N/2 \tag{5.10}$$

To extend the number of components in the real and imaginary sequences to N we invoke the complex conjugate property of spectral components (which is due to the time data being in the real sampling field)

$$X_{N-k} = X_k^* \qquad \text{for } k = 0,1,\ldots,N/2 \tag{5.11}$$

where the raised '*' above the symbol implies its complex conjugate, and

$$X_k = \text{Re}(X_k) + j\ \text{Im}(X_k)$$

The inverse DFT is now performed to give the time sequence $\{\tilde{x}_i\}$, where

$$\tilde{x}_i = \sum_{k=0}^{N-1} X_k \exp\left(j \frac{2\pi k i}{N}\right) \qquad (5.12)$$

$$i = 0,1,\ldots,N-1$$

$\{\tilde{x}_i\}$ is the time sequence whose frequency components occupy twice the frequency band of that of the sequence $\{x_i\}$. The corresponding spectral shape of $\{\tilde{x}_i\}$ would now be typified by that as sketched in figure 5.2(b).

Results

The above operations were performed by computer simulation, the value of N was taken as 1024 samples and F_s was set to 16kHz. The resultant time sequence was fed via the DAC in figure 4.1 and recorded for informal subjective evaluation.

The perceptual quality of the high frequency duplication process was found to be extremely unpleasant; this was especially true for the unvoiced sounds. The effect upon the voiced sounds was the impression of a signal dependent whistle, rather like the effect of aliasing of a speech signal. The processed signal quality did not appear to have any positive subjective attributes at all, although the speech intelligibility and speaker recognition was maintained to a reasonably high level. The poor perceptual quality was probably due to the fact that the duplicated spectrum had no similarities to

the original spectrum and the inclusion of the duplicated spectrum proved to be more like adding out-of-band distortion products rather than useful signal. The degradation appeared to be worsened for the consonants rather than vowels possibly due to the duplicated spectrum being amplified and not attenuated as for the vowels. The existence of the background tones during the vowel sounds was possibly due to the fact that with spectral duplication, the harmonic structure is interrupted at the 3.4kHz frequency ordinate. It may be likely that these tones can be eliminated by adjusting the bandwidth of the in-band spectrum in figure 5.2(a) to be a multiple of the pitch fundamental frequency on a short-term basis.(ref 70) Unfortunately, such a scheme would require a large amount of computation and would offset the simplicity afforded by this spectral duplication method. Due to this ensuing complexity, the high frequency regeneration scheme by spectral duplication was discontinued.

5.3 HFR by the Synthesis of an Out-of-Band Formant

The bulk of the energy of the voiced spectrum lies within the 3400Hz lowpass range and in general there can be some three formants within that frequency bandwidth. If it is required to synthesise higher frequencies of this lowpass range beyond the 3400Hz cut-off frequency for the voiced speech then it may be worthwhile to generate a high frequency signal such that the spectrum of the synthesised plus the 3400Hz signals shows a continuation of the formant structure.

The idea is to synthesise a signal from a resonant circuit of fixed frequency which endeavours to generate a fourth or at least an additional formant to be added to voiced speech signals only. It may not be considered advantageous, as far as perception is concerned, to vary the frequency of this formant in accordance with a defined rule of some description. The salient feature of this synthesised signal is thought to be how the gain of the formant should be relative to the available bandlimited version of the speech signal.(ref 76)

5.3.1 Experimental Procedure

Initially, the additional formant was generated at a fixed frequency of 4kHz and no gain control was used. The general system employed to spectrally enhance voiced speech spectrally by this method is shown in figure 5.3.

The resonant network is implemented using a 4th order all-pole recursive filter. This network is fed by the excitation signal derived from the input bandlimited signal x(t) which has been lowpass filtered to 3400Hz. For this arrangement, it is only required to add the synthesised high frequency signal during the presence of voiced speech. The operation of the switch S_1, is governed by the control of the voiced decision. The voiced decision is based upon the first shift autocorrelation function method which has been just described in full in section 4.7. The parameters of the voiced decision scheme were set to suit the bandlimited signal

$x(t)$ such that the switch S_1 closes (i.e. $d(t)=1.0$) when voiced speech is deemed to be present in $x(t)$. The switch S_1 opens such that $d(t)=0.0$ otherwise.

Before discussing the operation of the resonant circuit, we need to explain how the high frequency excitation is generated in order to drive the resonant circuit.

5.3.2 HFR Excitation

In section 3.5.1 we saw that Schroder and David(ref 47) proposed a vocoder for transmitting wideband speech (10kHz) over a 3.5kHz channel. The transformation from the shaped and bandlimited spectrum of the base band speech signal to a flat wideband spectrum required for excitation between 2 and 10kHz was achieved by a piece-wise linear network with an input-output characteristic of straight line segments. The transfer function of this network (shown in figure 5.4) resembled the leter 'W'. The transfer function has the form

$$y = 1 - 4\,|x'| \qquad \text{for} \quad |x'| \leq 0.5$$

$$y = 4\,|x'| - 3 \qquad \text{for} \quad 0.5 \leq |x'| \leq 1.0$$

and $$y = 0 \qquad \text{elsewhere} \qquad (5.13)$$

This function spreads the spectrum more for large input amplitude values than for small ones. In order to overcome the effect of different amplitudes causing different spectral spreading,

instantaneous logarithmic compression of the speech signal preceding the 'W' transfer function was used.(ref 47) The logarithmic compressor appears as shown in figure 5.5.

To ensure that the logarithmic compressor transfer function does not saturate, it is useful to find the maximum value of the magnitude of the input signal x(t). If the processing is not performed in real-time then value of each sample of the input data can be tested to find $|x|_{max}$. Once this has been found, the expression for the logarithmic transfer function can be ascertained, viz

$$x'(t) = \alpha.\mathrm{sgn}(x) \log_e \left[c\ |x| + 1\right]$$

where $$\alpha = 1/(\log_e \left[c\ |x|_{max} + 1\right]$$

and $$\mathrm{sgn}(x) = x/|x|$$

(5.14)

'c' is the compression factor; the smaller value of 'c' the more shallow, is the rate of compression; the larger values of 'c' correspond to harder compression. For our case, the value of 'c' chosen was 5 and the actual compression law is shown in figure 5.6(a). The output from the logarithmic compressor is now fed directly into the 'W' transfer function as shown in figure 5.6(b). There is no longer any possibility of overloading this 'W' function as the maximum value of the magnitude of $|x'|$ is 1.0. The combined transfer function is now shown in figure 5.7. The network is not quite the same as that described in the reference 47 as that was achieved by a diode arrangement.

Analytically, one can envisage the spectral spreading phenomenon by resorting back to the 'W' transfer function without the logarithmic compressor. The spectral width can be measured by the second moment of the power spectrum $P(\omega)$:

$$\overline{\omega^2} = \frac{\int_0^\infty \omega . P(\omega) d\omega}{\int_0^\infty P(\omega) d\omega} \qquad (5.15)$$

As a consequence of Parseval's theorem, the second moment of the power spectrum is related to the integral over the corresponding time function, $x(t)$ and its derivative as follows

$$\overline{\omega^2} = \frac{\int_{-\infty}^{\infty} \left(\frac{dx}{dt}\right)^2 dt}{\int_{-\infty}^{\infty} x^2(t)\ dt} \qquad (5.16)$$

For the particular network shown in figure 5.4, the value of (dy/dt) for the output is four times the value of (dx/dt) of the input due to the 4:1 slope of the network branches. However, for a signal whose absolute amplitude exceeds 0.5 at any time, the output signal will have a smaller amplitude than the input signal for such instants of time thus

$$\int y^2(t)\ dt < \int x^2(t)\ dt \qquad (5.17)$$

and with equation (5.16)

$$\overline{\omega^2_{out}} < \overline{\omega^2}_{in}$$

The exact amount of spectral spreading can be computed for any given input waveform. In the case of a triangular input waveform with extremal amplitudes of ±4, the spectral spreading be by a factor of 4

i.e. $$\sqrt{\overline{\omega^2_{out}}} = 4\sqrt{\overline{\omega^2_{in}}}$$

For a sinewave of amplitude 4, the spectral spreading factor can be shown to be equal to 4.4 i.e.

$$\sqrt{\overline{\omega^2_{out}}} = 4{:}4\sqrt{\overline{\omega^2_{in}}}$$

Experimentally, the combined logarithmic compressor and the 'W' transfer function was used to show how the spectrum of a voiced /I / in "sister" may be extended. Figure 5.8(a) shows the input signal and figure 5.8(b) shows the expanded output signal. It can be seen that the formant structure of the original is smoothed but the fine structure is retained even for the out-of-band spectral components.

Originally(ref 47), the non-linear distortion process, i.e. the use of the logarithmic compressor and the 'W' transfer function was applied in the analogue domain. When applying this process in the digital domain the non-linear distortion may introduce energy at frequencies higher than the Nyquist rate. It was therefore

recommended[ref 70] that the baseband be interpolated to at least double the original sampling rate before the distortion in order to avoid spectral aliasing which can cause roughness in the output speech. After the spectral expansion, the signal can be filtered to just below the original Nyquist rate before decimating the time sequence by 2 to drop the sampling frequency back down to the original value. The generation of the high frequency excitation now appears as shown in figure 5.9.

5.3.3 The Resonant Circuit

The resonant network of figure 5.10 will be responsible for providing the additional formant in order to enhance spectrally the high frequencies of bandlimited voiced speech. As previously stated, this network is implemented as an all-pole recursive filter as shown in figure 5.10(a).

The positions of the poles for this formant are indicated on the s-plane aa shown in figure 5.10(b)[ref 76]. In practice, the resonant network was modelled by two cascaded second order filters.

i.e. $$H(z) = H_1(z).\ H_2(z)$$

where $$H_1(z) = 1/(1 + a_1\ z^{-1} + a_2\ z^{-2})$$

and $$H_2(z) = 1/(1 + b_1\ z^{-1} + b_2\ z^{-2}) \qquad (5.18)$$

For a 16.0kHz sampled signal, the poles in the s-plane appear in the z-domain as shown in figure 5.11. Under the transformation

$$z = \exp(sT) = \exp(\sigma T) \angle \omega T$$

where $T = 1/F_s$ (5.19)

the positions of the poles in the z-domain are

$$\beta_1 = \exp(-100T) \angle (4000-200)T.2\pi$$

$$\beta_1^* = \exp(-100T) \angle -3800T.2\pi$$

$$\beta_2 = \exp(-100T) \angle 4200T.2\pi$$

$$\beta_2^* = \exp(-100T) \angle -4200T.2\pi$$

where $T = 1/16000$ (5.19)

The transfer function of the filter in figure 5.11 is:

$$H(z) = 1/\left[(z-\beta_1)(z-\beta_1^*)(z-\beta_2)(z-\beta_2^*)\right]$$

$$= 1/\left[(z^2-(\beta_1+\beta_1^*)z + \beta_1\beta_1^*)(z^2-(\beta_2+\beta_2^*)z + \beta_2\beta_2^*)\right]$$

$$= z^{-4}/\left[(1 + a_1 z^{-1} + a_2 z^{-2})(1 + b_1 z^{-1} + b_2 z^{-2}\right]$$

(5.20)

The numerator of H(z) corresponds to a delay of 4 sample intervals and the denominator corresponds to the product $H_1(z)\ H_2(z)$. The values of the coefficients are:

$$a_1 = -(\beta_1 + \beta_1^*) = -2\,\mathrm{Re}(\beta_1)$$

$$= -2\exp(\sigma_1 T).\mathrm{Cos}(\omega_1 T)$$

$$a_2 = \beta_1\,\beta_1^* = |\beta_1|^2 = \exp(2\sigma_1 T)$$

Similarly

$$b_1 = -2\exp(\sigma_2 T).\mathrm{Cos}(\omega_2 T)$$

and $$b_2 = \exp(2\sigma_2 T) \qquad (5.21)$$

In our case, the chosen values of the coefficients are:

$$\sigma_1 = \sigma_2 = 100S^{-1}$$

and $f_c = 4000\mathrm{Hz}$, $\Delta f = 200\mathrm{Hz}$

so $\omega_1 = 2\pi(f_c - \Delta f)$

and $\omega_2 = 2\pi(f_c + \Delta f)$

The filter was implemented as a subroutine called FORMANT: the listing is included within the appendix A4. In order to check the response of the resonant filter, the response was evaluated around the unit circle in the z-domain i.e. $H(z)\Big|_{z=e^{j\omega T}}$; the gain and phase versus frequency response of the filter is shown in figure 5.12. The subroutine RESPONSE was used for this purpose and its listing is also included in the appendices. As expected, the gain response shows a very pronounced peak at 4kHz with rapidly decaying

skirts where the frequency deviates from the centre frequency. The two poles are not resolved along the $j\omega$ axis as they are positioned too close together in the z-domain. The phase response also shows a small deviation from linearity at the 4kHz centre frequency.

The impulse response was determined by inputting a unity valued sample to the subroutine and plotting the output which is shown in figure 5.13. The plot of the impulse response contains a high frequency component and a low frequency and decaying envelope which is expected for a 2-pole damped network.

The parameter 'G' in subroutine FORMANT controls the amplification factor of the formant that is generated; it is this factor which controls the mixing proportions of the formant and the bandlimited signal so as to produce a high frequency enhanced output signal. This fixed amplifier is applied to the signal $y(t)$ in figure 5.3 which now completes the voiced high frequency regeneration system.

5.3.4 Results

The results of adding the synthesised formant to the 0.3 to 3.4kHz voiced speech had shown no apparent difference in the time waveform plots of figure 5.14(a) for the voiced /I / in sister. This is due to the fact that the relative level at which the synthesised formant added was about 50dB lower than the first formant of the same speech segment; the spectral plots of figure 5.14(b) clearly show this relative frequency distribution. In this case, the synthesised formant appears as a fifth formant. The frequency and gain at which the resonant circuit of figure 5.10 is driven, were held constant.

The synthesised formant is only introduced into the 300 to 3400Hz bandlimited signal when voiced speech is deemed to be present by the voiced speech detector (section 4.7). The value of G was set to 0.70×10^{-7}. This value of G was selected experimentally to generate a continuation of the spectral trend of the formants; e.g. for the spectrum in figure 5.15, this shows that the height of the peak of the synthesised formant falls roughly on an imaginary line connecting all of the peaks of the in-band formants.

The perceptual results of the processed signal were unfortunately a little disappointing; there was no evidence of any quality improvement as far as providing an impression of a wider bandwidth than 300 to 3400Hz; rather, the processed signal appeared degraded by the disturbance of a high frequency "whistle" during voiced speech. This sounded as though the bandlimited signal was partially aliased in frequency. These investigations were in fact only intended to be of an exploratory nature and the limitation of time prevented any further study into this area. The obvious recommendation at this stage is to determine how the gain of the resonant circuit must be adjusted to conform to the statistics of the in-band signal. This may mean that an adaptive algorithm is to be sought.

The networks used for the high frequency enhancement system for voiced speech will be found to be of use in other areas discussed further in this chapter.

5.4 HFR for Unvoiced Speech

In view of the results mentioned in the previous section, it was felt that it might be more advantageous to synthesise high frequency components for unvoiced speech beyond the 3400Hz cut-off frequency and leave the voiced speech signals unprocessed. Certainly, unvoiced speech signals do possess quite significant spectral energy components well above the 3400Hz cut-off frequency which are well worth investigating for spectral enhancement. One may hypothesise that voiced speech is far more sensitive to signal modifications than unvoiced speech as far as perceptual quality and naturalness of sound is concerned.

The following experiment, was used to test whether it would be at all worthwhile to attempt to enhance spectrally unvoiced speech in absence of any voiced speech processing. The arrangement in figure 5.16 was set up on the computer program to investigate this idea.

The voiced speech is fed via a 300Hz to 3400Hz band-pass filter and the unvoiced speech is processed with a 300Hz to 7600Hz filter plus a level control amplifier/attenuator module. The voiced and unvoiced segments were selected by visual waveform inspection (discussed in Chapter 4). The summation of the filtered voiced and unvoiced signals provided the output signal in this arrangement. The computer subroutines used to model the bandpass filters (FIR in this case) are included in the appendix A1.2. The number of taps used for each filter was 256 so that the delay in each branch of the circuit was synchronised.

The results of this experiment were obtained in the form of graph plots and informal subjective listening tests. When the amplifier was set to 0.5 gain, the output signal for /sI/ in "sister" appears as shown in figure 5.17(b). When this is compared with the input signal of the same specimen of speech as shown in figure 5.17(a), there appears to be only an amplitude difference between the waveforms of (a) and (b) during the sustained unvoiced /s/ signal. Only at the transition between the /I/ and /s/ does there show some attenuation due to the lowpass filtering of this part of the signal. This was interpreted as voiced speech by the 'V/UV' switch. When the signal was tested subjectively for the 10-word sequence with the amplifier gain set to unity, it was found that the unvoiced speech appeared to be slightly loud in comparison with the level of the voiced speech sounds. This impression was emphasised when the amplifer level was set to 2.0. However, when the amplifier was set to 0.5, it was then found that the subjective impression of the unvoiced speech was such that it blended more naturally with the bandlimited voiced speech than when any of the other amplifier settings were used. Furthermore, the quality of the output speech of figure 5.17(b) appeared to be almost the same as the quality of the speech bandlimited to 300Hz to 7600Hz and superior to the quality of speech bandlimited from 300Hz to 3400Hz. The result seems to indicate that the quality of voiced speech signals remains almost unaffected by the 300 to 3400Hz bandwidth limitation as most of their energy resides below the 3400Hz cut-off frequency. Unvoiced speech, on the other hand, does possess energy at frequencies above 3400Hz and it is therefore considered preferable to apply some form of high frequency enhancement to the bandlimited

unvoiced speech signals rather than to improve the quality of the bandlimited voiced speech. The following sections in this chapter attempt to adhere to that principle.

5.4.1 Development of the High Frequency Regenerator (HFR) for Unvoiced Speech Sounds

The general scheme for producing high frequency regeneration for unvoiced speech is shown in figure 5.18. The transmitter operates on the 300 to 7600Hz bandlimited speech to detect the presence of unvoiced speech.

This information is multiplexed with the baseband 300 to 3400Hz signal in the form of an injected pilot tone into a narrow frequency slot located at some frequency within the 300 to 3400Hz bandlimited signal. At the receiver, the voiced/unvoiced decision is demultiplexed from the baseband signal and is used to control the high frequency regenerator network. The baseband signal is also used to excite the high frequency network, the output of which is added to the received baseband signal to provide an artificially generated wideband signal at the output of the receiver.

At present, it is necessary to extract this voiced/unvoiced information from the 300 to 7600Hz wideband input speech and multiplex it with the bandlimited speech signal prior to transmission. Thus the current method is not a 'receiver-only' processor since it requires the availability of the wideband input speech but it possesses minimal processing requirements at the transmitter terminal.

Before discussing the high frequency regeneration network, the multiplexing of the side information will be explained in a little more detail. It was necessary to find a suitable frequency slot into which the pilot tone could be inserted without causing significant distortion in the received speech. A notch filter was used whose characteristic is shown in figure 5.19(a). The filter was implemented by a 1024 point non-recursive network with a finite impulse response illustrated in figure 5.19(b).

The centre frequency of the spectral slot was selected to be 2600Hz as it was considered to be above the average range of the second formant. The maximum rate of changing of the state value of the pilot tone, in order to indicate the voiced/unvoiced transitions in the speech signal, is defined by the shortest unvoiced sound. The duration of the shortest unvoiced sounds, which were the /t/ stop consonants, were found to be about 40 to 50mS from the 10-word sequence. The reciprocal of the shortest unvoiced sound duration determines the maximum significant frequency of the composite modulated pilot tone (assuming amplitude shift keying is used); in our case, the maximum frequency is equal to 25Hz. A frequency slot sufficient to accommodate the unvoiced duration information was implemented using a 50dB frequency notch of 60Hz width centred at 2600Hz.

The notch filter was fed by a speech signal of 300 to 3400Hz bandwidth; the output of the filter appeared perceptually unchanged with respect to input signal when tested by informal listening experiences. The filter was also tested by waveforms and spectra

shown in figure 5.19(e) to (f). This then suggests an acceptable space in the 300 to 3400Hz channel for conveying the voiced/unvoiced information to the receiver.

Returning back to the development of the high frequency regeneration network, we now will consider several methods for the spectral extension of unvoiced speech signals.

5.4.1.1 Bandpass Filtered Noise

It is often said that the high frequencies of unvoiced speech may be synthesised by introducing a noise source into the bandlimited speech signal. This method was indeed tried using a very simple arrangement by which the regenerator network in figure 5.18 was composed of a wideband noise source followed by a 3400 to 7600Hz bandpass filter. The noise was controlled by commands received from the transmitter such that it was only switched on when unvoiced sounds were present in the received signal.

The method was later modified such that the envelope waveform of the unvoiced speech was transmitted to the receiver. At the receiver, the envelope was multiplied by the bandpass noise signal. The transmission of the waveform envelope required about 300 to 500Hz additional channel bandwidth; the system now appears as shown in figure 5.20.

The results of the scheme of adding the bandpass filtered noise to the 300Hz to 3400Hz baseband unvoiced speech showed a detrimental effect both in quality and by a reduction in intelligibility,

especially with the 'unvoiced-voiced-unvoiced' utterances. This result was not improved at all by the transmission and multiplication of the noise by the waveform envelope. Our results for the HFR of unvoiced speech, at this stage, were compared with the notion developed by the BBC research dept(ref 71). Their system for regenerating high frequencies for unvoiced sounds was, in fact, a purely receiver-only arrangement as the operation of the process (shown earlier in figure 3.58) did not require multiplexed information to differentiate between voiced speech and unvoiced speech for the receiver. The system is again shown for the reader's convenience in figure 5.21.

It may be seen that the method does attempt to determine implicitly the mix of voiced and unvoiced speech from the bandlimited input signal by the amount of output from filter F_1, which controls the amplitude of the wideband noise source. The components of the variable noise source are selected by filter F_2 then added to the 300 to 3400Hz input speech. When the system was investigated, we found that the output speech signal lost little or no intelligibility when compared with the input bandlimited signal and perceptually some high frequencies did appear to be present, but the quality of the output voiced speech was poor and gave the impression that a low bit rate coding system was used in the channel.

Going back now to the scheme shown in figure 5.1 the next question that we considered was what can we place into the 'HFR-box' in order to attempt to generate more successfully high frequencies for the unvoiced sounds? This meant returning to non-linear transfer

functions to generate inter-modulation products and hence provide a broader spectrum output signal. As well as using the 'W' transfer function from section 5.3 to generate a wideband excitation signal for the spectral enhancement of unvoiced speech sounds, a different spectral broadening technique was used where the instantaneous baseband speech amplitude may be raised to the third power(ref 77), i.e. a cubic transfer function. These are considered in the following sections.

5.4.1.2 Use of the 'W' Transfer Function for Unvoiced Signals

In this method, the baseband 300 to 3400Hz received unvoiced signal was used as the input to an instantaneous logarithmic compressor followed by a 'W' transfer function. The output of the 'W' element was fed into a 3400-7600Hz bandpass filter as shown in figure 5.22(a).

5.4.1.3 Cubic Transfer Function(ref 77)

In this scheme, the 'W' transfer function was replaced by the cubic transfer function as shown in figure 5.22(b). The cubic transfer function was only applied for unvoiced signals.

5.4.1.4 Hybrid Regenerator

Using the high frequency regenerators of sections 5.4.1.2 and 5.4.1.3 a hybrid arrangement was produced by including (a) the 'W' transfer function preceded by the logarithmic compressor and

followed by the 3400-7600Hz bandpass filter; (b) the cubic transfer function plus a 3400-7600Hz bandpass filter and (c) the logarithmic compressor plus the 'W' transfer function and a 300 to 3400Hz bandpass filter. Regenerator (a) is incorporated when an /s/ sound is indicated from the transmitter. Regenerator (b) is used when the transmitter indicates an /f/ or /θ/ sound and finally regenerator (c) is applied when a /ʃ/ is selected at the transmitter. The reason for using model (c) for the /ʃ/ sounds is that we have noted in section 4.2 that the time waveform of a /ʃ/ sound appears very similar in shape to that of an /s/ sound with the exception of the zero-crossing rate being considerably lower than that of the /s/ sound which may be reproduced by bandpass filtering the output of the 'W' transfer function to a lower range than that for the high frequency synthesis of an /s/ sound.

5.4.1.5 'W' Transfer Function applied only for /s/ Sounds

In this case, high frequency regeneration was applied for /s/ utterances only using the network shown in figure 5.22(a). All other speech sounds (voiced and unvoiced) were left unprocessed and therefore, the receiver output yielded the baseband 300Hz to 3400Hz version for these signals. The high frequency regenerated signals applied for the /s/ utterances were slightly attenuated below the normal unvoiced to voiced amplitude ratio of the original 300 to 7600Hz signal.

5.4.1.6 Results

In all of the experiments described in sections 5.4.1.2-5, the relevant unvoiced speech signals were selected by visual waveform inspection. In addition to this, the zero-crossing rate detector (section 4.5) was used in sections 5.4.1.2, 4 and 5 for detection of the /s/ signals.

The results of the informal subjective listening tests of the methods utilising these non-linear transfer functions began to show some improvement over the baseband signal. This was perhaps in contrast with the noise-source techniques used previously for the high frequency regeneration. Method 5.4.1.2 had shown improvement, particularly with the /s/ sounds, with little reduction in intelligibility, although the /ʃ/ sounded more like an /s/ utterance. Method 5.4.1.3 showed a similar improvement over the baseband signal but more so with an /f/ sound. The hybrid system explained in section 5.4.1.4 had shown improvement for /s/, /f/ and /θ/ utterances but the /ʃ/ sound was still degraded. However, the technique 5.4.1.5 had shown the best overall improvement so far.

The waveform plots for the schemes discussed are illustrated in figure 5.23. The original waveform of the /s/ sound has a consistently high zero-crossing rate with a non-uniform envelope structure. Figure 5.23(b) shows that the /s/ signal is greatly attenuated and distorted by the bandpass filtering from 300Hz to 3400Hz. The noise source of figure 5.43(c) may perhaps show some improvement over the bandlimited /s/ of figure 5.23(b) but the

structure of the waveform bears very little resemblance to that of the original waveform, both in terms of zero-crossing rate and the amplitude envelope. If the time envelope of the original is multiplied by the noise source as shown in figure 5.23(d), then obviously the shape of the waveform now appears more like the original but the informal perceptual results appear to have shown that the transmission of the time envelope had very little effect on improving the quality of the processed signal. It does seem to suggest that the 'secret' of improving the quality of the processed signal lies in preserving or at least reproducing the original zero-crossing rate of the original signal. This was attempted by employing the 'W' transfer function of section 5.4.1.2 which gives rise to the waveform of figure 5.23(e). The waveform of the output from cubic transfer function used in figure 5.22(b) is shown in figure 5.23(f). For the hybrid system of section 5.4.1.4, it appears that the 'W' transfer function yields a slightly more regularly structured time waveform than does the cubic transfer function. The former thus seems more suited to enhance the /s/ sound whilst the latter appears more successful when artificially producing the high frequencies of the /f/ sound.

It was found that the technique of employing the 'W' transfer function for /s/ utterances only (yielding the waveform of figure 5.23(e)) and leaving all of the other speech sounds unprocessed produced the best results. We therefore considered that in order to afford additional sound improvement of the processed signal, the amplitude envelope of the original may be used in conjunction with the process of 5.4.1.5. It was considered that it may also be

advantageous to aim at shaping the regenerated spectrum for the /s/ signal such that shape of the output spectrum of the processed signal would be more like the shape of the original 300Hz to 7600Hz /s/ spectrum. We will now concentrate only upon the /s/ fricative sounds in the speech signal which we consider perceptually to contribute most significantly to high frequency components of wideband speech.

The reason for only processing wideband /s/ sounds only from all other speech utterances are as follows:

1. The bulk of the spectral content of voiced speech, particularly vowels, resides within the 300 to 3400Hz band;

2. Although stop consonants and voiced fricatives possess spectral energy outside the 300 to 3400Hz range, a significant amount of spectral energy still lies within this band. Further, it was earlier found that only a marginal improvement may be obtained if the bandpass filtered version of the stops and voiced fricatives be replaced by their wideband (300-7600Hz) counterparts (from the system in figure 5.16). This assumes that perfect high frequency regeneration is possible for these signals.

3. The relative frequency of the /s/ phoneme in English speech prose is 4.55%[ref 3] which is ranked high in the list of consonants (4th out of 24). The wideband version of the /s/ utterance also shows a large improvement in perception over the bandlimited case concerning both intelligibility and quality.

4. Unlike most other sounds, the bulk of the spectral energy of the /s/ phoneme lies outside the 300 to 3400Hz range with very little energy within that range. Thus the process of detection and separation of the /s/ signal from the 300 to 7600Hz is quite a straightforward task (section 4.3 and 4.5) and may be achieved reliably and without too much complexity.

5.4.2 Transmission of the Waveform Envelope of the /s/ Signal

The waveform of figure 5.23(e) shows the structure of the waveform for the synthesised /s/ signal. If the start and finish of the /s/ utterance were included, this would indicate an abrupt start and finish to each /s/ sound. The perceptual quality of the processed /s/ sound may be improved by incorporating a gradual rise at the leading edge and a gradual fall at the trailing edge of each processed /s/ segment. The leading and trailing edges were modified by exponentially rising and decaying envelope functions respectively. The envelope functions were arbitrarily chosen and are shown in figure 5.24.

For the leading edge of each processed /s/ segment, function 5(a) was used where:

$$y(t) = \exp(at) - 1 \qquad (5.22)$$

and $y(T_A)$ is set to 1.0

$$\therefore \quad a = \frac{\ell n\, 2}{T_A} \quad S^{-1} \qquad (5.23)$$

For the trailing edge of each processed /s/ segment, function 5.24(b) was used, viz:

$$y(t) = \exp(-bt) \tag{5.24}$$

and $$y(T_B) = \exp(-bT_B) = 0.1$$

$$\therefore \quad b = \frac{\ell n\ 10}{T_B}\ S^{-1} \tag{5.25}$$

For our experiments, the duration of each exponential smoothing envelope was taken as 16mS, such that

$$T_A = T_B = 16mS \tag{5.27}$$

The resultant processed waveform for the /s/ signal (figure 5.25) now appears to have a smoother rise and fall in amplitude at the start and finish of each /s/ segment. It was therefore expected that the perceived quality of the output signal for the /s/ sounds may be better blended with the unprocessed, bandlimited voiced speech. Unfortunately, this was found not to be the case, and appeared that although the /s/ signal still seemed to be conveyed by a wider bandwidth channel than 300 to 3400Hz, it still sounded somewhat unnatural when heard together with bandlimited voiced speech in the same speech passage.

Further experiments were then conducted to modulate the processed /s/ waveform of figure 5.23(e) by the amplitude envelope of the original 300 to 7600Hz /s/ signal at the input to the transmitter. The experiment was set up as shown in figure 5.26.

The time constant of the envelope detector was set to 10mS giving the resultant envelope signal corresponding the /s/ sound in "sister" as shown in figure 5.27. At the receiver, this is used to control the amplitude of the synthesised /s/ signal of figure 5.23(e).

Upon listening to the output signal, the method appeared to produce a better quality sound for the /s/ utterances than did the case using the exponentially rising or decaying windows. This is possibly due to the fact that the wideband /s/ sound is a continuous utterance, the high frequency regenerated /s/ signal appears to require more shaping than just the leading and trailing edges of the time waveform.

5.4.3 Shaping the Spectrum of the Processed /s/ Signal

As previously mentioned, a further step in the direction of improving the quality of the processed /s/ signal was to attempt to shape its spectrum such that it appeared more similar to the /s/ spectrum of the input signal.

The spectrum for a wideband /s/ signal at the input to the transmitter is shown in figure 4.3(b). The spectrum of a bandlimited /s/ signal is shown in figure 5.28(a). If the bandlimited /s/ signal is enhanced with the 'W' transfer function element of figure 5.22(a) with no spectral shaping, then the frequency plot will appear as illustrated in figure 5.28(b). It can be seen that whilst the original spectrum has a characteristic peak

in the spectral envelope centred at about 4.8kHz, there is no such contouring with the spectrum of figure 5.28(b). The flat spectrum output from the 'W' transfer function may be contoured using a filter with a spectral peak in its response. One such type of filter is precisely the same as that used to synthesise an additional formant to enhance the the high frequencies of voiced speech as described in section 5.3. The system used for generating a spectral peak in the processed /s/ signal is shown in figure 5.29 for the receiver terminal only. The transmitter terminal is the same as that in figure 5.26(a).

As illustrated in figure 5.26, the amplitude envelope of the /s/ signal and the decision of its presence can be transmitted along with the baseband signal, so in addition to the spectral shaping currently being applied, the amplitude of the synthesised /s/ signal is also modified by the amplitude of the original wideband /s/ signal.

The output of amplifier A_1 yields a signal with a flat spectral distribution between 3400Hz and 7600Hz whilst the output of the amplifier A_2 yields a signal with a prominent peak in its spectral magnitude. The adjustments of A_1 and A_2 control the mix of those spectral distributions.

The filter F_3 is a high Q resonant circuit of the same design as the formant filter used previously but with the settings re-adjusted from those in equation (5.21) to these as listed:

$$\sigma_1 = \sigma_2 = 1{,}100\ S^{-1} \qquad \text{(damping factors of poles)}$$

$$f_c = 4.8\ kHz \qquad \text{(centre frequency)}$$

$$\Delta f = 400\ Hz \qquad \text{(pole separation)}$$

$$A_1 = A_2 = 40.0 \qquad \text{(Amplification factors)}$$

(5.26)

These values were fixed throughout the whole of the 10-word test sequence. The resultant frequency response of the spectral magnitude and phase for the resonant filter F_3 of figure 5.29, is shown in figure 5.30; again, this was found by evaluating equation (5.20) around the unit circle in the z-plane. This was performed using subroutine RESPONSE as listed in appendix A.4.

As expected, the response of the resonant filter has a peak at the centre frequency of 4.8kHz with decaying skirts. However, one side of the response curve is not quite monotonically decreasing possibly due to spectral leakage effects from the negative frequency spectrum (c.f. section 6.4.1).

The spectra of an /s/ signal at the output of filters F_2 and F_3 is shown in figure 5.31(a) and (b) respectively. The spectrum of figure 5.31(a) appears as a bandlimited noise response whilst the spectrum of figure 5.31(b) obviously has the required peak in its

spectral amplitude envelope. The two signals are mixed together by amplifiers A_1 and A_2 with the settings as in equation (5.26) after which the signal is modulated by the envelope of the time signal from the transmitter. The spectrum of the output signal, shown in figure 5.31(c), may be compared with that of the original signal at the transmitter, which is shown in figure 5.31(d). It can be seen that the spectral contouring provides the required shaping for the spectral envelope although the fine structure, perhaps being less important for unvoiced fricatives, is not preserved.

The resultant time waveforms for the original and bandlimited /s/ signals in "sister" are shown in figure 5.32(a) and (b) respectively; the processed signal is shown in sub-figure (c). Due to the transmission of the amplitude envelope plus that fact that the prominence of the synthesised peak in the spectral magnitude is manifested as a consistently high zero-crossing rate, the processed signal seems to bear more resemblance to the original signal than does the 300 to 3400Hz bandlimited signal. The graphical results seem to be supported by the informal subjective listening tests which yielded the most encouraging results of all the methods described in this chapter so far. The provision of the spectral peak for the processed /s/ signals appears to remove the 'rasping' sound which were generated when the /s/ sound was processed without using the resonant filter.

Referring to the listening tests using our 10-word sequence, words such as "sister" and "S K Harvey" appear quite satisfactory but the /s/ utterance in the word "fist" still sounds slightly unnatural in

its blending with the voiced /I/ and the stop consonant /t/. It still appears, therefore, that more effort may be required to mix properly the regenerated high frequencies with the 300 to 3400Hz bandlimited signal.

Since this latest idea of figure 5.29 requires processing at the transmitter as well as at the receiver, it may be advantageous to examine the possibilities of replacing the analogue medium with a digital channel. We will need to examine some coding algorithms prior to transmission and preceding the reception of the signal. This is indeed the topic of our next section.

5.5 Digital Techniques in conjunction with Bandwidth Enhancement Post-Processing

In this section we are concerned with using currently developed codecs to facilitate transmission of the baseband signal together with the necessary multiplexed information for the system of figure 5.20. As it is the 300 to 3400Hz baseband signal that requires digitisation, it is not considered necessary to design a new coder to be tailored specifically for the high frequency regeneration system. The coders that are to be tested are precisely those described in Chapter 3(refs 18-27) which have the required parameter optimisations performed before being included in the channel of the bandwidth enhancement system.

The coders that were considered for transmitting the 300 to 3400Hz baseband signal and the necessary side information are APCM, DPCM, CFDM and CVSD. The investigation of the performance of the coders in this situation was undertaken by Mr S N Koh of the Department of Electrical Engineering Loughborough University.(ref 78)

5.5.1 Optimisation of APCM

The performance of APCM depends upon the input signal power level and the set of the step size multipliers used. The optimum values of the step size multipliers for different number of bits per sample were determined by Jayant and used throughout the investigations. To determine the optimum input signal power, each input sample was

divided by a fixed constant and the signal to quantization noise ratio was calculated by

$$SNR = \frac{\sum_{i=1}^{N_s} x_i^2}{\sum_{i=1}^{N_s} e_i^2} \qquad (5.27)$$

where N_s is the total number of input samples and $\{e_i\}$ was obtained by filtering the quantization noise (input minus quantizer output) by a 300 to 3400Hz bandpass filter. The noise obtained in this way is known as the in-band noise. The words "sister" and "father" plus the silence pause between the words were used to evaluate the SNR performance of the codecs. The optimum input signal power was found to be -25dB for the 300 to 3400Hz bandwidth speech signal. The performance curves for the APCM coder operating at transmission rates of 16, 24 and 32 kb/S against input signal power are shown in figure 5.33.

5.5.2 Optimisation of Differential Pulse Code Modulation

In the case of the ADPCM coder, the predictor in the feedback loop (figure 3.3) was implemented as an integrator with a leakage factor of 0.9. The optimum values of the step-size multipliers that were determined by Jayant(ref 20) were used. To optimise the value of the input signal power, the input sequence $\{x_i\}$ was again divided by different constant values and the signal to in-band quantisation noise ratio for each constant was calculated using the expression

(5.27); the results are plotted in figure 5.34. The value of the input signal power which was positioned at the mid dynamic range of the input power for the transmission bit rates investigated was again -25dB.

5.5.3 Adaptive Delta Modulation (ADM)

5.5.3.1 Constant Factor Delta Modulation (CFDM)

As previously discussed in Chapter 3, constant factor delta modulation is an instantaneously companded delta modulation with a one-bit memory Jayant quantizer. The step size adaptation strategy is given by

$$\Delta_{i+1} = \Delta_i \, M \qquad (5.28)$$

where Δ_i is the current step size, and

$M = M1$ if the present quantization output is the same as the previous quantizer output, or

$M = M2$ if they are difference (c.f. equation 3.8).

M1 should be greater than unity for the encoder to follow an abrupt increase in the magnitude of the input signal and M2 should be less than unity. It has been shown by Jayant(ref 20) that M1 and M2 should satisfy the following equation

$$M1.M2 \leq 1 \qquad (5.29)$$

for stability.

The value of Ml = 1.5 as determined by Jayant(ref 20) to maximise the signal to quantization noise. To optimise the input signal, the input sequence $\{x_i\}$ was again divided by a constant value and the signal to in-band quantization noise ratio was used (expression 5.27). The results are plotted in figure 5.35 where the optimum input signal power was found to be 0dB.

5.5.3.2 Continuously Variable Slope Delta Modulation (CVSD)

As discussed in Chapter 3, CVSD is functionally a syllabically compounded delta modulation; in this system, the quantizer step size varies at a much slower rate than the instantaneous variations in the speech simple sequence.

The coefficients α and β, from equations (3.18) and (3.20), were set to the values shown in table 5.1.

F_s(kHz)	α	β
16	0.94	0.99
24	0.96	0.99
32	0.97	0.99

Table 5.1
(ref 27)

The results of varying the input signal power are plotted in figure 5.36 and the optimum input signal power to the encoder was found to be 15dB.

5.5.4 Speech Enhancement System Incorporating the Digital Codecs

The digital speech transmission system with quality enhancement by high frequency regeneration pre- and post-processing is shown in figure 5.37. Prior to encoding the speech signal, the signal was first bandlimited from 300 to 3400kHz. It was then sampled at different rates and encoded by the four different digital codecs to obtain transmission rates of 16, 24 and 32 kb/S. The process of re-sampling an already sampled sequence is achieved by a combination of interpolation and decimation. The process is relatively straightforward and is discussed more fully in section 6.6.1.

Referring to figure 5.37, the envelope of the /s/ sounds was extracted from the 300 to 7600Hz version of the speech signal by passing this signal through an ideal full-wave rectifier followed by a lowpass filter of a finite impulse response and a bandwidth of 150Hz. The reason for using the FIR filter rather than a recursive filter is because the bandwidth to sampling frequency ratio is relatively small, hence the delay characteristics of the passband would be highly non-linear causing unwanted dispersion. However the signal delay of an FIR filter can be precisely known.

The output of the filter was sampled at a fixed rate of 400 samples per second and encoded by a two-bit ADPCM encoder to give a fixed transmission rate of 800 bits per second. The transmission channel was assumed to be ideal in all cases (i.e. no bit errors were introduced).

At the transmitter terminal, the output of both the encoders for the 300 to 3400Hz speech signal and the signal envelope were multiplexed together with the '/s/-decision' and transmitted. At the receiver the digital signal was demultiplexed and decoded by the respective decoder corresponding to that used at the transmitter to recover the bandlimited speech signal of 300 to 3400Hz and the 0 to 150Hz envelope signal. These recovered signals were then sent to the high frequency regeneration network outlined in section 5.4.1. The '/s/-decision' was used to determine whether or not to add the re-synthesised high frequency /s/ signal to the 300 to 3400Hz bandlimited signal. The material used for testing purposes for the digital system was the 10-word sequence as used in our previous experiments.

For comparison purposes, the digital codecs were tested without any form of high frequency regeneration applied at the receiver (i.e. only the 300 to 3400Hz bandwidth speech was processed). The generalised system is shown in figure 5.38.

5.5.4.1 Results of Using the Digital Codecs Alone

The aim was to study the use of digital codecs in conjunction with the high frequency regeneration network and to compare the quality of speech processed by the four different codecs at three different transmission rates of 16, 24 and 32 kb/S with and without high frequency regeneration.

Objective comparisons for the performance of the four codecs was based on their signal to in-band quantization noise and their dynamic ranges with each of the digital coders operating at its respective optimum point. As shown in figure 5.39, the SNR values against the three different transmission rates for each codec show that their relation is almost linear. It can be seen that the higher the transmission bit rate, the higher the resultant SNR.

In the case of APCM and ADPCM, the sampling rate is fixed at 8000 samples per second and as the transmission rate increases, a greater number of bits are available and may be used to represent each sampled value of the input signal. The quantization noise correspondingly decreases giving an increased SNR value. It is also noted that ADPCM is about 3dB higher in SNR than APCM. This is to be expected as ADPCM is a more efficient coding technique than APCM.

In the case of CFDM or CVSD, a one-bit quantizer is used. The increase in transmission rate is the result of the increased sampling rate. Signals sampled at a higher sampling rate produces an error sequence with a lower variance to be quantized and thus reduces its step size without causing overload. Decreasing the step size of the quantizer without overload always gives rise to smaller quantization noise and hence a higher SNR. This explains why the adaptive delta modulation had higher SNR values at higher bit transmission rates.

Comparing the dynamic range of the coders (i.e. the range of input power that provides 3dB below the peak SNR) it appears that the ADPCM has the widest dynamic range followed by CFDM, then APCM then CVSD. The ADPCM coder yields a dynamic range a few dB higher than CFDM but almost three times that of CVSD. The values of the dynamic ranges obtained when examining the coders are tabulated in Table 5.2.

	ADPCM		APCM	
TBR(kbps)	SNR(dB)	Dynamic range(dB)	SNR(dB)	Dynamic range(dB)
16	12.3	50	10.6	34
24	17.6	50	12.4	40
32	22.9	45	19.6	36

	CFDM		CVSD	
TBR(kbps)	SNR(dB)	Dynamic range(dB)	SNR(dB)	Dynamic range(dB)
16	11.0	47	13.7	10
24	17.3	43	18.7	17
32	22.9	40	22.4	17

Table 5.2 Comparison of the dynamic range offered by the different coders (ref 78)

Subjective listening experiences indicated that the higher bit transmission rate always resulted in better speech quality. ADPCM offered the best performance amongst the four codecs which is not surprising since ADPCM exploits the advantages of signal differentiation and multilevel coding as compared to APCM with multilevel coding only and delta modulation with signal differentiation only. The CFDM had more noise during the speech sounds but had less quantization noise during the silent intervals. This was expected since the step size multiplier had a small value of 0.6 during the silence pauses.

The APCM and CVSD yielded the worst subjective results from amongst the four coders. The APCM coder had a very high quantization noise for both speech sounds and the silent intervals. At 16kb/S, the CVSD coder suffered from noise as well as intelligibility reduction (although only informal subjective listening experiences indicated this result).* The result may be attributed to the very narrow dynamic range that the CVSD coder had as shown in figure 5.36.

5.5.4.2 Addition of High Frequency Regeneration

The high frequency regeneration network in figure 5.37 was now switched on and the perceptual experience of this indicated an apparent extension of bandwidth at the output of the receiver for all of the digital coding schemes considered. Most notable was the fact that the quality and intelligibility of the 16 kb/S-CVSD coder appeared very inferior without the use of the HFR network; when the HFR was applied for the /s/ sounds, the quality and intelligibility seemed to improve as far as informal listening tests were concerned. The coder noise associated with both the CVSD and CFDM systems appeared to become less obtrusive when the HFR process was applied although in comparison with the better coding scheme of ADPCM these were still considered fairly noisy whether or not the HFR processing was included. The noise generated by both of the delta modulators was more correlated with the speech signal than in the case of the APCM coding at the three transmission rates examined. The APCM coder generated a greater amount of granular noise and the subjective effect of this did not seem to improve even when the HFR process was applied. The continuous granular noise

disappeared only during the time when the /s/ sounds were synthesised which gave a disturbing switching effect from one signal to another.

In conclusion, it appeared that the 32 kb/S ADPCM system using HFR gave the best overall result and the 16 kb/S-CVSD system showed the most marked improvement for the output speech when HFR was included compared with the case without the HFR for /s/ sounds.

5.5.5 Future Recommendations of HFR

As we noted earlier in section 4.2, the spectrum of the /s/ utterance exhibited a peak in its envelope between about 4 and 5kHz. This prominent peak has been exploited by the resonant filter network of figure 5.29. However, the spectra associated with the /s/ signal may change in time especially perhaps during the onset and trailing positions of this utterance. This can be seen in the examples of the spectral plots for the leading portion and trailing portion of the /s/ utterance taken from the word "fist". These are shown in figures 5.40(a) and (b) respectively.

At present, no provision is made for this temporal variation as the pole-residues of the resonant filter in figure 5.29 are fixed. A suggested future investigation would be to track the amplitude of the spectral peak from the 300 to 7600Hz bandwidth input signal and transmit this to the receiver such that the height of the spectral peak of the synthesised /s/ signal may be varied in accordance with the input speech signal. Figure 5.41 shows the proposed HFR network

for future examination. The transmission of the strength of the spectral peak of the /s/ sound may be achieved simply by transmitting the energy within the 4.5 to 5.5kHz frequency band and using the envelope of that time signal to modulate the output of the resonance filter at the receiver. This should hopefully achieve the desired result and improve the blending of the HFR signal and the bandlimited signal.

5.6 Baseband Synthesis

Until now, we have devoted considerable attention to the enhancement of the high frequency components of speech which are otherwise suppressed by telephonic bandpass filtering from 300 to 3400Hz. As discussed earlier, the effect of telephone bandpass filtering is two-fold, i.e. to render the speech signal "tinny" as well as muffled. Reduction of the apparent muffling of the processed signal does in some cases increase the intelligibility of the processed signal. In particular, the distinction between an /s/ or an /f/ can be improved. On the other hand, if it is required to reduce the effects of highpass filtering then this is unlikely to render the processed signal more intelligible (ref 17), although here we are more interested in improving the quality of the processed signal.

The quality mismatch that exists between broadcast quality speech and "phoned-in" speech to the studio indicates the desire to attempt to improve the low frequency quality of the telephone bandwidth speech signal. Various schemes directed at this objective will now be described.

5.6.1 Low-Frequency Synthesis by a Local Oscillator Driven at Pitch Frequency

This method simply adds the output of a sinewave generator to the base band signal. The frequency of the generator is controlled by the pitch or fundamental frequency of the incoming bandpass signal while the amplitude of the oscillator is governed by the signal energy within the 300 to 1kHz band. The system, illustrated in figure 5.42, has the advantage of being devoid of any processing at the transmitter terminal, hence it requires no multiplexed side information.

The input signal x(t) is 'cleaned up' by a 300 to 3400Hz bandpass filter to produce x'(t) and the short-time energy of the 300Hz to 1kHz bandpass region was determined by a 15^{th} order Butterworth whose gain and phase response is shown in figure 5.43.

The subroutines used to model the filter are listed in the appendix A4. The filter was followed by a first order envelope detector with a time constant set to 10mS; the time constant was chosen so as to smooth out the pitch variations without causing too much "lag" in the response of the short-time energy detector.

The pitch frequency is determined using the cepstrum technique which was preceded by a sliding-block analyser. This is such that the relatively large number of samples required by the cepstrum to determine the average pitch frequency of a speech sample block does not prevent the use of a smaller time resolution for each instant that the pitch frequency measurement is updated. After the cepstrum

is determined, the function is truncated at lower and upper quefrency regions so as to filter out the pitch spike of the cepstrum. The position of the spike is then selected by a peak picker such that the pitch period may be ascertained.

The local oscillator is then driven at that pitch frequency and its amplitude is modulated by the short-time envelope $s(t)$. The resultant low frequency synthesised signal is then added to the 300 to 3400 Hz bandlimited signal to form the processed signal $z(t)$.

An additional module shown as the V/UV detector in figure 5.42 is used to switch on the low frequency synthesis network only when voiced speech is deemed to be present in $x'(t)$ and switch off the network otherwise. It may well be envisaged that the computation of the cepstrum itself may be used to perform the task of voiced/unvoiced decision but since the cepstral calculation is a fairly large load for computer execution time, a simpler method for the V/UV switch was sought. The scheme that was decided upon was the first shift autocorrelation coefficient method already explained in section 4.7 and the parameters selected for this method were those chosen for the 300 to 3400Hz input speech data. When the detector indicates a condition of voiceless speech at the input then $x'(t)$ is fed straight to the output as switch S_2 is opened. In this state, the cepstral computation does not take place but the short-time energy measurement and the sliding block process still continues.

5.6.1.1 The Sliding Block Strategy

The sliding block strategy used prior to the cepstrum computation is shown in figure 5.44. The "input-tape" contains the 300 to 3400Hz bandlimited data and the "output-tape" contains of low frequency synthesis processed data. The sequence of operations of the sliding block regime, during cycle "m", are as follows:

(1) The 1024 samples contained in block 'B' are shifted to the left by 64 samples such that if the samples in block 'B' are written as $\{b_i\}$; $i = 1,2,\ldots,1024$

then $$b_i = b_{(i+64)} \quad ; \quad i = 1,2,\ldots,960 \qquad (5.30)$$

(2) The top 64 samples of block B are filled by a new set of 64 samples from block 'A' which is read from the input-data tape; i.e.

$$b_{(961+j)} = x_j \quad ; \quad j = 1,2,\ldots,64 \qquad (5.31)$$

where $\{x_j\}$ represents the input sequence of block 'A'.

(3) The new samples of block 'B' are now used for the computation of the cepstrum which determines the average pitch frequency of the speech data within that block of samples. The pitch period is taken to correspond to the speech samples in Block 'C' i.e. the centre 64 samples of block 'B' such that

$$c_k = b_{(480+k)}$$

$$\text{for} \quad k = 1,2,\ldots,64 \qquad (5.32)$$

(4) In the diagram of figure 5.42 the voiced/unvoiced detector is used to disconnect the low frequency synthesis network when voiced speech is not present at the input. The V/UV decision scheme is based upon the first shift autocorrelation coefficient of the input speech and uses a speech block of 128 samples so that without altering the parameters chosen in section 4.7 the same block length was used. This constitutes block 'D', where

$$d_\ell = b_{448+\ell} \qquad \text{for} \quad \ell = 1,2,\ldots,128 \tag{5.33}$$

(5) If the samples of $\{d_\ell\}$ were tested to give a "voiced" condition then the 64 samples of block 'E' are those generated by the local oscillator driven at the pitch frequency, where:

$$e_j = s_j \, \mathrm{Sin}(2\pi F_o jT + \phi_{m-1}) \tag{5.34}$$

where F_o is equal to the reciprocal of the pitch period,

$T = 1/F_s$ (F_s = the sampling frequency)

and ϕ_{m-1} = the residual phase left from the previous set of samples in block 'E' i.e. during operation cycle (m-1).

The phase ϕ_{m-1} of the sinewave generator is retained such that the phase of the local oscillator is continuous even though the frequency is varied

$$\therefore \quad \phi_m = s_{64} \, \mathrm{Sin}(2\pi F_o . 64 . T + \phi_{m-1})$$

The amplitude modulation term, $\{s_j\}$, are those samples generated by the first order envelope detector. The principles of operation of the envelope detector are the same as those described in section 4.6 and need not be explained again here. The output of the envelope detector is further delayed by 480 samples before modulating the fundamental frequency oscillator so as to compensate for the shifting involved with the sliding block process.

(6) The samples of the modulated fundamental frequency, i.e. $\{e_j\}$ are then added to the delayed input block 'C' in figure 5.44 i.e.

$$z_j = e_j + c_j \qquad \text{for} \quad j = 1,2,\ldots,64 \tag{5.35}$$

before the $\{z_j\}$ samples are written to the output tape of processed data.

(7) When the V/UV detector indicates voiceless speech then the cepstral computation does not take place and the samples of block 'C' are transferred directly to the processed data tape i.e.

$$z_j = c_j \; ; \qquad j = 1,2,\ldots,64 \tag{5.36}$$

The sliding-block scheme still however continues to operate whether voiced speech is present or not.

At the finish of this step, the sliding-block strategy returns to step (1) until all the input speech data has been processed.

5.6.1.2 Pitch Detection by the Cepstrum Method

The method used to determine the average pitch frequency of the 1024 samples contained in block 'B' was the cepstrum method; this process will now be briefly reviewed as follows:

The samples contained in block 'B' are designated $\{b_i\}$ as in equation (5.31) and the DFT of that sequence is given by

$$B_J = \frac{1}{N_B} \sum_{i=1}^{N_B} b_i \exp\left[-\frac{2\pi i J}{N_B}\right] \qquad J = 1,2,\ldots,N_B \tag{5.37}$$

Since the time sequence is purely real, the double sided spectrum is in complex conjugate form i.e.

$$B_J = B^*_{N_B-J} \tag{5.38}$$

Therefore, only the values of B_J for $J=1,2,\ldots,N_B/2$ need be considered for further computation.

Next the Cartesian DFT co-ordinates are converted to Polar form viz

$$|B_J| = \left[Re^2(B_J) + Im^2(B_J)\right]^{1/2}$$
$$\angle B_J = \frac{180}{\pi} \mathrm{Tan}^{-1}\left[\frac{Im(B_J)}{Re(B_J)}\right] \qquad J = 1,2,\ldots,N_B/2 \tag{5.39}$$

The logarithm of the power spectrum is now taken with the phase set to zero for all frequency coefficients, i.e.

$$|B_J|_{dB} = 20 \log_{10}|B_J|$$

$$\text{and } \angle B_J = 0$$

$$J = 1,2,\ldots,N_B/2 \qquad (5.40)$$

To determine the cepstrum, the inverse DFT of the sequence $\{|B_J|_{dB}\}$ is taken after the full double sided spectrum is reformed.

$$|B_J|_{dB} = |B^*_{N_B-J}|_{dB}$$

$$\text{for } J = (N_B/2+1),(N_B/2+2),\ldots,N_B \qquad (5.41)$$

Since the sequence $\{|B_J|_{dB}\}$ is wholly real, the complex conjugate of each coefficient is equal to that coefficient itself, i.e.

$$|B_J|_{dB} = |B_{N_B-J}|_{dB}$$

$$\text{for } J = (N_B/2+1),(N_B/2+2),\ldots,N_B$$

$$\text{and so } b'_j = \sum_{J=1}^{N_B} |B_J|_{dB} \exp\left[\frac{2\pi jJ}{N_B}\right] \qquad (5.42)$$

where the sequence $\{b'_j\}$ is the double-sided cepstrum of the time sequence $\{b_j\}$.

As the original time sequence, $\{b_j\}$, and its double-sided log-power spectrum $\{|B_j|_{dB}\}$ are both wholly real, the double-sided cepstrum is thus symmetrical about the $N_B/2$ ordinate. Therefore, the sequence $\{b'_j\}$ for $j = (N_B/2+1), (N_B/2+2), \ldots, N_B$ is discarded.

The results of this process for a 64mS specimen of speech are shown in figure 5.45(b). There are two characteristic portions of the single-sided cepstrum; one being the slowly varying portion associated with the lower quefrency range between 0 and say 3-4mS; and second is the spike at about 9-10mS. The former is indicative of the spectral envelope distribution of the speech sequence in figure 5.45(a) and the latter is associated with the spectral fine structure or the average pitch period of the same speech sequence.

In order to select the position of the cepstral peak, a rectangular window was used to truncate the cepstrum below 5mS and above 15mS such that the resultant function is now shown in figure 5.45(c). This rectangular window corresponds to a maximum frequency of 200Hz and a minimum frequency of 66.7Hz so as to isolate the upper portion of the cepstrum from the lower portion. The width and position of the window can be varied to impose different frequency restrictions if required. Ascertaining the peak position is now a relatively trivial task. The reciprocal of the frequency value is numerically equal to the average pitch frequency F_0, of the sequence $\{b_j\}$, which now is used in equation (5.34) to drive the local oscillator at the fundamental frequency of the speech signal.

5.6.1.3 Pitch Detection by Data Reduction

Pitch detection by data reduction makes use of the time domain properties of the speech signal. It is a faster technique than the cepstrum method in terms of computer processing time. The original algorithm, developed by N J Miller,(ref 80) is accomplished in three phases. In the first phase, a data structure is constructed from speech samples using zero-crossings and energy measurements. This structure contains the candidates for the pitch period markers. Secondly, the number of candidate markers within this structure is reduced using syllabic segmentation, coarse pitch frequency estimations and discrimination functions. Finally, the remaining pitch period markers are corrected to compensate for errors introduced by the data reduction process.

For the data actually used in the experiments (section 4.2), there were no errors introduced by the data reduction process, therefore the last phase of this algorithm was not invoked. The simplified algorithm performed the exact function as the original algorithm caters for more complicated situations of input speech data.

A. Data Structure Construction and Reduction

The first phase of the algorithm involves the construction of a data structure from the speech samples. As the input signal is bandlimited from 300 to 3400Hz and the sampling rate is 8kHz, the representation of speech by its sampled values is enormously large. an alternative representation is made by the 'excursion cycles' of

the speech waveform. An excursion cycle consists of that part of the waveform between successive zero-crossings (c.f. time encoded speech, section 3.4.2.3). The first excursion cycle that occurs in a pitch period is termed its principal cycle.

The first step in the construction of the data structure is to calculate the energy contained in each excursion cycle. As the unvoiced portions of the speech signal have shorter time intervals between adjacent zero-crossings than those in the voiced portions and the amplitude of unvoiced speech is generally less than that of voiced speech, it follows that the energy of each unvoiced excursion cycle is less than that of the voiced excursion cycle. A threshold in terms of the percentage of the maximum cycle's energy can be set to eliminate the unvoiced excursions. This greatly reduces the number of excursion cycles to be handled in the ensuing phase. Each excursion cycle in the data structure is a possible candidate for the principal cycle. The number of candidates can be further reduced by invoking another property of the speech waveform i.e. principal cycles exhibit either positive or negative extreme amplitudes. However, within a given speech signal, most principal cycles exhibit the same sign. These excursion cycles with opposite sign to that of the excursion cycle with extreme amplitudes can be eliminated.

The last phase of the simplified algorithm consists of three tests; these are the pitch markers' minimum separation check, maximum separation check and the frequency tolerance check. As the fundamental frequency normally lies between 50Hz to 500Hz, pitch

marker separations of less than 2mS or more than 20mS are erroneous and hence re-examined. The last test compares the adjacent pitch periods. The adjacent pitch periods should not differ from one another by more than 30%.

Unlike other pitch detection algorithms, this one does not rely upon other techniques for voiced/unvoiced classifications. The very presence of pitch markers indicates the presence of voiced speech; hence the speed of operation is increased.

5.6.1.4 Results

The results of the low frequency synthesis network outlined in figure 5.42 are based upon informal subjective listening experiences. Waveform SNR values were not considered appropriate in this case as the structure of the voiced waveform was not preserved. Figure 5.46 shows waveforms of the voiced /I/ in the test word "sister". Figure 5.46(a) shows the original 0 to 3400Hz version; figure 5.46(b) illustrates the 300 to 3400Hz bandwidth version and figure 5.46(c) illustrates the 0 to 3400Hz processed signal. Also for comparison, the 0 to 300Hz filtered signal from the original is shown in figure 5.46(d) together with the 0 to 300Hz synthesised signal in figure 5.46(e).

As may be noted from figures 5.46(d) the original base frequency components are rich in second harmonic constituents as well as the fundamental frequency due to the positive and negative envelope waveforms of figure 5.46(d) being dissimilar. In the low frequency

enhancement system, only the fundamental component is added to the 300 to 3400Hz bandlimited signal in figure 5.46(c). The waveform structures in figure 5.46(a)(i) and (c)(i) are different and thus SNR values were not invoked. Subjective listening tests obviously reveal the apparent "tinnyness" of the 300 to 3400 bandwidth signal as compared to the 0 to 3400Hz original when using the 10-word test sequence. Upon perception of the low frequency processed signal, it was apparent that the base frequency components were again reinstated into the telephone bandlimited speech signal with an improvement in the quality of the resultant speech signal. For some of the test words, however, notably "S K Harvey" there was some unnaturalness associated with the addition of the base frequencies. This seemed to indicate that the synthesised low frequencies were not "blending" appropriately with the the 300 to 3400Hz bandlimited speech to make the processed signal sound like the original 0 to 3400Hz signal. Again this suggests that merely to incorporate the fundamental component into the bandlimited signal may not suffice in naturally restoring the full base components into the received speech signal. These subjective results were only taken for the cepstrum method to detect the pitch period, the data reduction method was tested in a subsequent system outlined in the following section.

The initial conclusion from these results indicate that better quality speech might be produced if the low frequency synthesis network of figure 5.42 could include the facility of adding at least a second harmonic component as well as the fundamental frequency to the telephonic bandlimited speech signal. This would be expected to

enable the synthesised low frequencies to represent more faithfully the low frequency portion of the original speech and also to be more naturally blended to the available bandlimited speech.

5.6.2 Inclusion of the 2nd Harmonic Plus Spectral Shaping

In addition to using the cepstral peak to generate the fundamental frequency, the second harmonic was also included as shown in figure 5.47. This second harmonic frequency was simply generated by a sinewave oscillator working at twice the pitch frequency. In order to elaborate upon the performance of the system further, the amplitude coefficients, A_1 and A_2, of the fundamental and second harmonic oscillators respectively were controlled according to the vocal chord spectral response. These amplitude coefficients were approximately determined by the equation (5.43).

$$A_n = k \left[\frac{\text{Sin } (n\pi\Delta fT)}{(n\pi\Delta fT)}\right]^2 \qquad (5.43)$$

where k = a constant value

This expression is derived from the publication "Nature of the Vocal Chord Wave" by R L Miller.(ref 81) This investigation revealed that the vocal chord wave can be approximated by a triangular waveform such that its Fourier components are given by the above expression. To be precise, the values of A_n in equation 5.4.3 need to be emphasized by +6dB/octave to account for the time differentiation caused by the radiation of the speech signal from the mouth. In our experiments, this emphasis was not included.

5.6.2.1 Vocal Tract Frequency Response

After the low frequency excitation is obtained, the next step is to determine the vocal tract frequency response. In this case, the initial quefrency region of the cepstrum was used to produce the vocal tract response. The cepstrum was first multiplied by the time window given by the equation:

$$\ell(nT) = \begin{cases} 1 & nT < \tau_1 \\ \frac{1}{2}\left(1 + \cos(\pi(nT-\tau_1)/\Delta\tau\right) & \tau_1 \leq nT \leq \tau_1 + \Delta\tau \\ 0 & \text{elsewhere} \end{cases} \qquad (5.44)$$

where $\tau_1 = 4.75$ mS i.e. the width of the flat region of the window

and $\Delta\tau = 1.25$ mS i.e. the width of the smooth truncation of the time window.

This time window isolates the lower portion of the cepstrum from the upper region. By taking the Discrete Fourier Transform of the windowed double-sided cepstrum followed by exponentiating the DFT, the frequency response of the vocal tract may be obtained. This operation is detailed in section 3.35 concerning the homomorphic vocoder.

5.6.2.2 Results

Once the low frequency excitation and the frequency response of the vocal tract were determined, the missing 0 to 300Hz band of the telephone speech signal was synthesised using the network of figure 5.47. The fundamental frequency and the second harmonic were modulated by the response of the vocal chord wave as in equation 5.43 then by the response of the vocal tract before being added to the telephonic bandlimited signal of 300 to 3400Hz. The scheme was carried out for the ten word sequence and figure 5.48 shows the regenerated 0 to 300Hz signal prior to being added to the telephonic bandlimited signal; the waveform corresponds to the /I/ in "sister". It can be seen that this waveform compares more favourably to the 0 to 300Hz version of the original signal than does the waveform of figure 5.46(d) using the synthesis of the fundamental frequency only. The waveform of the processed signal is shown in figure 5.49 which again corresponds to the /I/ in the word "sister".

The subjective listening experiences of this revised system for the 10-word sequence unfortunately and rather surprisingly did not yield any improvement in speech quality and naturalness of sound over that obtained from the original base frequency synthesis network of figure 5.42. However, at least the baseband was apparent in the processed signal and no degradation in sound quality was noticed compared to our first baseband synthesis network.

Due to the lack of perceptual improvement of this latter scheme, it was considered necessary to pursue the investigations further in order to determine just how the relative mix of the fundamental frequency and second harmonic is composed in the original zero to 3400Hz bandwidth signal. Referring to figure 5.50, the energy of the fundamental frequency was determined by a zero to 150Hz lowpass filter followed by a full-wave rectifier and a first-order filter with a 10mS time constant. The approximate energy of the second harmonic was ascertained using a similar arrangement with a 150Hz to 300Hz bandpass filter and a 20mS time constant first order envelope detector.

The fundamental frequency was obtained from the 300 to 3400Hz bandlimited speech using the data reduction pitch detection algorithm described in section 5.6.1.3. The two envelope waveforms, $s_1(t)$ and $s_2(t)$ in figure 5.50 were used to control the relative magnitudes of the regenerated fundamental frequency and its second harmonic. The summation of these two modulated low frequency signals was added to the 300 to 3400Hz speech signal.

The waveforms of the output of the fundamental frequency envelope detector $s_1(t)$ is shown in figure 5.51(a) for the vowel /I/ in the word "sister" and correspondingly $s_2(t)$ is shown in figure 5.51(b).

The regenerated low band waveform is also shown in figure 5.51(c) and the spectrum of the zero to 300Hz band of the original signal may be compared with that of the processed signal spectrum in figure 5.51(d) and 5.51(e) respectively. Finally, the waveform of the processed signal is shown in figure 5.51(f).

As with the previous arrangements for low frequency synthesis, this "ℓf vocoder" scheme was used to process the 10-word sequence and tested for listening purposes. The informal perception seem to indicate that the output speech sound from this latest system possessed a somewhat lower quality than that of the previous schemes. The processed sound, although certainly having a pronounced presence of low frequency components, did appear very "machine-like" and buzzy. For some of the words, most notably "S_oK Harvey" and "fist", it appeared that two people were speaking which resulted in a lack of speaker identity.

The possible reasons for the slightly disappointing outcome may be due to the fact that the relative delay between the 300Hz to 3400Hz telephone speech, the fundamental envelope $s_1(t)$ and the second harmonic envelope $s_2(t)$ were not synchronous. This obstacle can be overcome by replacing the first-order IIR filters with FIR filters to determine $s_1(t)$ and $s_2(t)$ plus a compensating all-pass FIR delay line inserted into the 300Hz to 3400Hz speech path.

Secondly, the data reduction pitch detection algorithm may not be as accurate as the cepstrum method since pitch peaks may be up to 20mS apart in the time waveform and the computed value of the pitch period is only updated at every principal excursion cycle, therefore the variation of the computed pitch period is discrete with a maximum possible time resolution up to 15-20mS. On the other hand, the cepstrum sliding block method ensured that the computed value of the pitch period was updated every 8mS irrespective of the separation between the pitch peaks in the speech time waveform.

When the experiment was repeated using the FIR filters to produce the envelope signals plus the cepstrum method to detect the pitch period, the output signal now appeared to have an improved quality compared to the previous method of performing this experiment. In fact, it produced a quality sound better than the output signals obtained from the two other baseband synthesis experiments. There still however appears to be much more work required to improve further the naturalness of the processed signal.

5.7 Discussion, Conclusion and Recommendations

In this chapter, we have discussed various methods of attempting to improve the quality of speech signals which have been bandlimited to within the telephone frequency range. As initially stated, quality improvement in this case means to render the processed signal perceptually as though it were conveyed over a wider-than-telephone bandwidth channel. The aim was to apply all or at least the bulk of the processing at the receiver without any processing or multiplexed side information at the transmitter. However, during the course of these investigations, it was found necessary to digress from this elusive goal. It was desired to examine what might be achieved by various forms of spectral extension algorithms before actually attempting to conform these algorithms into receiver-only systems.

The first investigation of high frequency regeneration was by spectral duplication. Although it was conceptually simple, it did not really offer evidence for a suitable solution as the resultant perceptual quality was found to be extremely unpleasant especially

for unvoiced sounds. This was possibly due to the fact that the spectral distribution in the 300 to 3400Hz band is generally quite unlike the spectrum in the range above 3400Hz.

The second experimental investigation was directed towards synthesising an out-of-band formant in voiced speech and again this did not prove to be advantageous as far as enhancing the speech quality was concerned but at least the process did not severely degrade the speech signal as in the case of the spectral duplication. The main drawback of this notion was that the resonant network generated a disturbing high frequency whistle during the voiced speech and this was supposed to be an additional formant to improve the quality. A fruitful outcome of this investigation is that the non-linear-transfer functions used for the spectral extension of the excitation function to drive the resonant circuit were subsequently used to facilitate high frequency regeneration for unvoiced speech sounds. The investigation shown in figure 5.16 seemed to indicate that perhaps it is better to enhance the high frequencies of unvoiced speech without permitting any processing to the bandlimited version of the voiced speech. Discerning unvoiced speech, particularly fricative sounds, from 300 to 3400Hz speech was found difficult and unreliable from the last chapter, therefore it was necessary to incorporate the decision process at the transmitter, as shown in figure 5.18 and transmit the decision by modulating a 2600Hz pilot signal within the 300 to 3400Hz bandlimited channel. Before conducting the experiments using the non-linear transfer function a simple method of adding bandpass filtered noise to telephonic bandlimited unvoiced speech was used;

the method was further refined by shaping the time waveform of the noise in accordance with the envelope of the wideband unvoiced signals. However, it seemed clear that neither of these ideas would be acceptable for enhancing the quality of telephone speech. This is not surprising since it is seen from the waveforms of figure 4.2 that unvoiced sounds, like voiced sounds, have characteristic time waveforms and associated spectra. It therefore appears unlikely to be able to replace successfully these unvoiced sounds by random signals such as the output of a noise generator. The non-linear transfer functions subsequently used do however attempt to extend the spectra of the received bandlimited signals by high frequency regeneration related to the bandlimited spectrum and hence they may be expected to yielded more successful results. The 'W' transfer function apparently yielded a slightly more regularly structured time waveform than the cubic transfer function of section 5.4.1.3 thus the former seemed more suited to enhance the /s/ sound whilst the latter seemed more successful when artificially producing an /f/ sound. The hybrid regenerator (section 5.4.1.4) also suggested better results for processing the /s/ and /f/ sounds. It did appear, however, that the waveform associated with the /ʃ/ sound had shown very little difference between the 300 to 7600Hz and the 300 to 3400Hz version. The /ʃ/ sound thus seemed better left unprocessed, i.e. no attempt was made to extend spectrally this sound over its telephonic bandlimited version. The same argument was applied to the stop consonants and the informal subjective listening tests described in section 5.4.1.6 seemed to support these findings.

Since it appeared simpler to detect the presence of /s/ sounds, the method described in section 5.4.1.5 may be the easiest to implement. As mentioned in Chapter 4, the /f/ and /θ/ sounds were difficult to detect reliably from equipment noise even at the transmitter operating upon the original wideband signal. These /f/ and /θ/ sounds may be generally too low in relative amplitude level to perhaps offer anything but a marginal listening quality improvement. Consequently, these sounds were found to be best left unprocessed at the receiver. The best method so far was found by just applying the non-linear 'W' transfer function process for the /s/ utterances. The research was now directed towards shaping the time waveform of the output of the 'W' transfer function during the /s/ sounds. Smoothing the leading and trailing edges of the HFR signal did not prove beneficial, but shaping the whole of the HFR waveform in accordance with the envelope of the original 300 to 7600Hz /s/ signal showed encouraging results on perceptual testing. The processed signal was further improved by incorporating a spectral peak in the regenerated /s/ signal using the very same resonant filter as that employed for the hf formant synthesis for voiced speech.

Having pursued this course of investigations, it was thought useful to consider the application of digital codecs in order to replace the analogue channel in figure 5.18. Apart from the multiplexed side information, the signal transmitted via the digital channel prior to coding is the 300 to 3400Hz speech signal itself, therefore currently developed codecs were used for the task. For the four coders tested, it was found that ADPCM generated the best SNR,

dynamic range and speech quality; its performance was followed by that of CFDM, APCM and CVSD. Moreover, when the comparison of using the coders with and without high frequency regeneration was made, this clearly indicated that a wider spectrum of speech signal was perceived when using the HFR network in figure 5.29 over that signal without any HFR method at the same transmission bit rate. Even at 16 kb/S, the ADPCM coder appeared to produce a more intelligible signal (regarding informal subjective listening tests) with high frequency regeneration for /s / sounds than the signal without the HFR. This effect seemed to be due to the fact that the "muffling" caused by band limitation and quantization noise generated by the low transmission bit rates was masked by the regeneration of the hf /s / sounds.

Having dealt with some aspects of high frequency regeneration and incorporating digital coders, still further investigations are required into the area of spectrally blending the high frequency regenerated fricative signals with the bandlimited voiced signal. Abrupt changes in signal bandwidth at the leading and trailing parts of speech utterances were sometimes perceptually annoying and unnatural. As mentioned in Section 5.5.5 one route possibly worth investigating is to transmit information regarding the level of the spectral peak of the /s/ utterance such that the HFR processed signal appears more natural. It may, on the other hand, be argued that the development of the system now tends towards that of Schröders voiced excited vocoder(ref 47) discussed in section 3.5.1 but the HFR method can be described as being more speech-utterance specific than that particular vocoder.

Turning our attention now to low-frequency regeneration of telephone bandlimited speech, the general method here was first to derive the fundamental frequency and then subsequently the second harmonic frequency from the signal available at the receiver. These frequencies plus the time and spectral shaping were also discerned from the 300 to 3400Hz signal at the receiver. Two pitch detection algorithms were utilised, namely the cepstrum method and the data-reduction scheme. The cepstrum approach proved to be more accurate but absorbed more computation than the data-reduction method.

The subjective results of the methods consistently showed that base frequency components were reinstated and blended with the 300 to 3400Hz speech fairly well for some of the words in the 10-word sequence, notably "sister" and "father". Using utterances such as "S K Harvey" and "talk" revealed that the lower frequency components began to sound unnaturally mixed with the band limited speech signal. This may be due to the incorrect amplitude of the first and second harmonic frequences due to the uncompensated delay of the envelope detectors of figures 5.42 and 5.47. Also this may even be caused by inaccuracies of pitch detection by either the cepstrum or data reduction method.

Further work was then carried out in order to apply the channel vocoder technique to the zero to 300Hz baseband signal at the transmitter as indicated in section 5.6.2. The exception to the channel vocoder rule is that not one but two local oscillators at the receiver were varied and controlled by the pitch frequency

detector. Unfortunately the quality of speech obtained from this experiment proved to be lower than that of the previous methods, again possibly due to delay compensation problems. When delay compensation (using FIR filters) was provided, the results appeared to be better than before although there was still a little unnatural buzziness associated with the voiced speech particularly with the trailing portion of voicing. This will need further investigation.

5.8 Note on Publication[ref 82]

A paper entitled "Speech Quality Enhancement by high frequency band regeneration", in co-authorship with Dr C S Xydeas (thesis supervisor), has been published in a conference on "Digital Processing of Signals in Communications" by the IERE, Vol. 49, April 1981. The paper is an earlier version of section 5.4 and presents waveforms of figure 5.23. Taped demonstrations were also included in the conference.

Chapter 6

6.0 Bandwidth Compression of Wideband Speech

6.1 Introduction

In the previous chapter, we had focussed our attention towards the goal of obtaining wideband quality speech at the receiver output when the receiver input was only supplied with 300 to 3400Hz telephone bandwidth speech. Of course, some of the notions explained deviated substantially from this underlying aim but it was felt that they nevertheless did serve to aid a full solution to that directive. In this chapter we now turn back to the perhaps more simple objective, namely to compress a wideband source speech signal (of say, 300 to 7600Hz) into a signal of 300 to 3400Hz capable of being propagated along a telephone analogue input-output channel. The job of the receiver is to expand the signal spectrally (as opposed to the spectral extension detailed in Chapter 5) back to a 300 to 7600Hz reconstructed speech signal. This generalised approach is illustrated in figure 6.1. The wideband speech signal $x(t)$ is processed by a frequency compression algorithm to yield the 300 to 3400Hz bandlimited signal $y(t)$. The signal then emerges from the channel as $\tilde{y}(t)$. It is then fed into the receiver which applies the frequency expansion process to produce the reconstructed signal $\tilde{x}(t)$. Unlike the spectral extension processes described in Chapter 5, which employed various non linear transfer functions, the frequency expansion process is a complementary stage to the frequency compression process at the transmitter, therefore, in all

cases to be discussed, if the wideband speech signal $x(t)$ is not available, i.e. the frequency compression process is not applied at the transmitter, then $\tilde{x}(t)$ cannot in general be generated at the receiver alone.

The notion of frequency compression/expansion processing to reduce the channel bandwidth requirements is by no means new. All of the early vocoder principles can certainly fall into this very broad category. Some of the earlier schemes that facilitate bandwidth compression e.g. the voice excited vocoder, VEV[ref 47]; the sub-band coder, SBC [refs 49,50,52]; the adaptive transform coder, ATC[ref 53] and the pilot controlled overtone reproduction method, PICOR[ref 63] can easily be operated to provide spectral compression of wideband speech into a telephone bandwidth channel. These systems, which were discussed in Chapter 3, work continuously on the whole signal i.e. they operate during voiced speech, unvoiced speech and silence pauses. It is felt, however, that these or newer systems can be made more speech specific if some adjustment is made in their operation to adapt to changes in the mode of the input speech. For the case of wideband speech, it may be advantageous to employ some form of switching in the processor when the speech changes from voiced to the unvoiced state and vice versa. We discussed earlier that it is unvoiced speech that possesses the bulk of the energy in the higher frequency region whilst voiced speech has its energy generally confined to frequencies below 3400Hz. As the wideband VEV[ref 47] was designed to transmit 10kHz speech over a 3.5kHz channel, it therefore applies frequency compression to all the speech signal, and so during voiced speech the more

significant 2 to 3500 portion is unduly processed to afford the transmission of those less significant higher frequency components. During silent interval pauses, the noisy higher frequency components are also transmitted as faithfully as possible which perhaps degrades the overall subjective impression of the processed signal at the output of the receiver. It is considered that if the unvoiced speech is spectrally compressed while the voiced speech and silence noise are left unprocessed, then the combined signal should appear to be conveyed via a wider bandwidth than 3400Hz but without the accompanying distortion to voiced speech and increased silent interval noise that the voice excited vocoder (or related systems) produce. This leads us to the underlying framework upon which the notions and systems to be discussed in this chapter are based. In our early development stages of the systems the processor and inverse processor were first tested on the entire speech signal. This framework will now be discussed in the following section.

6.2 The General Compression System

The general framework is illustrated in figure 6.2. The 300 to 7600Hz input speech signal x(t) is first split into two components v(t) and u(t) by the voiced/unvoiced decision switch (V/UV) in figure 6.2(a). The switch may comprise of any one of the four methods discussed in Chapter 4 depending upon its suitability and the requirements of the actual algorithm for the processor and inverse processor stages in figure 6.2 (a) and (b) respectively. The voiced speech, v(t) is simply bandpass filtered between 300 and 3400Hz (i.e. the same bandwidth as the propagation channel between

the transmitter and receiver). The unvoiced speech containing the perceptually important high frequency components beyond the 3400Hz cut-off frequency are spectrally compressed by the processor module from 300 to 7600Hz bandwidth into the 300 to 3400Hz range, to yield the signal $u'(t)$.

Before being transmitted, the signals $v'(t)$ and $u'(t)$ are summed together by a combiner, $v(t)$ and $u(t)$ are rendered mutually exclusive by the V/UV switch, so provided that the filter delay is compensated in the processing stage $v'(t)$ and $u'(t)$ will also occur mutually exclusively such that they should be easily separable. The combined signal together with the multiplexed V/UV decision is then transmitted in the 300 to 3400Hz analogue channel (or analogue input-output channel).

At the receiver, the signal emerges as $\tilde{\ell}(t)$ where the V/UV decision information is demultiplexed to leave $\tilde{y}(t)$. The receiver can then decide whether $v'(t)$ or $u'(t)$ was transmitted using the V/UV decision $\tilde{d}(t)$ to operate the switch S_2. During voiced speech, $\tilde{v}'(t)$ is filtered by the 300 to 3400Hz bandpass filter to produce $\tilde{v}(t)$ and while unvoiced speech is present, $u'(t)$ is spectrally expanded from 300 to 3400Hz to 300 to 7600Hz by the inverse processor to yield $\tilde{u}(t)$. Again, if the filter delay is compensated in the inverse processor stage then $\tilde{v}(t)$ and $\tilde{u}(t)$ are mutually exclusive in time and can be added together to form the reconstructed output signal $\tilde{x}(t)$.

The signal $\tilde{x}(t)$ is not an exact replica of $x(t)$ even when noise free channel is used as $\tilde{x}(t)$ occupies the original bandwidth although never all of it at any one instant.

In a practical configuration of this system, it may be found necessary to modify the arrangement of figure 6.2 to that as shown in figure 6.3. This modified arrangement avoids the use of a 'hard' switch prior to the bandpass filter F_1 and F_2 thus preventing switching spikes and 'ringing' that would occur at the output of the corresponding filters in figure 6.2. The disadvantage here is that both the processors and the filters are operating on the whole signal which is of course a less economical use of computation time.

Before discussing the details of the processor and inverse processor in figure 6.3, it is considered worthwhile to examine the validity of the model used for the wideband speech compression algorithm. For this purpose, the processor and inverse processor are replaced with a straight forward link and the channel bandwidth is extended to 300 to 7600Hz, i.e. the same spectral width as in the input speech. The net effect now is that the voiced speech is filtered from 300Hz to 3400Hz whilst the unvoiced speech is transmitted with the full bandwidth of 300 to 7600Hz. With a noise free channel, this investigation is indeed the same as the experiment discussed in section 5.4, and for this reason the experiment need not be simulated again. It was suggested from the earlier experiment that a subjectively acceptable impression of the output speech may be obtained if the attenuation factor applied to the voiced speech was

set to one half. Assuming unity in-band gain filters, this can be achieved by placing a 3dB attenuator in the unvoiced path of the network, as shown in figure 6.4.

At present, an "auxiliary-wire" is used to obviate the requirement of multiplexing the voiced/unvoiced switching information into the channel signal. The relevant details concerning multiplexing of the voiced/unvoiced decision have been already briefly considered in section 5.4.1.

So now having presented the underlying framework upon which the processor and inverse processor are based, we now proceed to discuss some of the notions employed to facilitate the spectral compression and expansion algorithms.

6.3 The Voiced/Unvoiced Band Switching System

The voiced/unvoiced band switching system (VUBS) is used for transmitting wide bandwidth speech (here 300 to 6000Hz) over a commercial telephone channel of 300 to 3400Hz. According to our bandwidth compression framework of figure 6.3, when voiced speech is present, 3400Hz is adequate but when unvoiced speech is present, we only transmit frequency components from 3 to 6 kHz[1], i.e. a bandwidth of 3kHz and one that can also be passed over a 0.3 to 3.4kHz channel.

---

[1] The edge frequencies f_{c_1} and f_{c_2} (here 3 and 6kHz) are system parameters selected for quality of perception, where $f_{c_1} - f_{c_1} =$ 3kHz.

At the transmitter the 6kHz speech signal $x(t)$ is bandpass filtered to give a signal $x_1(t)$ having frequency components from 300 to 3400Hz as shown in figure 6.3. The signal $x_2(t)$ is derived from $x(t)$ by filtering $x(t)$ to extract frequency components between 3 and 6kHz which are then heterodyned down with the aid of a 6300Hz carrier frequency so as to occupy the 300 to 3400Hz band. Thus $x_1(t)$ and $x_2(t)$ now have frequency components in the same frequency band. The decision concerning whether $x_1(t)$ and $x_2(t)$ is made by the voiced/unvoiced switch explained in section 4.4. Briefly, the switch(ref 74) filters $x(t)$ into two 1000Hz bands. 0 to 1000Hz, and 5000 to 6000Hz and estimates the energy in each of these bands over 5mS and 1mS, respectively. These energy levels are compared and $x_1(t)$ is transmitted if the energy in the lower frequency band exceeds that in the higher frequency band, while $x_2(t)$ is transmitted if the energy in the higher band is the greater. Should the energies be comparable, as in the case of some stop consonants and voice fricatives, $x_1(t)$ is selected. After the energy comparison, delays of 4 and 2mS are incorporated depending upon whether the switch is changing from voiced to unvoiced, or from unvoiced to voiced speech respectively. These delays prevent spurious switching. In general when voiced speech is present $x_1(t)$ is transmitted, while if unvoiced speech is detected $x_2(t)$ is transmitted.

The receiver must decide whether $x_1(t)$ or $x_2(t)$ was transmitted. This can be done in a number of ways. The most certain method, which involves some waveform distortion, is to insert a tone near the top of the speech band (say 2.6kHz as in section 5.4.1) whose

phase or amplitude informs the receiver whether $x_1(t)$ or $x_2(t)$ was transmitted. A more elegant solution is not to transmit the decision of whether $x_1(t)$ or $x_2(t)$ was transmitted, but for the receiver to deduce this information. Two measures are effective: the correlation coefficient (significantly lower for unvoiced than voiced speech) and the number of zero-crossings which are higher for unvoiced speech than voiced speech. These methods, however, were not tested. If $x_1(t)$ is received it is passed to the loud speaker via the filter F_2 in figure 6.3, while if $x_2(t)$ occurs it must be heterodyned up to occupy its original band of 3 to 6kHz before being passed to the loudspeaker. In the case of silence the $x_1(t)$ signal path is used, i.e. no heterodyning is employed.

The VUBS system does not recreate the original 6kHz speech but a signal that occupies the 6kHz band, although as for the fundamental system of figure 6.3, never all of the 6kHz band at any instant. By arranging for voiced speech to reside in the conventional telephone band, and unvoiced speech in the higher band of 3 to 6kHz, speech close in quality to the original 6kHz speech may be perceived.

6.3.1 Results

The VUBS system was investigated by computer simulation. An ideal telephone channel of 300 to 3400Hz was assumed using a FIR filter of 255 coefficients (see Appendix A1.2). The 10-word sequence described in section 4.2 was used in the experiments from which figure 6.5(a) shows a segment of the 6kHz speech waveform for /Is/ in the word "sister". When this segment is subjected to the

bandlimiting effect of the telephone channel the waveform appears as in figure 6.5(b). The unvoiced /s/ is severely attenuated and distorted whereas the voiced /I/ is only marginally affected. The VUBS transmits the signal shown in figure 6.5(c) with an amplified unvoiced component. At the receiver, the VUBS demodulator produces the signal shown in figure 6.5(d) where the close correspondence with the 6kHz speech of figure 6.5(a) is evident. The spectrograms for the complete utterance "sister" are shown in figure 6.6. for (a) the original wideband speech, (b) telephonic speech, (c) transmitted VUBS signal and (d) the reconstructed VUBS speech. From figures 6.5 and 6.6 we see the closer correspondence, both in the time and frequency domains, of the 6kHz speech and the reconstructed VUBS signal transmitted over the 3400Hz channel, compared to the conventional 3400Hz speech.

Informal listening experiences on this small sample of material seem to confirm the waveforms and spectrograms in that the VUBS speech is preferable to telephonic speech although more material needs to be processed so that formal subjective listening tests can be undertaken to fully evaluate the performance of the system.

Finally, before leaving the VUBS method, it is worth mentioning that although frequency translation was used before transmitting the unvoiced signal between 3 and 6kHz, there is no reason why frequency translation cannot be applied to voiced signals between zero and 3kHz so as to include the base frequency components and thereby

affording an improved quality processed speech. This latest refinement has not been tested and it is therefore an obvious candidate for future investigations using the VUBS process.

Having discussed the relatively simple and inexpensive VUBS method, our attention is now drawn towards techniques performed in the frequency domain which is the topic of our next section.

6.4 Bandwidth Compression Techniques in the Frequency Domain

As we found in the previous section, the subjective quality of the reconstructed VUBS speech was preferable to the telephone bandwidth speech. However there was some discernible unnaturalness associated with the VUBS speech, particularly with the transitions between voiced speech and unvoiced speech and vice versa. This may be due to the fact that although unvoiced speech has significant and sometimes strong spectral components above the normal telephone bandwidth, the lower frequency components which are not conveyed by the VUBS are still considered to be important as far as the subjective listening impression is concerned.

Since for the fricative sounds the lower frequency components between 300 and 3400Hz have less energy than the range of 3000 to 6000Hz (e.g. as shown by the spectrum of the /s/ utterance in figure 4.3(b)) it may be considered necessary to apply some form of frequency coding which allows more emphasis to the higher energy upper band and less emphasis to the smaller energy lower band. This is indeed our objective here. To develop this approach in the

spectral domain we invoke the Discrete Fourier Transform (DFT) and apply this to contiguous blocks of speech samples. The basic spectral processor is illustrated in figure 6.7.

At the transmitter, the sequence of samples representing the input speech block $\{x_i\}$ are multiplied by a window function $\{w_{1i}\}$ and forward Fourier transformed by the DFT to yield the complex frequency coefficients $\{X_k\}$. After the spectral compression algorithm has been applied, the modified frequency components $\{Y_k\}$ are inverse Fourier transformed and weighted by a second window function $\{w_{2i}\}$, complementary to the first window function, $\{w_{1i}\}$, to produce a spectrally compressed time sequence $\{y_i\}$. The receiver network of figure 6.7(b) is identical in structure to the associated transmitter with the exception that the spectral compressor is replaced by a spectral expansion processor. Ideally, the recovered sequence $\{\tilde{x}_i\}$ should have the same bandwidth as the input sequence with as little distortion as possible.

To start with, the combined network of figure 6.7(a) and 6.7(b) is used on its own to afford spectral modification in order to transmit wideband speech via a telephone bandwidth channel and later the network will be inserted into the more basic model shown in figure 6.3.

Before discussing the processing under the strategy of figure 6.7, we need to consider a suitable choice of window weighting functions to impose upon the blocks of speech samples. We will therefore need to digress from the 'main-stream' of this chapter to make a brief study of window weighting functions.

6.4.1 The use of Windows for Harmonic Analysis with the DFT

The Discrete Fourier Transform is a method of estimating the spectrum of a discrete sampled block of time data. The DFT handles the time signal in a way quite different from that found in more traditional frequency analysis. When a frequency analysis is performed by the use of filters (analogue or digital) there is a steady flow of time data into the filter, which in turn produces a steady flow of filtered time data at the output. In this way, the time signal is continuously being processed. The DFT, on the other hand, transforms the time domain into the frequency domain in one block at a time (c.f. section 3.4.1.4A). Furthermore, in order to obtain discrete frequency components, it is assumed that one block represents one period of the time signal hence the original input signal needs to be time limited before the analysis. We can now concentrate on the way this time limitation is performed, i.e. the use of different time weighting functions.

Before investigating the effects of different weighting functions, we will graphically[ref 83] review the DFT to aid our understanding of the process. As demonstrated in figure 6.8, the process involves three steps: 1) time sampling, 2) time limitation and 3) time convolution or frequency sampling. These steps are not actually performed by the DFT analyser but serve only to clarify the difference between the integral and the discrete transforms.

6.4.1.1 Time Sampling

The continuous time signal and its frequency spectrum, obtained using the integral transform, are shown in figure 6.8(a) where:

$$G(f) = \int_{-\infty}^{\infty} g(t)\ \exp(-j2\pi ft)dt \qquad (6.1(a))$$

and

$$g(t) = \int_{-\infty}^{\infty} G(f)\ \exp(j2\pi ft)dt \qquad (6.1(b))$$

The expressions (6.1(a)) and (6.1(b)) are a Fourier Transform pair.

The time signal is sampled by multiplying it with an infinite series of impulses with a separation of T. The Fourier Transform of the sampling function is another series of impulses with a separation $F_s=1/T$. When this is convolved with the spectrum of figure 6.8(a), it results in the frequency spectrum of figure 6.8(c). The time sampling gives rise to the possibility of aliasing of frequencies which can be avoided by the correct use of an anti-aliasing filter.

6.4.1.2 Time Limitation

A rectangular window of length t_L is used to limit the number of samples in the time function as shown in figure 6.8(d). The Fourier Transform of this time window is a $\sin(x)/x$ function as also shown. The result of the time window multiplication is shown in figure 6.8(e), where the number of samples N, is given by $N = t_L/T$. In the frequency domain the time window multiplication is reflected

as a convolution of the two frequency spectra, introducing ripples into the frequency spectrum. The sin(x)/x function determines the filter characteristic of the analysis of each frequency component of the signal which gives rise to a leakage of power from one frequency component into the neighbouring frequencies. In the case of a rectangular time window the filter characteristic (sin(x)/x) will have a poor selectivity and limit the useful dynamic range to less than 40dB. Since the high levels of the sidelobes of the sin(x)/x are determined by the discontinuities of the rectangular time window, the use of a continuous and smooth time window (as detailed shortly) will diminish this effect.

6.4.1.3 Time Convolution

The time signal shown in figure 6.8(a) can be used for digital calculation. However, the corresponding frequency spectrum, being continuous cannot be calculated. What is required is the sampling of the frequency spectrum which can be achieved by multiplying with a sampling function in the frequency domain. The sampling function is shown in figure 6.8(f) where the impulses are separated by $\Delta f = 1/t_L$ corresponding to N samples within one period of the frequency spectrum, where

$$\frac{F_s}{\Delta f} = F_s \times t_L = \frac{t_L}{T} = N \qquad (6.2)$$

In the time domain the frequency sampling corresponds to a convolution of the time limited signal with impulses with a

separation of t_L. The effect of this is to produce a periodic time signal with t_L equal to the period. The resulting Discrete Fourier Transform is shown in figure 6.8(g).

Sampling in the frequency domain is associated with the so called "picket-fence effect". Since we do not take into account the full continuous spectrum, but only samples of it, this corresponds to observing the spectrum through a picket fence. If there are very peaked components in the spectrum we might not observe the correct maximum value, but only the lower values on the slopes of the peak. In the case of a frequency analysis using the rectangular window, the maximum error that this effect imposes is 3.9dB which can be reduced using a smoother window to a maximum of 1.4dB. The picket fence effect is sometimes referred to as the scalloping loss.(ref 84)

The overall result therefore is that the DFT gives an estimate of the true spectrum of the original signal to a degree which depends upon the aliasing of frequencies, the side lobe effect and the "picket-fence effect".

6.4.1.4 Time Weighting Functions

There are numerous(ref 84) types of window weighting functions which may be used prior to the DFT process some of these are now listed, together with a brief description of their transforms. These are all illustrated in figure 6.9.

(i) Rectangular Window - this transforms to a $\sin(x)/x$ function (Kernel)

(ii) Triangular Window - this transforms into a $[\sin(x)/x]^2$ function, it is the convolution of (i) of half extent.

(iii) Trigonometric ($\text{Cos}^{\alpha}(x)$) Windows - For a Hanning window, α is equal to 2, and its spectrum can be obtained by summing the spectra of a rectangular window and a single cycle of a cosine wave. This forms three kernels, one at the origin and two on either side. The sidelobes of the adjacent kernels are each about half the size and are of opposite phase of the sidelobes of the central kernel. The summation of the three kernels' sidelobes being in phase opposition tend to cancel the sidelobe structure and hence improves the performance of the Hanning window compared to the rectangular one.

(iv) Hamming Window - The Hamming Window can be thought of as a modified Hanning window. The Hanning window gives inexact cancellation of the sidelobes from the summation of the three kernels. A window may be constructed by adjusting the relative size of the kernels to achieve a more desirable form of cancellation. The Hamming window is given by:

$$w(n) = \alpha + (1-\alpha)\ \text{Cos}\left(\frac{2n}{N\pi}\right) \qquad (6.3)$$

where n is the time index,

α is a constant

and N is the number of data points in the sequence.

Perfect cancellation in the spectrum occurs when $\alpha = 25/46$ resulting in deep attenuation at the missing first sidelobe position. As there is a small discontinuity at the boundary of the time window, this results in a $1/\omega$ (-6.0dB per octave) rate of fall off of the sidelobe maxima. There is no such discontinuity with the Hanning window therefore the rate of fall off of the sidelobe maxima of this window fall off at the much more rapid rate of $1/\omega^3$ (i.e. -18dB per octave) since the discontinuity resides in the second derivative.

(v) Blackman Window - The general trigonometric summation of:

$$w(n) = \sum_{m=0}^{N/2} A_m \, \mathrm{Cos}\left(\frac{2\pi mn}{N}\right) \qquad (6.4)$$

subject to the constraint $\sum_{m=0}^{N/2} A_m = 1.0.$

defines the Blackman Window which reverts to Hanning and Hamming windows when only A_1 and A_2 are non-zero. The Blackman Window achieves more sidelobe cancellation when higher order coefficients are introduced. When exact coefficients are used to place zeros in the position of the sidelobe maxima, then equation (6.4) becomes an 'Exact-Blackman' Window.

(vi) Blackman-Harris Window - These are families of windows used for minimum sidelobe levels obtained by gradient search techniques.

Constructed Windows

These windows are products, sums and convolutions of simple functions. In general, they tend not to be good windows and occasionally they are very bad windows, but they do have certain desirable features, not the least of which is the attraction of simple functions for generating the window terms.

(i) Hamming Window - This is the sum of a rectangular and a Hanning window.

(ii) Riesz Window - This is a polynomial window. It is continuous and the simplest polynomial. Its transform rolls off as $1/\omega^2$.

(iii) Riemann Window - The Riemann Window is in fact a central lobe of a $\sin(x)/x$ kernel. Its transform is similar to that of the Riesz window (not shown).

(iv) De la Valle-Poisson Window - This is a piece-wise cubic curve obtained by self convolving four rectangular windows of 1/4 extent. With this window, there is a trade-off of main lobe width with sidelobe level (not shown).

(v) Tukey (Cosine Tapered) Window - This window attempts to set the data smoothly to zero at the boundaries while maintaining a flat weighting for the data in the central region of the

window. The transform of such a window exhibits an erratic array of sidelobe levels arising from the product of two component transforms, namely the cosine lobe and rectangular window.

(vi) Bohman Window - The Bohman window is the convolution of two 1/2 duration cosine lobes resulting in a roll-off of $1/\omega^4$ of the sidelobe maxima.

(vii) Poisson Window - This is a two-sided exponential window. There is a discontinuity at the boundary and as this becomes smaller, the sidelobe structure of the transform merges into the assymptote of the sidelobe maxima.

(viii) Hanning-Poisson Window - This is the product of both the Hanning and Poisson windows as the name suggests. It is similar to the Poisson window having a very large main lobe width (not shown).

(ix) Gaussian Window - From the generalised uncertainty principle, one cannot simultaneously concentrate both on a signal and its transform. If T_s is the mean square time and W_s the mean square bandwidth then

$$T_s \, W_s \leq \frac{1}{4\pi} \qquad (6.5)$$

The equality only being satisfied for a Gaussian pulse, therefore this might be a reasonable candidate for a window. When using the Gaussian window, it is, however, necessary to

truncate the tails. By restricting the pulse to a finite length window the minimum time-bandwidth condition is no longer maintained. If the truncation is greater than the 3-sigma point, the error should be small and the window should be a good approximation to the minimum time-bandwidth.

(xi) Dolph-Chebyshev Window - This is used in conjunction with antenna design which endeavours to achieve a narrow main lobe pattern for a given level of sidelobes. The window function uses mapping from the m^{th} order Chebyshev polynomial and m^{th} order trigonometric polynomial.

(xii) Kaiser-Bessel Window - This determines for a restricted energy signal the function of restricted time duration T_m which maximises the energy in the band of frequencies W_m.

(xiii) Barcilon-Temes Window - This determines the function which minimises the energy outside the band of frequencies W_B. Hitherto, this goal has been achieved by maximising the concentration of the transform at the main lobe.

6.4.1.5 Harmonic Analysis

We can now investigate the properties and consequences of some of the windows cited above with the use of the Discrete Fourier Transform. We have already discovered that one does not observe the continuous spectrum of the signal but only samples of it. When the time limited signal contains exactly an integer number of periods,

the frequency samples will be taken at the centre of the main lobe and will also fall into all the zeros between the sidelobes. Hence, the true peak value of the spectrum is measured. When the rectangular window and the Hanning window are used for weighting a single sinusoid of an integer number of periods, as shown in figure 6.10(a), a slightly broader bandwidth is shown for the Hanning window main lobe. This is because the DFT assumes periodicity of the time limited signal (1 period being equal to the time limit length) where the repetition of the signal will produce a perfect infinite sinusoid, from which any trace of the time limitation has disappeared. This gives a perfect line spectrum when a flat window is used but the Hanning window imposes amplitude modulation of the sinusoid which causes a widening of the spectrum. The peak amplitudes are the same in both cases.

Figure 6.10(b) shows a situation where the phase is such that the signal starts and stops at zero. When this signal is repeated in time, a perfect sinusoid is no longer obtained. Although the signal is continuous at the boundaries, there is a discontinuity in the first derivative. Furthermore, the signal will have a dc component. Accordingly, the frequency spectrum obtained using the flat weighting now shows the samples within the sidelobes. At low frequencies there are high amplitudes due to spectral leakage from negative frequencies while at high frequencies the amplitudes are relatively low due to the continuity of the signal. The main lobe is no longer sampled at its maxima therefore the amplitude of the signal spectrum appears to reduce by 3.9dB but the sidelobes are sampled at their maxima which is evident by the spectral spreading

effect of the observation. When Hanning weighting is applied, the modulated signal will be zero and have zero slope at the boundaries and very low dc level. This is seen from the frequency spectrum, which clearly shows the improved filter characteristic of the Hanning window and the reduced amplitude error of only 1.5dB.

In figure 6.10(c) the same signal has been analysed, but with a different phase. When this signal is repeated in time, a maximum discontinuity will be found at the boundaries, but now the signal will not have a dc component. For the rectangular window the high amplitudes are found at high frequencies due to the discontinuity at the boundaries. Again, using the Hanning window, the modulated signal will not be much different from that of figure 6.10(b) and the analysis essentially gives the same result.

The effects of boundary discontinuities have been illustrated by comparing the use of flat and Hanning weightings only, because with detection of single tones in broad band noise, nearly any window (other than the rectangle) is as good as any other. However, the detection of more than one tone in the presence of noise is affected dramatically by the choice of window used.

The windows are now tested for two-tone detectability, one tone being placed at 10 F_s/N and the other at 16 F_s/N, equal to 5 bins separation (1 bin is equivalent to 1 DFT point, i.e. F_s/N). The amplitudes of the tones are set to 1.0 and 0.01 respectively, giving 40dB separation. In this case, the two tones may be easily discriminated even by the use of a rectangular window.

The signal is now slightly modified such that the larger signal resides mid-way between two DFT bins, in particular at 10.5 F_s/N. The smaller signal remains in the 16^{th} bin position. The sidelobe structure of the rectangular window will now swamp the main lobe of the smaller one. The sidelobe structure will be only 25dB down at 5.5 bins from the centre of the rectangular window response. Hence the second signal would not be detected by the use of the rectangular window at all. There will be also some asymmetry around the main lobe at 10.5 bins due to the coherent addition of the sidelobe structures of the pair of kernels located at ± 10.5 bins, i.e. self leakage.

The other windows may also be applied to demonstrate two-tone detectability. This is presented(ref 84) such that the large signal is at the location corresponding to the worst case resolution.

In conclusion, it reported that the Blackman-Harris windows will perform best in detection of nearby tones of sufficiently different amplitudes. Another optimal window is the Kaiser-Bessel window but the Blackman-Harris window is simpler to implement. In comparison, the Hanning window separated the two-tones by a 3.0dB null, which is marginal in regard to the 20.0dB obtained with the Blackman-Harris window or the 19.0dB for the Kaiser-Bessel window.

It is therefore found that the choice of the window has a considerable effect upon multi-tone detection via the DFT. Maximum dynamic range of multi-tone detection requires the transform of the window to exhibit a highly concentrated central lobe with a very low

sidelobe structure (which is analogous to "shading" in antenna distribution design). Perhaps in future, the sidelobe structure may be suppressed by adaptive filtering techniques.

6.4.1.6 Application of Windowing for Bandwidth Compression

Although quite a comprehensive study has been made on the effects of different types of windows used in conjunction with the DFT, it was found that the continued use of the rectangular window and Hanning window suited our needs for the spectral analysis performed here. In figure 6.7, spectral analysis is performed by the processor itself to apply bandwidth compression in the frequency domain; after the spectral modification, it is necessary to invert the DFT back into a time sequence both at the transmitter and receiver. If one assumes for the moment that no spectral compression takes place at the transmitter of figure 6.7 then the time sequence at the output of the transmitter will be modulated by successive window functions which are applied to each block of samples. It is therefore necessary to apply inverse windowing to restore the blocks of time samples to a flat weighted sequence (i.e. to remove the modulation). Rounded windows reduce the data to low values at the block edges, so if the processor of figure 6.7 does indeed apply the spectral modification to the signal blocks then the error signal, or difference between the input and output time signals, will be exaggerated at the block edges by the inverse windowing. This results in gross block-end distortion, or periodic clicks, in the output signal. This problem can be completely alleviated by the use of the rectangular window which obviates the inverse windowing procedure.

The use of the cosine tapered (Tukey) window may also be used to remove the block-end distortion if the windowed blocks are overlapped such that the net weighting applied to the whole signal is constant. However, it was found(ref 84) that tapered windows did not perform well when tested by the criteria in section 6.4.1.5, and also overlapping blocks of samples will reduce the bandwidth compression efficiency of the complete algorithm in figure 6.7.

Another reason for the choice of the rectangular window in this case, is that it will shortly be seen that the block size, N, will be allowed to take on relatively large values (i.e. 256 or 512) such that the frequency resolution of the DFT, i.e. the value of 1 bin, will be fairly small. The swamping effect of the rectangular window sidelobe structure will therefore not be too drastic. It will also be appreciated that the absolute value of the spectral estimate of the signal is not the prime consideration but it is the spectral trend of the signal that is our premium.

In another example of our spectral analysis requirements, the DFT will be performed outside the processor path of figure 6.7, i.e. as a side branch. Frequency analysis in this instance is performed purely for observation purposes to monitor the signal spectrum which may then be plotted for some segments of the speech utterances shown in figure 4.3. The inverse DFT, and inverse windowing, should not generally be required thus the block-end distortion problems associated with the processor DFT will not be encountered here. We are free to choose any type of rounded window, knowing that reducing the weighting of data at the window edges is now acceptable. Again,

we note that the absolute value of the spectral estimate of each frequency component of the signal is not critical; we are still more concerned with the trends associated with the signal spectrum. With this in mind, plus other considerations regarding computational ease, the simple Hanning window was invoked. Although we have found that it is not an optimal window, the performance of the Hanning window regarding spectral "smearing" still exceeds that of the rectangular window.

So, to summarise the approach adopted here; if spectral analysis is implicit within the signal bandwidth compression/expansion process, then the rectangular window is used and if the DFT is used to obtain a spectral plot of the signal at any stage outside the processor path, then the Hanning window is applied.

6.4.2 Bandwidth Compression using the Second Order Spectrum

Having discussed in detail the consequences of time limitation and weighting the blocks of samples prior to the DFT process, we now return to the bandwidth modification algorithm of figure 6.7. In particular, we may proceed to detail the spectral compression expansion process; the first of our schemes considers the possibility of performing the forward DFT twice in succession i.e. to obtain the second-order spectrum.

The use of the second order spectrum has been performed by Niederjohn and Curtis(ref 85) for enhancing the intelligibility of speech in wideband random noise using spectral subtraction

techniques. Their method involved two forward and two reverse Fourier transformations with processing in both the first and second order spectral domains. The processing in the first order spectrum included a square root operation applied to each magnitude coefficient prior to the second forward DFT. In the second order spectrum, a "gating" operation was performed by setting to zero the magnitude of the five low "frequency" harmonics. The square root operation of the first order spectrum causes a concentration of noise energy relative to speech energy in a region near the origin of the second order spectrum if the square root operation is accompanied by a sign reversal of the real parts of all the odd numbered harmonics. The "modulation" in the first order spectral domain has the effect of reversing the scale of the axes in the second order spectrum. The concentration of noise in the second order spectrum may be removed by the gating process before the corresponding inverse operations are applied to the signal. As far as noise reduction is concerned, it was stated(ref 85) that the system did not appreciably increase the intelligibility of the signal but it did improve the "listenability".

We now consider the plausibility of the same approach for the bandwidth reduction of speech signals but perhaps with some modification to suit the objective. The postulated form of the transmitter terminal is shown in figure 6.11(a) and the corresponding receiver process is shown in figure 6.11(b). If the input signal is typified by $x(t)$ in figure 6.11(c)(i) and $X(f)$ representing its Fourier Transform in figure 6.11(c)(ii) then the square root function in figure 6.11(a) should have a smoothing (or

flattening) effect to $X(f)$ resulting in the function $\sqrt{X(f)}$ as sketched in figure 6.11(c)(iii). If this signal is now taken as the input to the second Fourier transformer (DFT2) then, being a relatively flat signal it should produce second order "frequency" components concentrated into the lower end of the second order frequency axis, f', as shown in figure 6.12(c)(iv). The input signal may now be transmitted more efficiently in terms of the second order frequency coefficients such that the higher f' values may be truncated by a window function operating in the second order frequency domain. At the receiver, the f' coefficients are inverse Fourier transformed and the magnitude of the first order frequency coefficients are squared before being inverse Fourier transformed again to yield the reconstructed time signal $\tilde{x}(t)$.

Before the complete system is investigated, we need to examine how well the second order spectrum packs the signal components down to the lower end of the axis. This was done by taking three voiced speech specimen blocks and applying the transmitter process up to $Y(f)$ Figure 6.12(a) shows the first speech block corresponding to $x(t)$, figure 6.12(b) and (c) show the magnitude spectrum and the square root of the magnitude spectrum respectively. It can be appreciated that the square root operation does not change the relative shape of the magnitude spectrum but halves the dB scale and also the phase (not shown), at every frequency ordinate.

The second order spectrum of the square-rooted first order spectrum is shown for the first speech signal specimen in figure 6.12(d). It can be seen that this second spectrum, $Y(f)$, does not appear bandlimited at all and has almost a completely random structure i.e.

no evidence of formants or a regular fine structure. The process was repeated for the other speech blocks of data shown in figure 6.13(a) and figure 6.14(a). The corresponding second order spectra are shown in figures 6.13(d) and 6.14(d) respectively. These also support the results for the earlier speech signal in figure 6.12. It therefore seems unlikely that this process can afford any bandwidth compression properties as it stands.

The reason for the rather discouraging result is perhaps due to the simplified description of the stages illustrated in figure 6.11(c)(i) and figure 6.11(c)(ii) not being valid. The distribution shown for the magnitude spectrum $|X(f)|$ is more representative of the spectral trend rather than the complete spectrum. More typical speech spectra, including the fine structure, are shown in figures 6.12(b), 6.13(b) and 6.14(b). It appears the spectral fine structure is responsible for preventing the square root operator from allowing sufficient smoothing of the spectrum. Therefore, it is now expected that the second order spectrum will not in general produce a bandlimited appearance.

If the second order spectrum had shown any consistent band limitation, then truncating a large portion of this spectrum in the upper "frequency" region would cause an inevitable error in the first order spectrum. This error is spread right across the band of the first order spectrum which is then amplified by the squaring operation in the receiver. The inverse Fourier transformer would then yield the reconstructed time signal with a noise component occupying all frequencies which may be subjectively very disturbing when listening tests are conducted on the output speech signal.

One method for rendering the model shown in figure 6.11(c) more realisable is to separate the spectral fine structure and the spectral envelope such that the spectral envelope (now being smoother than the full spectrum) may be treated by the second order spectrum process. The spectral fine structure may then be transmitted independently as an excitation signal. However, the system described is now becoming closely aligned with Homomorphic techniques such as the cepstrum vocoder method explained in section 3.3.5. It therefore appears that the complexities involved with this notion would now be relatively high and so it was decided at this stage to direct our investigations towards perhaps more simple methods of speech bandwidth compression. We now return to the more conventional first order spectral domain in which to develop a process termed "Frequency Mapping".

6.4.3 Spectral Compression by Frequency Mapping (FMAP)

In this section, we describe a technique of frequency mapping (or frequency warping) in order to facilitate bandwidth compression of speech signals. The notion has also been utilised to correct helium-speech from divers(ref 86) such that when the divers' speech signal reached the receiver, the formants had undergone an upward frequency shift due to the nature of the helium mixture in the environment in which the divers were speaking. To restore the formants to their original position, spectral warping was applied to perform this task without altering the pitch frequency.

For our application, the frequency mapping process is conducted as follows:

Contiguous blocks of speech samples are Fourier transformed by the DFT. The magnitude and phase of the frequency components are found such that every 'c^{th}' component is selected, the remainder being discarded. The selected components are translated together into a smaller and lower frequency band and inverse Fourier transformed by the IDFT to produce the frequency mapped time sequence. The process will now be explained in more detail.

As in section 5.2.1, we start by considering the block of N speech samples represented by the sequence $\{x_i\}$. The DFT of $\{x_i\}$ is the sequence $\{X_k\}$ whose components are

$$X_k = \frac{1}{N} \sum_{\ell=0}^{N-1} x_i \exp\left(-j \frac{2\pi k i}{N}\right) \qquad (6.5)$$

$$k = 0,1,\ldots,N-1$$

The magnitude of X_k is

$$|X_k| = \left\{Re^2(X_k) + Im^2(X_k)\right\}^{1/2}$$

$$k = 0,1,2,\ldots,N/2 \qquad (6.6a)$$

where Re(.) and Im(.) are the real and imaginary values of (.).

The phase of X_k is

$$\theta_k = \frac{180}{\pi} \text{Tan}^{-1} \left[\frac{\text{Im}(X_k)}{\text{Re}(X_k)} \right] \qquad (6.6b)$$

$$k = 0,1,2,\ldots,N/2$$

and equations (6.6a) and (b) are the polar representation of X_k. Again, because of the symmetrical nature of $\{X_k\}$, there are (N/2)+1 distinct components in each of the magnitude and phase components of X_k, the values of X_k with k = N/2+1 to N-1 are discarded.

Frequency compression is now applied to the $\{|X_k|\}$ and $\{\theta_k\}$ sequences. For simplicity, assume that the compression is linear and of value c. This means that one sample of $\{|X_k|\}$ and $\{\theta_k\}$ is selected every c samples, and the remaining c-1 samples are set to zero. We now frequency map those samples not set to zero so that the mapped samples occupy a lower frequency band than the original speech signal. By this means frequency compression is achieved. Representing the mapped frequency sequence $\{Y_\ell\}$

$$|Y_\ell| = \begin{cases} X_\ell & ; \quad \ell = 1,2,\ldots,\text{INT}\left[\frac{N/2}{c}\right] \\ 0 & ; \quad \ell = (\text{INT}\left[\frac{N/2}{c}\right] + 1),\ldots N/2 \end{cases}$$

and

$$\lambda_\ell = \begin{cases} \theta_\ell & ; \quad \ell = 1,2,\ldots,\text{INT}\left[\frac{N/2}{c}\right] \\ 0 & ; \quad \ell = (\text{INT}\left[\frac{N/2}{c}\right] + 1),\ldots N/2 \end{cases}$$

where $|Y_\ell|$ and λ_ℓ are the magnitude and phase of Y_ℓ, respectively, and INT[(.)] is the rounded down integer value of (.). For example, if N is even and c=2, $X_2, X_4 \ldots, X_{N/2}$ are mapped to Y_1, $Y_2, \ldots, Y_{N/4}$ and $X_1, X_3, \ldots X_{(N/2-1)}$ are omitted.

Having completed the frequency mapping (and returning to the general case for the value of c), the $\{|Y_k|\}$ and $\{\lambda_k\}$ sequences are converted into real and imaginary sequences whose components are given by

$$Re(Y_\ell) = |Y_\ell| \; Cos(\lambda_\ell)$$

$$Im(Y_\ell) = |Y_\ell| \; Sin(\lambda_\ell)$$

for $\ell = 0, 1, \ldots N/2$. To extend the number of components in the real and imaginary to N we invoke the complex conjugate property of spectral components namely

$$Y_{N-k} = Y_k^* \quad , \quad k = 0,1,\ldots,N/2 \tag{6.7}$$

where the raised * above the symbol implies its complex conjugate, and

$$Y_k = Re(Y_k) + j \; Im(Y_k)$$

The inverse DFT is now performed to give the time sequence $\{y_n\}$, where

$$y_i = \sum_{k=0}^{N-1} Y_k \exp\left(j \frac{2\pi k i}{N}\right) \tag{6.8}$$

$$i = 0,1,\ldots,N-1$$

$\{y_i\}$ is the time sequence whose frequency components occupy a lower frequency band compared to those of $\{x_i\}$. Specifically the highest frequency component in $\{y_i\}$ is

$$f_c = \frac{F_s/2}{c} \tag{6.9}$$

where F_s is the sampling frequency. For example, if we start with speech bandlimited to 7.6kHz and sampled at F_s = 16kHz, a value conveniently greater than the Nyquist rate then f_c = 8/c. If c=2 the 7.6kHz speech has been mapped into 4kHz.

6.4.3.1 Frequency De-Mapping

The receiver decodes the signal to produce the compressed speech sequence $\{\tilde{y}_i\}$, where a tilde ($\sim$) above a symbol signifies its existence at the receiver. The first step in frequency de-mapping $\{\tilde{y}_i\}$ into the final output sequence $\{\tilde{x}_i\}$ is to apply the DFT to $\{\tilde{y}_i\}$,

$$\tilde{Y}_\ell = \frac{1}{N} \sum_{i=0}^{N-1} \tilde{y}_i \exp\left(-j \frac{2\pi \ell i}{N}\right) \tag{6.10}$$

$$\ell = 0,1,\ldots,N/2$$

and then to convert $\{\tilde{Y}_\ell\}$ into real and imaginary sequences. From these we obtain the polar co-ordinates,

$$|\tilde{Y}_\ell| = \left\{Re^2(\tilde{Y}_\ell) + Im^2(\tilde{Y}_\ell)\right\}^{1/2} \tag{6.11}$$

$$\ell = 0,1,\ldots,N/2$$

and

$$\tilde{\lambda}_{\ell} = \frac{180}{\pi} \text{Tan}^{-1} \left[\frac{\text{Im}(\tilde{Y}_{\ell})}{\text{Re}(\tilde{Y}_{\ell})} \right] \qquad \ell = 0,1,\ldots,N/2 \tag{6.12}$$

The spectrum now has to be "stretched out" by restoring all the frequency components back to their original spectral locations. The receiver has à priori knowledge of these positions as it is provided with the mapping characteristic employed at the transmitter. Thus no side-information need be transmitted related to the spectral positions. The sequence to be formed consists of the stretched-out components and the estimated components. Let us first consider the magnitude sequence $\{|\tilde{X}_k|\}$. Its stretched out components are given by

$$|\tilde{X}_k| = |\tilde{Y}_{k/c}| \tag{6.13}$$

$$k/c = 0,1,\ldots,\text{INT}\left[\frac{N/2}{c}\right]$$

where k/c is an integer.

The estimated components occur for non-integer k/c values, and the magnitude of these estimated components are determined by a straight line interpolator that uses stretched-out components on either side of the components to be estimated. Specifically

$$|\tilde{X}_{k+u-1}| = |\tilde{X}_{k-1}| + u\, G_m \tag{6.14}$$

$$u = 1,2,\ldots,c-1$$

where

$$G_m = \frac{|\tilde{X}_{k+c-1}| - |\tilde{X}_{k-1}|}{c} \tag{6.15}$$

$$m = \text{INT}\left[\frac{k}{c}\right] + 1 \tag{6.16}$$

and $k = 1, c+1, 2c+1, \ldots, N/2$

The combined mapping, de-mapping and interpolation process applied to the magnitude frequency components are illustrated in the schematic diagram of figure 6.15.

The phase of the stretched out components in $\{\tilde{X}_k\}$ is known and is

$$\tilde{\theta}_k = \tilde{\lambda}_{k/c} \tag{6.17}$$

$$k/c = 0,1,2,\ldots \text{INT}\left[\frac{N/2}{c}\right]$$

The phase assigned to the estimated components, i.e. for those components where k/c is non-integer, is also determined by a straight line interpolator that operates in the same manner as for the magnitude components. Again equation (6.14) is used with $|X_{k-1}|$ and $|\tilde{X}_{k+u-1}|$ replaced by $\tilde{\lambda}_{k-1}$ and $\tilde{\lambda}_{k+u-1}$ respectively. Likewise in equation (6.15), $|\tilde{X}_{k-1}|$ and $|\tilde{X}_{k+c-1}|$ are replaced by $\tilde{\lambda}_{k-1}$ and $\tilde{\lambda}_{k+c-1}$.

The stretched out and interpolated magnitude and phase sequences $\{|\tilde{X}_k|\}$ and $\{\tilde{\theta}_k\}$ are converted to real and imaginary sequences, where

$$\begin{aligned} \mathrm{Re}(\tilde{X}_k) &= |\tilde{X}_k| \ \mathrm{Cos}\ \tilde{\theta}_k \\ \mathrm{Im}(\tilde{X}_k) &= |\tilde{X}_k| \ \mathrm{Sin}\ \tilde{\theta}_k \end{aligned} \tag{6.18}$$

for $k = 0, 1, \ldots, N/2$ and using

$$\tilde{X}_{N-k} = \tilde{X}_k^* , \qquad k = 0,1,\ldots,N/2 \tag{6.19}$$

we form the recovered speech sequence $\{\tilde{x}_i\}$ according to

$$\tilde{x}_i = \sum_{k=0}^{N-1} \tilde{X}_k \exp\left(j \frac{2\pi k i}{N}\right) \tag{6.20}$$

Notice that the rectangular window weighting function has been effectively used in all instances where the DFT and IDFT are invoked.

6.4.3.2 Frequency Mapping System

Initially, the frequency mapping system was implemented in perhaps its simplest mode i.e. a linear mapping law was used and the process was applied to the entire speech sequence. The system is schematised in figure 6.16.

In our case, the lower frequency of the input signal is 300Hz although the system can operate for frequencies down to 0Hz (dc). The linear frequency mapping stage is represented pictorially in

figure 6.16 by a frequency transfer function of zero to f_1 input band and zero to f_2 output band. The transfer function is reversed for the de-mapping process. Initially, the values of f_1 and f_2 were taken as 5.0kHz and 2.5kHz respectively giving a corresponding value of c equal to 2.

Figure 6.17(a) shows a block of voiced speech data used as the input to the frequency mapping transmitter and figure 6.17(b) shows the corresponding spectrum. The frequency mapped and de-mapped spectra are shown in sub-figures (c) and (d) respectively. The output spectrum at the receiver is shown in figure 6.17(e) with the corresponding time sequence as illustrated in figure 6.17(f). It may be noted that the sampling frequency of the input speech is only 10kHz as this data had been recorded prior to that in section 4.2 and was used to conduct the experiments in the early stages of the project to test the feasibility of this system.

When the input and output time sequences are compared, one can see that the output time sequence bears little resemblance to the input signal and has virtually lost most of its recognisable pitch structure. This is possibly because the estimated frequency components interpolated in the low frequency bands cause errors that are very significant to the structure of the output time waveform. It may be argued that the preservation of the structure of the time waveform is not of paramount importance in terms of listening tests as the ear is phase insensitive[ref 33] (compare the output of a non-linear phase channel such as a Butterworth filter).

This latter argument was not borne out by the linear frequency mapping system as when listening tests were conducted, the output "speech" was unintelligible and sounded like a "burble".

In order to avert the problem of generating very perceivable errors in the low frequency region of the output signal, it was considered necessary to preserve the low frequency band of the signal at the expense of coarser interpolation for the high frequency band. To implement this notion, a piecewise linear mapping law was used which effectively uses 1:1 mapping from zero to f_B and c:1 mapping from f_B to f_1 as depicted in figure 6.18.

The value of f_1 was maintained at 5.0kHz and the value of f_2 was increased to 3.0kHz. Initially the value of f_B was set to 2.5kHz such that the base band signal between zero and 2.5kHz was transmitted without any effective processing and the region 2.5 to 5.0kHz was frequency mapped into 2.5 to 3.0kHz band with the compression ratio c, equal to 5, where at the receiver four out of every five frequency components were interpolated. The system was now operated in a similar fashion to the voice-excited vocoder(ref 47) where again the baseband was transmitted intact whilst the upper frequencies of the signal were vocoded. The essential difference is that the frequency mapping system is based on block processing whereas the vocoder is operated continuously on the speech signal. The frequency mapping system therefore exchanges time resolution for frequency resolution. The number of frequency components in the frequency mapping system depends upon the transform size and is generally larger than the number of vocoder channels in the VEV system.

The result for the piecewise-linear frequency mapping system is shown in figure 6.19. The system was applied to the input data of figure 6.17(a). Figure 6.19(a) shows the inverse mapped spectrum at the receiver and figure 6.19(b) shows the output spectrum of the processed input signal corresponding to the original spectrum of figure 6.17(b). Figure 6.19(c) shows the output time waveform from the receiver. It can be seen that the original waveform structure is now more evident in the output signal. However there is still a presence of rather large block-edge distortion components, which incidentally is not a problem with the voiced-excited vocoder.

When listening tests were conducted on the output signal, it seemed apparent that frequencies beyond 3.0kHz were present when the output signal was compared with the same input speech signal lowpass filtered from zero to 3.0kHz. Unfortunately this rather marginal attribute was severely masked by the very annoying block synchronous "clicks" that accompanied the output speech signal. This was of course expected since they were evident from the time waveform of figure 6.19(c).

At this stage it was considered appropriate to investigate the cause of these block synchronous clicks before proceeding any further with the optimisation of the bandwidth compression properties of the frequency mapping system. Undoubtedly, the frequency mapping process causes interpolation errors in the spectrum for both the magnitude and the phase components. The interpolation error worsens as the compression ratio, c, increases for whatever part of the spectrum that the compression is applied. The question now

arises, what is the effect of the spectral interpolation error upon the recovered time sequence? The answer is considered to be inherent in the inverse DFT process, i.e. equation (6.20) which can be written in matrix form:

$$\left[\tilde{x}_i\right] = \left[W_N^{ki}\right]\left[\tilde{X}_k\right] \qquad (6.21)$$

where $[\tilde{x}_i]$ and $[\tilde{X}_k]$ are column vectors containing the N time samples and N frequency coefficients respectively and $[W_N^{ki}]$ is a square matrix of order N containing the complex unit vectors, $\exp(-j\frac{2\pi ki}{N})$. For a particular case of N=8 the matrix equation may be visualised as shown in figure 6.20. Each of the arrows in the square matrix represents a complex unit vector relative to the axes shown. It can be seen that as the time ordinate, i, of $\{\tilde{x}_i\}$ increases, the rate of rotation of the vector with respect to each frequency ordinate, k, of $\{\tilde{X}_k\}$ also increases.

For low values of i, e.g. i = 0, the complex vectors are all facing the same direction and for the first sample, $\tilde{x}_0$, the frequency coefficients will be summed coherently (refer to equation 3.78). Conversely, at the centre of the time sequence, e.g. i = 4, the complex vectors are in opposition such that the addition of the frequency coefficients $\{\tilde{X}_k\}$, for determining $\tilde{x}_4$, will tend to cancel each other. Consequently, if the frequency coefficients $\{\tilde{X}_k\}$ contain a correlated spectral noise component such as the interpolation errors incurred in the frequency de-mapping procedure, then the noise will tend to concentrate at the origin of the block

of time samples with a small residual noise contribution in the centre of the block $\{\tilde{x}_i\}$ which was indeed indicated in our experiences.

In order to attempt to reduce these annoying block synchronous clicks in the output speech signal, it may be considered subjectively more tolerable to spread the interpolation errors across the whole of the recovered time sequence. The possibility of achieving this objective is by "upsetting" the ordered pattern of arrows (or phasors) in the square matrix of figure 6.20. If the phase of the estimated frequency components of $\{\tilde{X}_k\}$ are assigned random values rather than interpolated values in the frequency de-mapping process then the effect would be to randomise the pattern in figure 6.20 and so to decorrelate the errors arising from the interpolated magnitude values of $\{\tilde{X}_k\}$, thus avoiding their concentration at the edges of the sequences.

Instead of using the straight line interpolator for the phase of the estimated frequency components, the phase $\tilde{\theta}_k$ was obtained by

$$\tilde{\theta}_{k+u-1} = 180(R - 0.5) \qquad u = 0,1,\ldots,(c-1) \tag{6.22}$$

Where R is a uniformly generated random number between 0 and 1.

Equation (6.22) is used together with (6.17) to form the complete phase spectrum of $\{\tilde{X}_k\}$. So when k/c is an integer, equation (6.17) is used and when k/c is non-integer we use equation (6.22) to form the values of $\{\tilde{\theta}_k\}$.

The magnitude coefficients $\{|\tilde{X}_k|\}$ from equation (6.13) and (6.14) plus the phase values from equations (6.17) and (6.22) are converted into cartesian coordinates and then into the new recovered time sequence by the IDFT in equation (6.20).

It was found that the random-phase refinement did show a reduced level of block synchronous clicks in the recovered speech signal which was evident in the listening tests performed on the output speech signal.

Now that the level of block synchronous clicks was reduced, the more general lower level noise and distortion which appeared present in the output signal from the frequency mapping algorithm was unmasked. During the passage of sustained fricative sounds, particularly the /s/ utterance, the subjective impression of wide bandwidth output signal clearly seemed to dominate any deleterious effects of the aforementioned system. It was therefore considered at this stage to process unvoiced sounds only and allow the voiced sounds and silence pauses to be transmitted without any bandwidth modification. We now of course return to the general compression system of figure 6.3 in which the processor and inverse processor are replaced with the transmitter and receiver of figure 6.16 respectively. The complete system now appears as shown in figure 6.21. The input speech of 7.6kHz bandwidth is separated into voiced and unvoiced paths. The voiced speech is filtered from 300 to 3400Hz, while unvoiced speech is frequency mapped into this bandwidth. A delay 'D' is inserted into the unvoiced path to allow for the difference in delay between the voiced and unvoiced paths.

The wideband input speech signal is separated into blocks of 8mS each and identified as voiced or unvoiced according to the first-shift autocorrelation strategy outlined in section 4.7. Although the V/UV decisions are made every 8mS, the blocksize used in the frequency mapping algorithm corresponds to a duration of 16mS (i.e. N = 256 samples and F_s = 16kHz). Thus if within the 16mS interval either the first or the second 8mS block is deemed to be 'voiced', then the decision to apply spectral compression is identical to the decision made during the preceding 16mS block (refer to equations 4.8-4.10).

6.4.3.3 Compression Law

Hitherto, a piecewise linear law has been assumed for the frequency mapping compression characteristic which was suited to the voiced speech upon which the system was originally tested. It was hence found that the system performed better if the lower frequencies of the voiced speech signal were preserved at the expense of a greater compression ratio applied to the higher frequencies. For wideband unvoiced speech it is perhaps the higher frequency band which has the greater significance. By resorting to the wideband spectrum of a prominent unvoiced sound such as that for the /s/ utterance shown in figure 4.3(b), the mapping law c = 4, 1,20 corresponding to the frequency bands 0 to 3.5, 3.5 to 5.5 and 5.5 to 8kHz as illustrated in figure 6.22 was used. Although a different compression law can be used for each unvoiced sound, a simple approach is to use this fixed law which is suitable for the utterance /s/ (a sound which is most seriously affected by the conventional 300 to 3400Hz bandlimiting) and apply the law to all unvoiced sounds in our data sequence.

6.4.3.4 Results and Conclusions

The time waveform of "si" from the word "sister" is shown in figure 6.23(a) for 300 to 7600Hz speech.

When the speech signal is bandlimited between approximately 300 and 3400Hz the /s/ in "si" is significantly attenuated and distorted as illustrated in figure 6.23(b), whereas the /I/ is substantially preserved. The comparison of the waveforms of figure 6.23(a) and figure 6.23(b) are of course the same as those carried out for the VUBS in section 6.3 but are repeated here for the reader's convenience. The frequency mapping approach is directed at unvoiced speech with the result that the 300 to 3400Hz mapped speech takes the form of figure 6.23(c). Observe that the mapped /s/ has a similar envelope to the original /s/, and that its zero-crossing rate is higher than that of the voiced /I/. The de-mapped time waveform for the "si" is shown in figure 6.23(d); notice that the waveform of the /s/ resembles the original wideband /s/ waveform far more closely than the bandlimited version of figure 6.23(b).

The spectrum of the de-mapped /s/ sound (under the mapping law of figure 6.22) is shown in figure 6.24(c). When this spectrum is compared with a 300 to 3400Hz bandlimited version of figure 6.24(a) it can be appreciated that the spectral peak in the /s/ signal has been essentially preserved by the frequency mapping process which utilises the transmission bandwidth shown by the spectrum of figure 6.24(b). Figure 6.25 shows the spectrograms corresponding to 300 to 7600Hz speech and the recovered 300 to 7600Hz speech using the frequency mapping process.

Informal listening experiences confirmed that by frequency-mapping the unvoiced sounds, improvements can be achievable in both intelligibility and quality.

In conclusion here, we argue that enhancement of speech intelligibility and quality can be attained by bandpass filtering voiced speech from 300 to 3400Hz, and frequency mapping unvoiced speech occupying the frequency band 300 to 7600Hz into the 300 to 3400Hz band. Although all unvoiced sounds are frequency mapped, the law advocated is particularly appropriate for /s/, a sound that makes a significant contribution to the intelligibility of speech.

When comparing the frequency mapping process with the VUBS method, the informal subjective listening experiences suggests that the frequency mapping recovered signal is preferable to the VUBS signal, particularly for the /s/ sound. This may be expected since the spectrum of the frequency de-mapped /s/ signal is a closer resemblence to the original than the 3 to 6kHz bandlimited portion generated by the VUBS method. The penalty, however, is that the frequency mapping system complexity far exceeds that of the VUBS approach.

Having now developed the scheme of wideband frequency mapping we now apply further refinements to the process in order to gain even greater improvements to the recovered speech signal.

In order to aim to meet this goal, we develop an adaptive algorithm of frequency mapping which will be detailed in the following section.

6.4.3.5 Adaptive Frequency Mapping

The frequency mapping used up to now was designed for the /s/ utterance (in particular the /s/ spectrum) which is the phoneme that has the highest relative frequency of occurrence of any of the unvoiced fricative consonants (ref 3). Other unvoiced sounds, particularly the unvoiced stop consonants, are not reproduced with sufficient accuracy although their perception was nevertheless enhanced. In order to improve upon this deficiency, an adaptive frequency mapping scheme is advocated which extends the frequency mapping scheme to improve the reception of unvoiced sounds, not just /s/. Specifically, we arrange to preserve the spectra of unvoiced speech in the vicinity of the energy peaks of the unvoiced spectra. In such regions, the mapping compression ratio, c, is set to unity. High values of c are therefore employed in the spectral troughs where frequency de-mapping errors are considered to be perceptually less important. Our intention now is to derive a set of mapping laws where the unvoiced spectrum will be simultaneously frequency mapped by each law, and the law which provides the best match to the spectrum will be the one employed. Two questions arise; how to select the set of fixed mapping laws and having selected these laws, how to decide which law is the most appropriate for a particular unvoiced utterance.

a) The Mapping Laws

Some spectral density functions of the unvoiced fricative consonants /s/, /f/, /θ/ and /ʃ/ and the stop consonants /t/, /k/ and /d/ are shown in figures 4.3 and 4.5; the fricative consonant /h/ plus the

dipthongs /tʃ/ and /dz/ are also examined to give an approximate indication of their spectral structure, as shown in figure 6.26. Table 6.1 displays the frequency compression factors and their corresponding frequency bands that were selected to enhance the perceptual quality of the recovered speech signal. Also in Table 6.1 is the relative frequency of occurrence of the unvoiced sounds, emphasising the importance of the phonomes /s/ and /t/.

The reasoning concerning the choice of the mapping law for /s/ has already been described. The fricative /f/ appears to have almost a flat spectrum over telephonic frequency band, while above 3400Hz the spectrum falls at about 5dB/octave. The frequencies up to 6800Hz were frequency mapped with c=2 and the frequencies above 6800Hz were rejected. The spectral density function for /ʃ/ indicates that frequencies in the range 2500 to 5500Hz should be preserved. Listening tests revealed, however, that using c=1 over the band 300 to 3400Hz, i.e. simple bandpass filtering, gave the best perceptual results. When mapping was employed for frequencies above 3400Hz, the interpolation noise was found to be perceptually more annoying than the muffling imposed by the lowpass filtering.

The stop consonants /t/, /k/ and /d/ tended to exhibit two or three significant formants. The lower formant is associated with the initial step and the large amplitude excursion in the time waveform. The upper formants are indicative of the rapid excursions in the trailing of the time signal (especially the /t/ in figure 4.2(a)). Accordingly, the spectrum is divided up into four zones with c having values of 1 or 4, as documented in Table 6.1.

Phoneme	Consonant	Relative frequency of occurrence (%) (ref 3)	Mapping-Law	
			c	Frequency band(Hz)
/s/ sip	Fricative	4.5	4 1 4	0 - 3500 3500 - 5500 5500 - 8000
/f/ father	Fricative	1.84	3	0 - 6800
/θ/ thin	Fricative	0.37	0	6800 - 8000
/ʃ/ shift /h/ he	Fricative Fricative	0.82 1.81	1	300 - 3400
/t/ to	Stop	7.13	1 4	300 - 1200 1200 - 4000
/k/ key	Stop	2.71	1 4	4000 - 5000 5000 - 8000
/d/ day	Stop	4.31		
/tʃ/ chew	Dipthong	0.52	3	0 - 6800
/dz/ jar	Dipthong	0.44	0	6800 - 8000

Table 6.1 Mapping Laws Associated with Different Unvoiced Sounds

Conversational speech produces many variants to the spectral density functions shown in figures 4.3, 4.5 and 6.26, and there are other sound such as voiced stop consonants and voiced fricative consonants etc which might be subjected to frequency mapping. What we have established is the set of mapping laws given in table 6.1 and we require the system to select the most appropriate law for the sound

to be frequency mapped. Thus we mean that the frequency mapping is adaptive, in the sense that it selects the best mapping law at its disposal for the unvoiced sound that it is processing. Often different laws may be sequentially employed during a particular unvoiced sound when the adaptive mapping law procedure is invoked. Three stages are involved in selecting the mapping law. The initial one is of course to decide whether to frequency map or to lowpass filter the speech samples. If frequency mapping is required then the block of speech samples is mapped by all of the mapping laws that are available. Finally, the law that gives the best fit must be identified. The first and last stages involve the making of decisions, and we now describe how these decisions are made.

b) When to use Adaptive Frequency Mapping

As with the fixed mapping strategy, the wideband input speech signal is again separated into block of 8mS each and identified as voiced or unvoiced according to the first shift autocorrelation method outlined in section 4.7. The V/UV decision is made every 8mS although the frequency mapping procedure operates on 16mS of speech. Again, if within the 16mS interval either the first or the second 8mS block is deemed to be 'voiced' (i.e. when there is a transition between voiced and unvoiced speech in adjacent 8mS blocks), then the decision is the same as for the previous 16mS block. The decision criterion is undoubtedly simple and is therefore likely on occasions to produce an incorrect decision. This is however not a disaster; if 7600Hz voiced speech is inadvertently mapped, or if a fricative is bandlimited, the

recovered speech is little different from 3400Hz. We therefore retain the simple decision criterion as it is easy to implement and in general it makes the correct decision. However when it does fail then no catastrophe occurs when examining the decision on the 10-word sequence of section 4.2.

c) Selecting the Mapping Law

Once the decision is made that frequency mapping is to be used, the 16mS block of speech is mapped by each of the four mapping laws given in table 6.1. The mapped signals are stored in buffers 1, 2, 3 and 4 until they are ready to be inverse frequency mapped, see figure 6.27. Each law is denoted L_1, L_2, L_3 and L_4 according to table 6.2. The first 'mapping law' is simply is bandpass filter followed by a Discrete Fourier transformer which corresponds to linear compression from 300 to 3400Hz and a rejection of frequency components above 3400Hz. The other mapping laws L_2 to L_4 correspond to actual spectral compression. Commencing from the first buffer (Buffer 1), the bandlimited or frequency mapped speech is recovered by the local decoder to give the spectral speech components $\{\hat{X}_{k,j}\}$ where k = 0, 1, ... N/2 and j = 1, 2, 3 or 4 corresponding to the frequency mapping and de-mapping law used. If j= 1 then of course bandlimiting is used and the spectral sequence $\{\hat{X}_{k,1}\}$ will not be decoded. In the environment of a noise-free transmission channel between the transmitter and receiver, the output of the local decoder represents the recovered speech spectrum that would be produced at the receiver output for each mapping law

LAW	c	FREQUENCY BAND (Hz)
L_1	1	300 - 3400
L_2	4 1 4	0 - 3500 3500 - 5500 5500 - 8000
L_3	2 0	0 - 6800 6800 - 8000
L_4	1 4 1 4	300 - 1200 1200 - 4000 4000 - 5000 5000 - 8000

Table 6.2 Notation for the four compression characteristics for the adaptive frequency mapping strategy

used. What we require is to use the mapping law which will generate the minimum interpolation noise possible from the four types of compression characteristics available at the transmitter. This aim is achieved by comparing the locally decoded spectrum with the original wideband spectrum and finding the maximum spectral signal to noise ratio, SP-SNR. The spectral SNR is computed according to

$$\text{SP-SNR} = 10 \log_{10} \left\{ \frac{\sum_{k=0}^{N/2} |X_{k,j}|^2}{\sum_{k=0}^{N/2} \left[\frac{|\hat{X}_{k,j}|}{\sigma_o} - \frac{|X_{k,j}|}{\sigma_I} \right]^2} \right\} \qquad (6.23)$$

where $X_{k,j}$ and $\hat{X}_{k,j}$ are the k^{th} frequency component in the input speech and locally frequency de-mapped signal from law L_j respectively; σ_I and σ_o are the rms values of the input and frequency de-mapped signals; and N = 256.

To find the SP-SNR_j for j = 1 to 4, the switches S_3, S_4 and S_5 move synchronously and in sequence from position 1 to position 4 where for each position the SN-SNR_j is computed. The law with the largest SP-SNR is selected for frequency mapping. The contents of the buffer associated with this law proceed via switches S_3 and S_6 to be inverse Discrete Fourier Transformed by the IFFT to yield the frequency mapped time sequence $\{y_i\}$. This sequence, and the analogue side information to instruct the receiver which mapping law was used, is then multiplexed and transmitted with the bandpass voiced speech signal.

Observe in figure 6.27 that there is a direct connection between switches S_3 and S_4. However, if the characteristics of the transmission channel are known this link can be broken, the contents of the buffer are inverse Fourier transformed via S_3 and S_6, and the resulting time sequence subjected to a model of the channel. The output from the model is Fourier transformed by the FFT and applied to switch S_4. By this means the selection of the law will be appropriate for both the speech signal and the transmission channel. In our experiments we used an FIR lowpass filter for the channel, for which the arrangement of the adaptive frequency mapping system shown in figure 6.27 is applicable.

The receiver de-multiplexes the received signal, and from it determines if frequency mapping was used and if so, the law employed. The speech block is then Fourier transformed by the FFT and the appropriate frequency de-mapping law is applied. A uniformly distributed random phase between $\pm 180^{o}$ is again given to the estimated frequency components where interpolation had been used in the frequency de-mapping procedure. This is the same operation used in the fixed frequency mapping case (equation 6.22). The sequence is then inverse Fourier transformed by the IFFT to recover the wideband output speech signal.

Side-Information Multiplexing

For the experiments conducted in order to simulate the operation of the adaptive frequency mapping procedure, space division multiplexing was effectively used to convey the side information from the transmitter to the receiver. This meant using a noiseless auxiliary 'wire' in parallel to the 300 to 3400Hz analogue channel which does indeed represent an ideal situation. In a more practical arrangement, the information regarding which of the 4 laws that are to be used on each of the 16mS speech blocks may be transmitted in the form of a 2-bit FSK pilot tone embedded in a narrow frequency slot, say 2600Hz, as was suggested with the HFR method in section 5.4.1. This procedure was not, however, tested by our simulation.

d) Results and Discussion

The system as shown in figure 6.27 was tested on the 10-word sequence obtained from section 4.2. For utterances containing just /s/'s and voiced speech, the adaptive frequency mapping system behaved as the fixed mapping method which just employed Law L_2 and bandpass filtering only. During the transitions between the /s/ utterance and voiced speech (and vice-versa), there are instances when a different mapping law was selected and in so doing offered a better spectral blend between the voiced and unvoiced transitions than the single change from fixed frequency mapping to bandpass filtering between one block and the next.

This result was demonstrated by the SP-SNR values obtained for the utterance /s/, /ka/ from "S K Harvey" as shown in table 6.3 and figure 6.28. During the sustained /s/ sound, law L_2 is selected for every block except at the last block where law L_4 appears to be the better choice. For the first block of the silence, law L_3 is preferred possibly because there is still a small amount of the /s/ signal still present.

With regard to the /k/ utterance, the leading edge which contains a large step function was preferentially bandpass filtered whilst the second block containing the trailing edge of the /k/ and hence the higher frequencies was mapped by law L_4. Under fixed frequency mapping conditions, law L_2 would have been selected for the leading and trailing positions of the /k/ resulting in the reduction of spectral SNR of around 8 and 6dB respectively. During the voiced /a/, the V/UV switch changed to 'V' position and none of the laws were invoked.

Block Number	Phoneme	SP-SNR(dB)				Law Selected
		L_1	L_2	L_3	L_4	
1	/s/	-18.2	15.7	5.87	14.2	L_2
2	/s/	-22.1	15.9	6.07	12.3	L_2
3	/s/	-13.3	15.5	6.49	6.97	L_2
4	/s/	-6.50	10.8	7.51	9.72	L_2
5	/s/	-2.11	6.50	6.44	7.03	L_4
6	inter-phoneme	-3.06	9.33	12.0	5.13	L_3
7	silence	12.3	7.13	8.99	11.6	L_1
8	/k/	14.9	6.38	9.14	10.7	L_1
9	/k/	10.5	6.77	7.71	12.5	L_4
10	inter-phoneme	17.1	0.24	7.03	6.79	L_1
11	silence	17.2	4.12	4.93	8.64	L_1
12	/a/					Voiced
13	/a/					Voiced
14	/a/					Voiced
15	/a/					Voiced
16	/a/					Voiced

Table 6.3 Variation of spectral SNR against block size for the four mapping laws and the adaptive mapping selection.

The waveforms in figure 6.29 also illustrate the advantage of the adaptive frequency mapping process. Figure 6.29(a) shows the time waveform of the trailing position of the /k/ utterance contained in "S K Harvey". The power spectral density function and phase spectrum for the 7.6 kHz speech are also displayed in figure 6.29, sub-figures (b) and (c) respectively. The bandpass spectra associated with law L_1 are not displayed as they are those of the wideband speech from 0.3 to 3.4 kHz. Sub-figures (d) to (f) show the recovered wideband time signal, its power spectrum and phase spectrum, when law L_2 is used. The third and final rows corresponding to sub-figures (g) to (l) show the time waveforms and spectral responses when frequency de-mapping is performed using laws L_3 and L_4 respectively. The centre column in figure 6.29

displays the power spectra, here the wideband speech spectrum shows large peaks in the vicinity of 1.5 to 2.5 kHz and 4 to 5 kHz. Law L_1 retains the first peak and rejects the second while law L_2 retains the second peak and makes a crude representation of the spectrum in the region of the first peak. Frequency compression of $c = 3$ is applied over the range of both peaks when law L_3 is used whereas law L_4 preserves the second peak and does as well as L_3 over the first peak. Our SP-SNR criterion of equation (6.23) is not solely concerned with how well the spectrum in the vicinity of the peak is reproduced but with the overall spectral match. According to table 6.3 (block 9), the double peak mapping law was selected. Thus the wideband unvoiced speech signal whose time waveform, power and phase spectra were as shown in row 1 were reproduced by a signal whose properties are displayed in the last row. In this example frequency mapping was selected by a narrow margin over the bandpass filter case, law L_1.

The mapping laws selected in the adaptive mapping process for some unvoiced sounds are displayed in figure 6.30, where the block numbers are relative to the beginning of the utterance. Sub-figure (a) shows successive blocks commencing with the second /s/ in "sister" and ending with the end of /t/. The /s/ extends over the first seven blocks where law L_2 is favoured except at the very end of the sound. Blocks 8 and 9 are for inter phoneme silence intervals, while the last two blocks apply for /t/. The sounds /f/ in "fist" and in "father", are shown in (b) and (c), respectively. For the latter case, the first 10 blocks are applicable to /f/, blocks 11 and 12 are inter phoneme silence, while blocks 13 to 15

are the lead-in to the /a/ sound. The variation of /k/ in "talk" is displayed in (d). In (e), blocks 1 to 11, 12 to 13, and 14 to 18 relate to the /s/, inter silence, and /k/ in "S K Harvey" respectively. The /ʃ/ phoneme in "shift" applies to (f) and /t/ in "talk" is shown in (g). The variation of the laws from the beginning of /f/ to the end of /t/ in "shift" is displayed in (h), where blocks 10 to 12 are effectively inter silence. The laws for /θ/ and /k/ in "thick" are presented in (i) and (j) respectively. The adaptive mapping for /s/ through to /t/ in "fist" is given in (k) where blocks 13 to 15 relate to inter silence. The /s/ followed by /t/ applies to both (a) and (k) with a similar selection of mapping laws. Sub-figures (l) and (m) both apply to "spent". The /s/ and /p/ are in (ℓ) where block 3 and 4 relate to the intra silence periods, and the response to (m) relates to /t/.

From these samples we see that some sounds use almost only one law, e.g. /s/ and /t/ which rely on laws L_2 and L_4 respectively. In general, two or three mapping laws may be used over the duration of an unvoiced sound. During inter phoneme silences, law L_4 tends to be preferred.

The spectrograms for the utterances "sister", "father", "S K Harvey", "shift", "thick", "fist" and "talk" are shown in figure 6.31; fig. 6.31 row 1 shows the spectrogram of the original signal. It is perhaps not so clear from this display that the /k/ utterance of the wideband signal needs any spectral compression at all as it would seem that a 300 to 3400Hz channel would be adequate to transmit it. This may be because the spectral weighting used for

the spectrograph display was insufficient to detect the third formant in the /k/ signal. If a higher spectral weighting emphasis or greater marking depth had been used then the spectra of the voiced speech would have been abnormally biased towards the high frequency region.

The performance of the adaptive frequency mapping system of figure 6.27 was further tested by informal subjective listening experience of the output speech signal from our 10-word input sequence. These tests clearly indicated far better speech quality than the 300 to 3400Hz bandlimited speech signal and the adaptive version of frequency mapping achieved a superior performance than the fixed mapping case. The output signal from the adaptive strategy indicated better results for all of the unvoiced utterances including the transitions between voiced and unvoiced speech and vice versa. Even the /st/ from "fist" showed an improved and more natural sound which in the past has been notoriously difficult to achieve.

In conclusion, then, we can say that the adaptive frequency mapping algorithm had consistently shown better results than by using the fixed mapping law. However, the price paid is the substantial increase in complexity, especially at the transmitter terminal. This is also reflected in the increase in the computation time required to simulate the adaptive system.

In the future, it may be envisaged that a greater flexibility can be incorporated into the adaptive frequency mapping system by including more mapping laws available for spectral compression based upon

statistically averaged spectra from much more data. In our example for the /k/ signal of figure 6.29, the /k/-spectrum would probably be suited to a three-formant preservation mapping law rather than just a two-formant law. It remains, however, to be seen how much more improvement in output speech quality can be obtained by pursuing these directives.

Having now discussed in detail a fairly proficient spectral compression scheme, we will now go on to another process of bandwidth compression of perhaps a simpler nature. This process will still be based in the frequency domain and is the topic of our next section.

6.4.4 Bandwidth Compression in the Frequency Domain using an Artificial Phase Function

In the previous section, we found that the frequency mapping algorithm operated in the Fourier transform domain, that is, using the general structure of figure 6.7, the speech signal was manipulated by way of the magnitude and phase of its frequency components. The frequency mapping algorithm dealt with the magnitude and phase components in parallel, i.e. whatever the processing was applied to a specific magnitude component, it is also applied to the phase of that component. The only divergence of the treatment for the magnitude and phase components occurred at the final stage of the receiver where the magnitude of the estimated frequency components in the frequency de-mapping procedure was interpolated from neighbouring known components and the phase was assigned a random value between $\pm 180^{o}$.

In this section, we consider separating the magnitude and phase spectra right at the earliest opportunity i.e. after the DFT operation in the transmitter terminal. Spectral compression can now be applied to the magnitude and phase as separate entities before ultimately recombining them at the receiver terminal to reconstruct the speech signal. It does perhaps seem sensible to treat the magnitude and phase spectra separately in this way because if the magnitude and phase spectra, shown in figure 6.29(b) and (c) for the /k/ in "S K Harvey", are compared it can be seen that they are structurally very different. In particular, the adaptive frequency mapping scheme of the previous section was modelled around the magnitude spectral distribution whereas the model completely disregards the phase characteristics.

What is proposed now is a relatively simple way to keep the magnitude spectrum completely intact while applying the bandwidth compression using phase-processing only. In general, phase processing offers the possibility of frequency compression by transmitting only the magnitude of each frequency component and not its phase. A theoretical maximum of 2:1 compression (halving the bandwidth) should be attainable by this process. As described in section 2.1.2, the notion of phase-rejection is commensurate with the lack of phase sensitivity of the human perception mechanism (ref 33). We have also noted in section 3.4.1.5 that earlier research(ref 56) involved DFT analysis of successive segments of speech and discarding the phase spectrum but it resulted in periodic clicks in the recovered speech signal. This is not surprising since if one refers back to the investigation of the IDFT

(figure 6.20) then for low values of time index, i, e.g. i = 0, the complex vectors are facing the same direction, so for the first time-sample, the frequency coefficients will be summed coherently giving a maximum at the start of the block. The same argument applies at the end of each processing block.

In reference (56) we saw from Chapter 3 that attempts were made to overcome the problem of the block synchronous peaks by gradually re-introducing the phase spectrum with minimal additional increase in channel capacity. Another method to reduce the unwanted periodic components employed pitch synchronous DFT processing of the speech signal where the block-edge distortion produced by the total or partial rejection of the phase was masked by the high amplitude component of the pitch excitation present in the voiced speech.

In accordance with our approach to overcome the block synchronous distortion produced by the frequency mapping schemes, we will attempt to apply total phase rejection at the transmitter and introduce a random phase spectrum at the receiver terminal. If one considers the phase spectrum of figure 6.29(c), it can be seen that the original phase spectrum already appears very random-like in structure, so, even from that standpoint our proposed algorithm can still be loosely deemed a signal-specific notion, just as the frequency mapping schemes were magnitude spectrum specific (although rather more stringent).

6.4.4.1 Operation of the Phase Processor

Our initial system is basically that of figure 6.7(a) and (b) using the aforementioned phase processing to afford the required spectral compression and expansion. The phase processing system which we will term the random phase processor becomes the arrangement as shown in figure 6.32(a) and 6.32(b).

The incoming signal, x(t), is segmented into contiguous blocks of samples and Discrete Fourier Transformed by the DFT where the phase spectrum is set to zero. The magnitude frequency coefficients corresponding to the upper band of 3400 to 6500Hz are then transferred into the lower band of the phase array corresponding to the 300 to 3400Hz frequency range. The new phase coefficients are then assigned to a pseudo-random sign (which is known to the receiver) before being inverse Discrete Fourier Transformed by the IDFT. We will see later that the reason for using random sign assignment to the phase values is to reduce the peaks at the block edges of the transmitted signal.

The band compressed signal is then sent to the 300 to 3400Hz channel which emerges at the receiver as shown in figure 6.32(b). The received signal is 'cleaned-up' by a further 300 to 3400Hz bandwidth filter before being again divided into blocks of samples. The blocks are transformed into the frequency domain by the DFT where the phase of the frequency coefficients are re-allocated to their original sign (i.e. de-randomised). The 300 to 3400Hz phase spectrum is then shifted back to occupy the 3400 to 6500Hz magnitude

spectrum. The expanded magnitude spectrum is combined with an arbitrary phase spectrum containing uniformly distributed random values between $\pm 180^{0}$ and extending between 300 and 6500Hz whence the resultant signal is transformed back into the time domain by the IDFT to reconstruct the wideband speech signal.

Having now discussed the general operation of the transmitter and receiver, we will elucidate each function in more detail.

a) The transmitter

Here, the signal is divided into contiguous blocks of N samples each and the sequence $\{x_i\}$ is transformed into the frequency domain. The DFT of which is

$$X_k = \frac{1}{N} \sum_{i=0}^{N-1} x_i \exp\left(-j \frac{2\pi k i}{N}\right) \qquad (6.24)$$

$$k = 0,1,\ldots,N/2$$

This is of course the same expression as that used in the DFT stage of the frequency mapping process (equation 6.5). Accordingly, the magnitude of X_k is

$$|X_k| = \left\{Re^2(X_k) + Im^2(X_k)\right\}^{1/2}$$

and the phase of X_k is

$$\theta_k = \frac{180}{\pi} \mathrm{Tan}^{-1} \left[\frac{Im(X_k)}{Re(X_k)}\right] \qquad (6.25)$$

$$k = 0,1,\ldots,N/2$$

We now separate the magnitude and phase components by setting all the phase values θ_k to zero, i.e.

$$\theta_k = 0 \quad \text{for} \quad k = 0,1,\ldots,N/2 \tag{6.26}$$

Restating equation (5.4), the frequency resolution (frequency separation) of each DFT component, Δf is equal to F_s/N where F_s is the sampling frequency. Therefore the nearest value of k which corresponds to a frequency 'f' kHz (if F_s is in kHz) is

$$k_f = \text{NINT}(f/\Delta f) \tag{6.27}$$

Where NINT(.) corresponds to the nearest integer value of (.).

The values of $|X_k|$ corresponding to zero to 3400Hz are retained, i.e.

$$|Y_k| = |X_k| \qquad k = 0,1,\ldots,k_{3.4} \tag{6.28}$$

For the initial experiment using this process, the values of corresponding to 3400 to 6500Hz were transferred into the phase array according to:

$$\phi_k = |X_{(k+k_{3.4})}| \tag{6.29}$$

Having completed the magnitude to phase array transfer the $\{|Y_k|\}$ and $\{\phi_k\}$ sequences were initially converted into real and imaginary

sequences without any sign randomising such that the components were now:

$$Re(Y_k) = |Y_k| \text{ Cos } \phi_k$$

and
$$Im(Y_k) = |Y_k| \text{ Sin } \phi_k \qquad (6.30)$$

$$k = 0,1,\ldots,N/2$$

To extend the number of components in the real and imaginary sequences to N, we again invoke the complex conjugate property of spectral components of wholly real time functions, namely

$$Y_{N-k} = Y_k^* \qquad (6.31)$$

$$k = 0,1,\ldots,N/2$$

where as before, the raised * above the symbol implies its complex conjugate.

$$Y_k = Re(Y_k) + j\ Im(Y_k) \qquad (6.32)$$

The inverse DFT is now applied to give the time sequence $\{y_n\}$ where

$$y_i = \sum_{k=0}^{N-1} Y_k \exp\left(j \frac{2\pi k i}{N}\right) \qquad (6.33)$$

$$i = 0,1,\ldots,N-1$$

$\{y_i\}$ is the time sequence whose frequency components occupy a lower frequency band compared to those of $\{x_i\}$. Specifically, the highest frequency component in $\{y_i\}$ is 3400Hz.

When this experiment was performed, the resulting time sequence, $\{y_i\}$ was found to peak at the block edges as shown by the waveform of figure (6.33). The reason for this phenomenon is considered to be due to the fact that the phase sequence $\{\phi_k\}$ in equation (6.29) has values that typically range between 0 and 35^o only.

First, let us hypothetically allow $\phi_k = 0$ for all $k = 0, 1, \ldots N/2$ which for the moment is used instead of equation (6.29). We then re-apply the IDFT viz:

$$y_i = \sum_{k=0}^{N-1} |Y_k| \exp(j\,\phi_k).\exp\left(j\,\frac{2\pi k i}{N}\right) \qquad (6.34)$$

The initial value of y_i is:

$$y_i = \sum_{k=0}^{N-1} |Y_k| \qquad i = 0,1,\ldots,N-1$$

Since the amplitude spectrum will have all positive values, y_0 will be the maximum value of the function y_i. Similarly, for $i = N-1$ (i.e. the last sample in the block), the exponential term will be large, making the value of the summation large. This can also be explained by our matrix demonstration of figure 6.20. The shape of the resulting processed block will now indeed peak at the block edges. This peak will correspond to the join between successive segments and will thus result in periodic peaks in the transmitted waveform, $\{y_i\}$.

As far as listening tests are concerned, the above phenomenon should have no consequence, provided that the receiver can successfully apply the inverse process to the transmitter and so remove the block synchronous peaks. An ideal analogue transmission signal containing block synchronous peaks will not deteriorate the overall perfomance of the system unless it is required to digitise the transmitted analogue signal. In such a case, the large spikes in the waveform to be encoded would pose a difficult task for the digital coder and in general would lower its performance. Furthermore, an encoder having a constant SNR for a wide range of input power values will produce an increase in quantization noise during the encoding of these peaks and while the inverse post processor should remove the block synchronous distortion, the quantization noise will remain periodically reinforced in the time waveform. It is therefore desirable to reduce the level of the block synchronous peaks in the analogue transmitted waveform. This aim was found to be achieved by reverting back to equation (6.29) and replacing it by:

$$\phi_k = a\ |X_{(k+k_{3.4})}| - 180^0 \qquad (6.35)$$

where $k = k_{0.3}, k_{0.3} + 1, \ldots, k_{3.4}$

Also $\phi_k = 0$ for $k = (k_{3.4} + 1), (k_{3.4} + 2), \ldots, N/2$

The value of 'a' in equation (6.35) is constant for each processing block (i.e. for all k), and is chosen such that

$$a = 360^0 / |X|_{max}$$

where $|X|_{max}$ is the maximum value of $|X_k|$ for the range $k_{3.4}$ to $k_{6.5}$. This allows for the values of the phase, ϕ_k, to be 'filled' by $|X_k|$ and effectively to spread the ϕ_k values between $\pm 180^o$.

The value of 'a' may be transmitted to the receiver once for each processing block as side information. For our data, the constant value of a = 9.0 was used for every processing block.

When the new compressed spectrum was inverse Fourier transformed by equations (6.30 to 6.33), the block synchronous spikes were still apparent although now they were reduced compared to the case when equation (6.29) was used to form ϕ_k.

We have already noted from the frequency de-mapping procedure in section 6.4.3.2 that if a 'randomness' were to be introduced into the phase spectrum the effect would be to smooth the peaks from the block edges in the time waveform. We can thus assign a random sign to ϕ_k to decorrelate the phase values occupied by the upper band $\{|X_k|\}$ from equation (6.35) and so reduce the block synchronous spikes that occur in the inverse transformed time sequence for transmission. The sign of the values of the phase, ϕ_k, are randomised according to

$$\lambda_k = (-1)^m . \phi_k \qquad (6.36)$$

$$\text{for} \quad k = 0,1,\ldots,k_{3.4}$$

where m = INT $(R_1-0.5)+2$ and INT(.) is the round down integer value of '.' while R_1 is a uniformly distributed pseudo-randomly

generated source having values between 0 and 1. R_1 is also duplicated at the receiver terminal but it need not be the same sequence for each block.

Now, the IDFT is again invoked such that

$$\begin{aligned} Re(Y_k) &= |Y_k| \ Cos(\lambda_k) \\ Im(Y_k) &= |Y_k| \ Sin(\lambda_k) \end{aligned} \qquad (6.37)$$

$$k = 0,1,\ldots,N/2$$

with $$Y_{N-k} = Y_k^* \qquad (6.38)$$

$$k = 0,1,\ldots,N/2$$

and $$y_i = \sum_{k=0}^{N-1} Y_k \exp\left(j \frac{2\pi ki}{N}\right) \qquad (6.39)$$

The sequence $\{y_i\}$ is now transmitted to the receiver via the 300 to 3400Hz bandlimited channel as shown in figure 6.32.

b) The Receiver

The receiver operates on the filtered version of $\{y_i\}$ sequence of samples now denoted by $\{\tilde{y}_i\}$. The block of N samples are again converted into the frequency domain by the DFT process:

$$\tilde{Y}_\ell = \frac{1}{N} \sum_{i=0}^{N-1} \tilde{y}_i \exp\left(-j \frac{2\pi ki}{N}\right) \qquad (6.40)$$

$$\ell = 0,1,\ldots,N/2$$

and the polar representation is obtained as

$$|Y_\ell| = \left\{Re^2(\tilde{Y}_\ell) + Im^2(\tilde{Y}_\ell)\right\}^{1/2} \qquad (6.41)$$

and

$$\tilde{\lambda}_\ell = \frac{180}{\pi} \text{Tan} \left[\frac{Im(\tilde{Y}_\ell)}{Re(\tilde{Y}_\ell)}\right] \qquad (6.42)$$

$$\ell = 0,1,\ldots,N/2$$

The magnitude spectrum now has to be restored by transferring the 300 to 3400Hz range of the phase spectrum into the 3400 to 6500Hz region of the magnitude spectrum. However, before doing this transfer, the received phase spectrum has to be recovered by 'de-randomising' the sign of the phase coefficients, viz:

$$\tilde{\phi}_k = (-1)^m \tilde{\lambda}_k \qquad (6.43)$$

$$k = 0,1,\ldots,N/2$$

and as with equation (6.36)

$$m = INT(R_1 - 0.5) + 2$$

The phase spectrum is rescaled, such that

$$|\tilde{X}_{(k+k_{3.4})}| = (\phi_k + 180)/a \qquad (6.44)$$

$$k = k_{0.3},(k_{0.3}+1),\ldots,k_{3.4}$$

where 'a' is as previously defined for equation (6.35). The values of $\{|\tilde{X}_k|\}$ in equation (6.44) are added to the lower band:

$$|\tilde{X}_k| = |\tilde{Y}_k| \qquad k = k_{0.3},(k_{0.3}+1),\ldots,k_{3.4} \tag{6.45}$$

Equations (6.44) and (6.45) form the full magnitude spectrum of the recovered signal up to 6500Hz.

Finally, the entire phase spectrum is replaced by a uniformly distributed randomly generated source having values between $\pm 180^{o}$. The phase is now given by:

$$\tilde{\theta}_k = 360^{o}(R_2 - 0.5) \tag{6.46}$$

where R_2 is a uniformly distributed randomly generated number between 0 and 1 which is not necessarily the same as the sequence R_1.

The sequences $\{|\tilde{X}_k|\}$ and $\{\tilde{\theta}_k\}$ are converted into real and imaginary sequences, i.e.:

$$Re(\tilde{X}_k) = |\tilde{X}_k| \; Cos(\tilde{\theta}_k)$$

and

$$Im(\tilde{X}_k) = |\tilde{X}_k| \; Sin(\tilde{\theta}_k) \tag{6.47}$$

$$\text{for} \quad k = 0,1,\ldots,N/2$$

and using

$$\tilde{X}_{N-k} = X_k^* \quad \text{for} \quad k = 0,1,\ldots,N/2 \tag{6.48}$$

the speech sequence, $\{\tilde{x}_i\}$, is recovered according to

$$\tilde{x}_i = \sum_{k=0}^{N-1} X_k \exp\left(j \frac{2\pi k i}{N}\right) \tag{6.49}$$

for $i = 0,1,\ldots,N-1$

The resulting sequence forms the recovered speech signal.

c) The Channel

The channel used in the simulation was modelled by a finite impulse response (FIR) filter of bandlimits 300 to 3400Hz. This type of channel is chosen because it is required to possess a linear phase characteristic such that the phase spectrum of the transmitted signal $\{y_i\}$ can be faithfully recovered after compensating for the channel delay, i.e. the recovered sequence $\{\tilde{\lambda}_k\}$ should be a close approximation to the sequence $\{\lambda_k\}$. In practice, the group delay characteristic must be precisely constant such that synchronisation between the transmitter and receiver can be preserved. Experimentally we found that if the channel had caused a delay error by just ± 1 sampling interval then it was not possible for the receiver to recover $|\tilde{X}_k|$ from $\tilde{\lambda}_k$ using equations (6.43) and (6.44). This problem can be obviated by using a digital channel in place of the FIR filter.

6.4.4.2 Results

The results, obtained by computer simulation, are based upon informal subjective listening tests again on the 10-word speech sequence. Initially no voiced/unvoiced switching was used by the random-phase processing system such that the entire speech sequence was phase-processed. The block length N for the compression algorithm was taken as 256 samples corresponding to a 16mS duration as the magnitude spectrum did not appear to change appreciably within that interval.

The waveform for the utterance /I/ in "sister" is shown in figure 6.34 which shows a large amount of distortion as compared to an original waveform of an /I/ sound from say, figure 4.17(a). The distortion was also evident upon subjective listening tests which showed a large amount of periodic distortion (hoarseness) in the processed voiced speech. This does indicate a rather unsuccessful outcome by the random-phase process.

6.4.4.3 Application of the Random Phase Processor to /s/ Sounds

It could be discerned that particularly for the /s/ fricative sounds, the results appeared to yield a better quality signal than the 300 to 3400Hz bandlimited version. It was at this stage decided to apply the random phase process just to the /s/ sounds and leave the remaining signal unprocessed. The phase processor now fits into our general framework as shown in figure 6.3; the processor of figure 6.32(a) becomes the transmitter of figure 6.3(a), and the

inverse processor of figure 6.32(b) now becomes the receiver terminal of figure 6.3(b). The V/UV switch is used to select all /s/ sounds from the 300 to 7600Hz wideband speech using the zero crossing rate technique outlined in section 4.5.

It may well be argued that the random phase processor is rather complex in operation and the complexity is not really justified in applying bandwidth compression and expansion for just /s/ sounds. The reasons for processing the wideband /s/ sounds only out of the entire speech signal were outlined in section 5.4.2 i.e. it was this sound that suffered the most distortion by bandlimiting it from 300 to 3400Hz. Thus a significant improvement in perception concerning intelligibility and quality may be made by spectrally compressing and expanding the /s/ signal to enable it to be transmitted via a bandlimited medium. We have also noted that the /s/ sound occurs relatively frequently in English prose.

The process of detection and separation of the /s/ signal from 300 to 6500Hz speech was achieved quite reliably from our data using the zero-crossing rate measurement method as the detector used was heavily biased against erroneously deciding that voiced speech was an /s/ sound. The reason for this is that if voiced speech is inadvertently phase-processed, the recovered signal would exhibit a gross periodic distortion as found earlier in our experimentation (figure 6.34).

If the /s/ detector switch changes state inside one 16mS processing block, then that block is considered as "voiced-speech" and does not

undergo any phase processing. Therefore to phase process a 16mS block, the entire 16mS must be considered to comprise of an /s/ utterance. The block diagram of the complete system is now displayed in figure 6.35 which shows that the random-phase process is applied to /s/ sounds only, i.e. the switches S_1 - S_4 are operated by the /s/-detector. The random-phase processor and inverse processor are exactly those illustrated in figure 6.32(a) and figure 6.32(b) respectively. The 'non-/s/' speech, i.e. all of the signal including voiced speech, stop consonants and silence, is processed simply by bandpass filtering from 300 to 3400Hz by FIR filters at the transmitter and receiver terminals. FIR filters are used to bandlimit the non-/s/ speech as the delay is known and can be compensated in the processing branch of the system. Therefore the signals may be synchronised at switches S_2 and S_4. The other switches S_1 and S_3 are so placed as to prevent any unwanted switching transients from occurring at the output of the bandpass filters when the /s/-detector changes state.

Finally from figure 6.35, it is noted that the receiver is required to provide bandwidth extension only at the times when the transmitter performs bandwidth compression i.e. when an /s/ sound is detected. The information can be multiplexed with the transmitted signal in the form of a modulated pilot tone in a dedicated spectral slot, again say, between 2 and 3kHz, similar to the method discussed in Section 5.4.1. The multiplexing arrangement was not however tested experimentally, and the side information was merely transmitted separately from the 300 to 3400Hz channel.

6.4.4.4 Results

a) Time and Spectral Plots

As with the previous bandwidth compression systems, figure 6.36(a) to (d) shows waveforms of the /Is/ in the word "sister" for four different cases. Figure 6.36(a) shows the waveform of the original 300 to 6500Hz signal, figure 6.36(b) is a 300 to 3400Hz FIR bandlimited signal, figure 6.36(c) shows the waveform of received signal from the 300 to 3400Hz channel of figure 6.35. Finally, figure 6.36(d) illustrates the /Is/ signal at the output of the receiver.

Figure 6.36(b) indicates that the 300 to 3400Hz bandlimited signal conveys a large proportion of information regarding the /I/ as here the waveform structure is virtually unaffected. As previously found, the unvoiced /s/ appears to be severely attenuated and distorted. The plot of figure 6.36(c) shows that the random phase processed signal does convey some information about the /s/ sound; when the /s/ signal is subsequently re-expanded at the receiver, the resultant output signal bears a closer resemblance to the wideband input signal than does the 300 to 3400Hz bandlimited version.

The spectral section of figure 6.37(a) shows the magnitude and phase spectrum of 16mS of a wideband /s/ utterance. The phase spectrum may be again noted for its random-like structure. Figure 6.37(b) also shows the magnitude and phase spectrum of the bandlimited version. The characteristic peak in the magnitude spectral envelope

at about 4500Hz has been heavily suppressed. Figure 6.37(c) shows the magnitude and phase spectrum of the input signal to the receiver which corresponds to $\{|\tilde{Y}_k|\}$ and $\{\tilde{\phi}_k\}$ in equations (6.41) and (6.43) respectively. The magnitude and phase spectrum of the recovered wideband output signal is shown in figure 6.37(d) where the magnitude spectrum of the processed signal appears to be a better representation of the input spectrum than the bandlimited spectrum.

b) Informal Subjective Listening Tests

The results of applying random phase processing only upon the /s/ fricative sounds appeared to yield better quality speech than the 300 to 3400Hz bandlimited speech. The result was far better than that gained by processing the entire speech signal. The phase-processed /s/ signal also tended to give a subjective impression that the bandwidth used for the transmission channel was greater than 300 to 3400Hz. In some parts of the processed speech, particularly the first /s/ utterance in "sister" and the /s/ in the word "fist" sounded slightly unnatural and "raspy". However, this situation was improved by attenuating the random phase processed /s/ sounds by fifty per cent. The informal listening tests on the small sample of data used (i.e. the 10-word sequence) did seem to confirm the time waveforms and spectral sections in that the random phase processing system was preferable to the bandlimited speech. Therefore, much more material needs to be processed so that formal subjective listening tests can be undertaken to fully evaluate the performance of the system. If a hardware realisation is feasible, conversational tests would provide a clearer indication of the acceptability of this process as well as the processes previously discussed.

c) Segmented Spectral SNR (SP-SNRSEG)

Apart from the evaluation of the performance of the system by the informal subjective listening test of the output speech signal, the signal to noise ratio may sometimes be invoked. The SNR, however, is more suited to the performance evaluation of high bit rate waveform coders whose task it is to preserve the original waveform as accurately as possible rather than to preserve the speech parameters. When the phase spectrum is altered,as for example with vocoders, the time SNR may be negative as the time waveform is no longer preserved. In these cases, including the random phase processing system, the SNR criterion fails to provide an acceptable performance measure.

A different measure used for these experiments was the spectral SNR which was adopted for the adaptive frequency mapping strategy in section 6.4.3.5, in particular from equation 6.23. The spectral SNR is considered to be more indicative of the performance of a bandwidth compression system that endeavours to recover the original magnitude spectrum rather than the time waveform. The random phase processor is such a system and is therefore posed as a suitable candidate for the application of the spectral-SNR which is computed as follows:

The power spectrum is calculated for overlapping blocks of samples (using 50 per cent overlap) for the input and output signal. These spectral values are then locally averaged with a sliding frequency window of 300Hz bandwidth. The error (or difference) spectrum, i.e. the output minus the input spectral values at each frequency, is

then formed for each block. The SNR is then obtained by taking the ratio of the mean spectral energy of the input and the error spectrum. The average SNR is then determined for all blocks within the entire sequence of speech samples.

Specifically, if $\{|X_{k,j}|\}$ and $\{|\tilde{X}_{k,j}|\}$ are the input and output spectral frequency coefficients for the j^{th} block of input and output signals respectively, then the values must be first locally averaged by the 300Hz sliding frequency window. The frequency samples within the 300Hz window are summed and the result is divided by the width of the window to form the smoothed frequency sample values, $<|X_{k,j}|>$, located at the centre of that window. The window is advanced by one frequency sample and the process is repeated:

$$<|X_{k,j}|> = \frac{\sum_{p=k-L}^{k+L} |X_{p,j}|}{2L} \qquad (6.50)$$

for $k = L,(L+1),\ldots,(N_1/2-L)$

where $L = \frac{k_{0.3}}{2}$ (corresponding to half the window length, i.e. 150Hz bandwidth)

and N_1 = blocklength.

At the start and finish of the spectrum, the window width reduces according to the number of samples available for averaging, such that:

$$\langle|X_{k,j}|\rangle = \frac{\sum_{p=0}^{K+L} |X_{p,j}|}{k + L} \qquad (6.51)$$

for $k < L-1$

and

$$\langle|X_{k,j}|\rangle = \frac{\sum_{p=k-L}^{N_1/2} |X_{p,j}|}{(N_1/2-k) + L} \qquad (6.52)$$

for $k > N_1/2 - L$

The smoothed spectral values for $\{|\tilde{X}_{k,j}|\}$, giving $\langle|\tilde{X}_{k,j}|\rangle$ are computed by the same operations as equations (6.50 to 6.52).

The spectral error signal is then formed, i.e.:

$$E_{k,j} = \langle|\tilde{X}_{k,j}|\rangle - \langle|X_{k,j}|\rangle \qquad (6.53)$$

for $k = 0,1,\ldots,N_1/2$

and the block spectral SNR is formed for the j^{th} block as:

$$S_j = 10 \log_{10} \left\{ \frac{\sum_{k=0}^{N_1/2} <|X_{k,j}|>}{\sum_{k=0}^{N_1/2} E_{k,j}} \right\} \qquad (6.54)$$

For the entire sequence of M_1 samples, there are N_A analysis blocks available to determine corresponding values of S_j where $j = 1, 2 \ldots N_A$ and:

$$N_A = \text{INT} \left[\frac{M_1 - N_1}{(1-b)N_1} \right] + 1 \qquad (6.55)$$

INT [.] denotes the round down integer value of '.', b is the block overlapping factor and N_1 is the blocksize which need not be the same as the blocksize N used for the random phase processing algorithm.

The average segmented spectral SNR (SP-SNRSEG) is thus given by

$$\text{SP-SNRSEG} = \frac{1}{N_A} \sum_{j=1}^{N_A} S_j \qquad (6.56)$$

The measured values for the segmented spectral SNR when applied to the word "sister", having a blocksize N_1 = 256 samples, with the overlap factor b = 0.5 (i.e. 50% overlap), are as follows:

i) Applying a 300 to 3400Hz FIR filter to the entire speech signal with a corresponding output waveform shown in figure 6.35(b).

SP-SNRSEG = 13.9dB

ii) Random phase processing of the /s/ fricative signals with a corresponding output waveform shown in figure 6.35(c)

SP-SNRSEG = 15.7dB

The spectral SNR, S_j from equation (6.54), of each overlapping spectral block, j, for the word "sister" is shown in figure 6.38. Figure 6.38(a) shows the spectral SNR for the bandlimited signal and figure 6.38(b) shows the spectral SNR for the random phase-processed signal. The spectral SNR is about 25 to 30dB for the voiced segments of both signals (a) and (b) as in each case the signal is bandpass filtered at the same bandlimits. The spectral SNR for the /s/ utterances in figure 6.38(a) is between about 0.1 to 0.3dB, whereas this increases to around 5 to 9dB in figure 6.38(b).

6.4.4.5 Discussion and Conclusion

It has been argued that enhancement of speech intelligibility and quality may be obtained by bandpass filtering voiced speech to 300-3400Hz and phase processing the 300 to 6500Hz bandwidth /s/ speech signals into the 300 to 3400Hz band. No other sound is phase processed since the /s/ sound is considered to make a vital contribution to the quality and intelligibility of wideband speech. If a non-/s/ block happens to be inadvertently phase-processed it is likely that the input waveform would exhibit a high zero-crossing rate such as that for the trailing edge of a /t/ consonant. The high zero-crossing rate of the waveform to be phase processed is governed by the /s/ detector switch and when such a waveform is phase processed, one should not expect any serious degradation although the procedure of testing 'non-/s/' unvoiced speech was not performed experimentally.

Another consequence of using the phase-processing method is that the approach used to reduce the block-end distortion was by generating a pseudo-random phase spectrum for the transmitted signal. This notion supports the procedure adopted for the freqency de-mapping process in section 6.4.3.2, i.e. when a random phase was assigned to the estimated components, the block-end distortion effects were significantly reduced.

The systems so far discussed in this chapter for wideband processing of unvoiced segments may be compared at this stage using the informal subjective listening experiences obtained from the 10-word

sequence of section 4.2. We consider that the quality of the processed speech obtained from the random phase method sounds superior to that obtained from the VUBS method of section 6.3 but perhaps inferior to the fixed and adaptive versions of the frequency mapping systems. The adaptive frequency mapped speech appears to fare the best over all the systems explained so far in this chapter. This is reflected by the complexity which is complementary to the processed speech quality for the four types of systems. The VUBS method is the simplest to implement and the adaptive frequency mapping process is the most complex system.

As far as objective comparisons are concerned, we have argued that the spectral SNR is perhaps the better indicator of the performance of a system rather than the more familiar time waveform SNR measure.

The spectral SNR as obtained for the random phase processing system may be applied to other systems in exactly the same manner. This is a subject to which we will be returning later in section 7.4.1.

We will at this point return to consider processing in the time domain and outline a further spectral compression scheme. The notion is based upon the approach by Malah(ref 2) which already has been reviewed in section 3.4.2.2. The modified process will be the topic of the following section.

6.5 Time Domain Harmonic Scaling (TDHS) of Unvoiced Speech Signals

The discussion of the TDHS process in Chapter 3 showed that the system was devised for the intention of processing the entire speech

signal such that a 3400Hz bandwidth signal would be linearly compressed by a factor of two or three. In this section, we will adopt the same idea but will direct the application to the objective of wideband speech compression. In particular we will concentrate on processing the unvoiced speech sounds only. We may therefore dispense straight away with any form of pitch tracking that was needed in the method described in section 3.4.2.2. In order to apply the TDHS procedure, we return to the structure of the basic compression system of figure 6.3(a) and (b) where the processor is now the time domain harmonic compressor and the inverse processor in the time domain harmonic expander. Notice that as the processing is facilitated in the time domain, the DFT and IDFT operations are no longer needed.

The time domain operations employed by the transmitter and receiver are the same as equations (3.87) and (3.88) which are rewritten here for the reader's convenience.

6.5.1 Time Domain Harmonic Compression (TDHC)

Using a 2:1 frequency compression, the time domain operation is

$$y^{1/c}(\ell) = x(\ell-N_p) + h_N(\ell)\left[x(\ell) - x(\ell-N_p)\right] \qquad (6.57)$$

$$\text{for } \ell = 0,1,\ldots,N_c-1$$

where $x(\ell)$ is the input time sequence

$y(\ell)$ is the frequency compressed output time sequence

N_p is the number of samples in the defined "pitch" interval of $x(\ell)$

$(N_c + N_p)$ is the number of samples contained in the triangular window h_N

and c is the frequency compression factor.

Although we are only concerned with processing unvoiced speech, it is more convenient to illustrate the processing applied to periodic voiced speech in figure 6.39 bearing in mind that the actual defined "pitch period" of the unvoiced speech is an arbitrarily fixed blocksize equal to N_p samples.

The illustration of figure 6.39 shows the operation of the TDHC processor using the equation (6.57). The original algorithm is written as

$$y^{1/c}(\ell) = x(\ell)\, h_N(\ell) + x(\ell-N_p)\, h_N(\ell+ N_c) \qquad (6.58)$$

which when using the triangular window function, h_N and using $N_p = N_c$, the equation (6.58) reduces to (6.57). It can be seen that for every two samples of $x(\ell)$ spaced by one pitch period apart only one sample of $y(\ell)$ is generated, i.e. samples a and a' combine to form a", similarly samples b and b' combined to form b". The new sequence $y(\ell)$ has only N_c samples per $(N_p + N_c)$ input samples of $x(\ell)$, consequently, to maintain temporal continuity of the output signal, the sampling rate of $y(\ell)$ is reduced by a factor of c such that $T' = cT$. Thus the new sampling interval is twice that of the original for our case when $c = 2$.

The operation of the frequency compressor system is similar to that of an interpolator where a single pitch interval of output speech is produced by interpolating samples from two pitch periods of input signal, the samples to be interpolated are spaced one period apart. The new pitch period is reduced in sampling rate by one half, corresponding to a time dilation factor of two which afford linear frequency division by the same factor.

For our case, the values of N_p and N_c were chosen to be 128 samples each corresponding to a duration of 8mS with a corresponding window length (N_p+N_c) of 16mS.

6.5.2 Time Domain Harmonic Expansion (TDHE)

Using a 1:2 expansion corresponding to the 2:1 compression process at the transmitter of figure 6.39 the time domain operation is:

$$\tilde{x}^{s}(\ell) = \tilde{y}(\ell-N_p)\, h_N(N_s-\ell) + \tilde{y}(\ell-2N_p)\, h_N(2N_s-1)$$

$$\text{for } \ell = 0,1,\ldots,(N_s-1) \qquad (6.59)$$

where $\tilde{y}(\ell)$ is the receiver input sequence

$\tilde{x}(\ell)$ is the frequency expanded output time sequence

N_p is the number of samples in the defined pitch interval of $\tilde{y}(\ell)$

$2N_s$ is the number of samples contained in the triangular window of h_N

and s is the frequency expansion factor.

The expansion process of equation (6.59) is applied at the receiver of figure 6.3(b), the process itself being illustrated in figure 6.40.

It may be seen from figure 6.40 that for every period, N_p of $\tilde{y}(\ell)$, two periods of $\tilde{x}(\ell)$ are produced, i.e. samples a and a' combined to form a" also samples b and b' combine to form b". When N_s samples of output signal $\tilde{x}(\ell)$ have been generated [by N_s pairs of the input signal, $\tilde{y}(\ell)$], the signal $\tilde{y}(\ell)$ is advanced by N_p samples and the process is repeated. Since $N_p = N_s/2$, each period of $\tilde{y}(\ell)$ is used twice to form the output signal $\tilde{x}(\ell)$, so again to maintain temporal continuity, the time scale of $\tilde{x}(\ell)$ is compressed such that $T" = T'/s$ which increases the sampling rate by a factor of s. This effectively performs linear multiplication upon the frequency scale of the output signal, $\tilde{x}(\ell)$. If we choose s=c then T"=T, i.e. the bandwidth of the output signal from the TDHE process is equal to that of the input signal to the TDHC process, although we have seen that the transmission bandwidth is reduced by a factor of c.

As with the compression process, a triangular window weighting function, h_N, is used at the expander and so the equation (6.59) can be slightly simplified to (6.60) which is the expression used for our simulation purposes.

$$\tilde{x}^s(\ell) = \tilde{y}(\ell-2N_p) + h_N(N_s-\ell)\left[\tilde{y}(\ell-N_p) - \tilde{y}(\ell-2N_p)\right] \qquad (6.60)$$

$$\text{for } \ell = 0,1,\ldots,(N_s-1)$$

where the symbols have the same meanings as those in equation (6.59).

When the process of equation (6.60) was applied to expand the spectrum of unvoiced speech at the receiver, the chosen value of s was 2 (i.e. frequency doubling). N_p was set to 128 samples of receiver input speech signal, corresponding to 16mS, with N_s set to 256 samples of receiver output speech signal. This still corresponds to 16mS due to the higher sampling rate.

6.5.3 Results

As stated during the explanations of the spectral compression and expansion procedures, the processing was applied to unvoiced speech only using the structure of figures 6.3(a) and (b). The unvoiced speech segments were selected by visual inspection of the wideband input speech waveform as described in section 4.3. All other speech sounds including voiced speech and silence pauses were simply bandpass filtered at the transmitter of figure 6.3(a) prior to transmission through the 300 to 3400Hz analogue channel.

As with some of our previous systems, we demonstrate the TDHS process using a 16mS block of wideband /s/ utterance as shown in figure 6.41(a) with the accompanying spectrum as illustrated in figure 6.41(b).

The 16mS block corresponds to (N_p+N_c) samples from which N_c samples are formed by the TDHC processor, as shown in figure 6.41(c). If the sampling interval is doubled, the processed block occupies a 16mS time interval shown in figure 6.41(d), which has a corresponding spectrum as shown in figure 6.41(e). The spectrum has

a slightly lower cut off frequency than 4.0kHz because it already has been bandpass filtered from 300 to 3400Hz by the channel filter, otherwise the TDHC spectrum should span one half of the original folding frequency of the input signal, i.e. 8/2kHz.

At the receiver, the application of expression (6.60) yields the output signal as shown in figure 6.41(f) after the sampling rate has been increased by the factor of two. The output signal has a corresponding magnitude spectrum of figure 6.41(g) where it can be seen that the important spectral peak is preserved with good integrity and shows good promise regarding the listening quality of the processed signal to which we will come shortly in this section.

The time waveform plots for the original, bandlimited and transmitted and output processed signals are shown in figure 6.42(a) to (d) respectively for the utterance /sI/ in "sister". It can be seen that the waveform of the transmitted signal has a lower zero-crossing rate than that of the original waveform but the amplitude envelope is fairly well preserved, this was also true for the frequency mapping and voiced-unvoiced band switching system but not true for the random phase process. The receiver output signal shown in figure 6.42(d) illustrates a good correspondence with the input signal of figure 6.42(a) both in terms of waveform structure and zero-crossing rate. Although fixed length block processing has been used for this process, there is no evidence at all of any block edge discontinuity or distortion since the DFT and IDFT are not employed in the TDHS method.

6.5.3.1 Comparison by Informal Listening Tests

The informal subjective listening tests seem to show that the TDHS processed signal conveys more high frequency information than does the bandpass filtered signal of figure 6.42(b) and perhaps sounds more like the original signal of figure 6.42(a). It may be noted, however, that there is a quite significant amount of unnatural "buzziness" associated with the TDHS processed signal, particularly with the /s/ utterances, although this effect does not appear in the time waveform plot of figure 6.42(d). The spectrum does however show evidence of a harmonic structure due to block segmentation which is responsible for the "buzzy" quality of the processed signal. This unnaturalness unfortunately down-grades the subjective performance of the TDHS method when applied in the arrangement of figure 6.3.

As an additional point of interest, we may also consider the perceptual impression of the transmitted signal of figure 6.4(c). Although it was bandlimited from 300 to 3400Hz, i.e. the same as that for the signal in figure 6.42(b), it did sound significantly different from that signal. It was still quite intelligible although the quality was impaired which rendered the impression that the speaker had a "lisp" when uttering the /s/ sounds. The perceptual quality of the transmitted signal for any of the frequency compression notions discussed so far in this chapter that employ the basic structure of figure 6.3 did not vary significantly from one system to another. They were all quite different from the bandlimited signal and in every case, the bandlimited signal provided a better subjective impression of quality than the transmitted signals for any of the systems discussed. One

observation that seemed apparent from this latter investigation was that if a listener did not possess as a suitable receiver terminal to apply the appropriate spectral expansion algorithm to the speech, then he would still be able to receive intelligible speech but at a reduced quality than the signal produced by the receiver output.

At this stage, it is again emphasised that all of our subjective listening experiences are based only on a small amount of speech material (i.e. that obtained in section 4.2) therefore all of our findings will need to be more thoroughly examined to fully evaluate the systems discussed.

6.5.4 Discussion and Conclusion

In this section, we have discussed an approach based on the time domain harmonic compression and expansion algorithm proposed by Malah(ref 2). The system was applied in the form of figure 6.3(a) where wideband speech may be spectrally compressed into the 300 to 3400Hz bandwidth and re-expanded at the receiver of figure 6.3(b). The operation of the system was fairly simple in comparison with the previously discussed frequency compression methods using the DFT and IDFT. Also, in contrast to Malah's original application of bandwidth compression of voiced and unvoiced speech, we do not employ pitch detection as we only apply the process to unvoiced speech.

As far as the results are concerned, we have already noted that as with the frequency mapping and voice/unvoiced band switching system, the amplitude envelope of the processed transmitted waveform closely

resembled that of the original waveform. This is possibly because during the spectral compression of the signal, one may envisage that the shape and structure of the signal spectrum is unaltered whereas the frequency scale is warped, i.e. expanded linearly or non-linearly. Therefore we would expect that the shape and structure of the time waveform to undergo little change except a corresponding warping (compression) of the time scale; the reverse argument will apply at the spectral expansion process at the receiver. This argument has already been presented in more general terms during the discussion of the analytic-rooter (refs 38,39). In section 3.3.10 it was ascertained that during frequency scaling of the analytic signal, the waveform envelope was scaled in amplitude only without undergoing any temporal compression.

We have further noted that perceptually, the transmitted signal sounds less favourable in terms of quality than the equivalent bandlimited signal of the same bandwidth. The distortions imposed by bandpass filtering are more tolerable than those produced by the spectral compression processes; our listener(s) may be more acclimatised to the former rather than to the latter. Certainly, examples of the bandlimited speech are far more abundant than cases of the spectrally compressed speech although if some of the ideas of spectral compression are put to use commercially then this may no longer be the situation.

Coming back to the subjective appreciation of the TDHS receiver output speech, we have found that although the impression of a wider bandwidth channel was apparent over that of the 300 to 3400Hz

actually employed, the processed /s/ utterances sounded somewhat "buzzy" in comparison to the original signal. This may be due to the fact that in the development of the TDHS algorithm formed by Malah[ref 2], it was assumed that the signal block contained harmonically related components to the block interval which is true when a good pitch tracker is used for voiced speech. In our case, however, no such periodicity is applicable, and therefore, as a plausible argument, the TDHS process may form harmonically related spectral error components which are subjectively manifested as a background "buzz-noise". Clearly, the problem requires a more thorough investigation in order to render the wideband TDHS process more viable. The problem may be circumvented by applying the same refinement Malah applies in order to compensate for the errors caused by the pitch detector applied to voiced speech. This allows for the deviation of the pitch harmonics from the centre frequencies of the contiguous analysis filters upon which the TDHS process is derived [ref 2].

Another limitation of the TDHS process is that only linear frequency compression and expansion can take place by the time domain operations of equations (6.58) and (6.60). The algorithm may be elaborated by developing time domain operations which inherently impose non-linear frequency scaling in favour of subjectively significant bands, as with the adaptive frequency mapping scheme. As concluded in section 3.4.2.2, the system may be developed more simply by combining sub-band coding of section 3.4.1.2 with the TDHS process to apply different compression ratios in different bands according to some subjective criteria.

Having now discussed the TDHS mechanism, this rounds off our presentation of wideband compression systems in this chapter and we will now briefly turn our attention to the channel coding of the 300 to 3400Hz transmitted signals produced by these schemes.

6.6 Digital Coding of the 300 to 3400Hz Speech Channel

In this section we briefly consider the characteristics of the 300Hz to 3400Hz bandwidth channel used to transmit the spectrally compressed signal to the receiver. We have so far assumed that the channel has an ideal, linear-phase and noiseless characteristic, but we can now examine more realisable forms of this channel. In particular, we will investigate the consequences of conveying the spectrally compressed signal by digital means and look at ways of rendering this as efficient as possible giving good performances at low cost and using conventionally available digitisers as we have done with the high frequency regeneration schemes of Chapter 5. Experiments have been conducted by encoding the transmitted signals of three types of bandwidth compression systems; namely the VUBS system; the fixed frequency mapping system and the adaptive frequency mapping scheme. The work has been carried out in the department of Electronic and Electrical Engineering at Loughborough University of Technology by C C Evci,[refs 87,88] W K Cham[refs 89,90] and V K C Tse[ref 91] respectively. Since the transmitted signals for all the three systems had shown very similar traits, e.g. compare figures 6.5(c) and 6.23(c), it is thus considered sufficient to explain the procedures and results of the digitisation of one spectral compression scheme only.

The work undertaken involving the fixed frequency mapping scheme plus the digitally encoded channels[refs 89,90] was the first to be fully completed and therefore, this is the system that we have chosen to be overviewed. For this examination, four digital encoders were compared for three different types of signals and were operated at three different digital transmission bit-rates. This involves a maximum possible number of 36 output signals to be assessed, although, as we will see later, the number of actual examinations was limited to thirty.

The three different signals used to test the digital coders are demonstrated by the arrangement shown in figure 6.43. In the first case, the input signal is 300 to 3400Hz speech, and for the second case, the input speech signal is allowed to have a 300 to 7200Hz bandwidth. For the last case, the input signal is initially pre-processed by the fixed frequency mapping algorithm to compress spectrally the 300 to 7200Hz speech into a bandwidth of 300 to 3400Hz before being encoding by the digitizer. The output from the channel decoder is post-processed by the inverse frequency mapping process to accomplish the corresponding spectral expansion of the speech signal. The pre-processor shown in figure 6.43(c) is the unvoiced speech spectrum compressor of figure 6.21 employing the fixed frequency mapping procedure, and likewise, the post-processor in figure 6.43(c) is performed by the unvoiced speech spectrum expander of figure 6.21 and this ultises the inverse fixed frequency mapping process. The signal to be encoded has the same bandwidth as commercial speech of 300 to 3400Hz. However, it does possess

slightly different properties to that of normal 300 to 3400Hz speech and therefore the encoders used to digitise the bandwidth compressed speech need to be slightly re-optimised from their operating conditions in figure 6.43(a). The signals to be encoded in figures 6.43(a), 6.43(b) and 6.43(c) are digitised by one of the four encoders - APCM, ADPCM, CFDM and CVSD. Since we have explained the operation of the four coders in section 3.2 and have already applied them to the spectral enhancement schemes in section 5.5 we will not devote any further attention to the optimisation of the coders here. The coders themselves had been tested at a number of different input signal power levels and their optimum operating points (i.e. the signal power which produces the maximum segmented SNR values) had been sought. Before we go on to discuss the comparisons of the three different digital speech transmission systems in figure 6.43, we will consider the input sampling rate requirements.

6.6.1 Sampling Frequency Conversion (ref 89)

The three systems of figure 6.43 were compared under three different transmission bit rates - i.e. 16, 24 and 32 kbits/sec. When a delta modulator is used, the channel transmission bit rate is equal to the sampling rate. This implies that the 16, 24 and 32 kbits/sec channel transmission rates require input data of sampling rates 16, 24 and 32kHz respectively which is achieved by adjusting the sampling rate of the input signal. The conversion of the sampling fequency is performed by a combination of linear interpolation, filtering and decimation of the speech samples. If the old and the

new sampling frequencies are integer multiples of each other, then a single linear interpolation or linear decimation is sufficient to accomplish the conversion. However, if the old and new sampling frequencies are not integer multiples of each other then the conversion between them needs to be performed in two steps. This involves converting the old sampling frequency, F_s, to an intermediate frequency, F_s', and then converting the intermediate to the new sampling frequency, F_s''. The value of F_s' can either be the LCF (largest common factor) or LMS (least common multiple) of F_s and F_s''. If F_s is equal to 16 kHz and F_s'' equal to 24 kHz then the value of F_s' can be 8 kHz or 48 kHz which are the LCF and LCM values of F_s' and F_s'' respectively.

If the LCF is used, then the sampling frequency is carried out by first decimation and then interpolation. On the other hand, if the LCM is used, this is carried out initially by interpolation followed by decimation. The interpolation process introduces errors and hence it requires lowpass filtering to suppress them before decimation can be applied. However if the decimation can be performed first, the subsequent interpolation process will still cause errors in the new sequence and therefore will again require lowpass filtering. The computation time required by each of the LCF or LCM method is the same and so from this standpoint, neither method is more preferable to the other.

If the old sampling frequency, F_s, is equal to 16 kHz and the signal bandwidth extends up to 7.2 kHz as in the case of the system shown in figure 6.43(b), the LCF method will cause aliasing by the

decimation process and this will not be removed by a linear interpolation process. Hence, the LCM method must be used in cases such as this to increase the sampling rate to 24 kHz.

So, using all these factors regarding the conversion of the sampling frequency of the input signal, we can construct a table (table 6.4) to summarise the processes required in order to perform this conversion procedure.

Upper Cut-Off Frequency of Input Signal (Hz)	Old Sampling Frequency (kHz)	New Sampling Frequency (kHz)	Conversion Process
3400 Bandlimited Speech	16	8	2:1 sample decimation
		24	2:1 decimation followed by 1.3 linear interpolation
		32	2:1 linear interpolation
7200 Wideband Speech	16	8	Not applicable (new sampling is less than Nyquist rate)
		24	1:3 linear interpolation, 8kHz lowpass filter, then 2:1 decimation
		32	1:2 linear interpolation
3400 Frequency Compressed Speech	16	8	2:1 decimation
		24	2:1 decimation followed by 1:3 linear interpolation
		32	2:1 linear interpolation

Table 6.4 Sampling Frequency Conversion

When a multilevel codec is used, i.e. APCM or ADPCM, the input sequences were reduced in sampling rate from 16 kHz to 8 kHz (by the process indicated in Table 6.4) and then encoded by a 2, 3 or 4-bit Jayant quantizer (ref 22). The encoded bit stream was transmitted at 16, 24 and 32 kbits/sec respectively.

For the case of the wideband transmission system, as shown in figure 6.43(b), the bandwidth of the input speech is 300 to 7200 Hz of which the Nyquist rate is 16 kHz and therefore it is not possible to reduce the original sampling rate at all without causing aliasing of the input signal. The smallest number of levels of a multilevel quantizer is two, therefore, the minimum channel transmission rate that the wideband system has to be is 32 kbits/sec (unless bit-sharing is used between samples). In our comparisons, the use of a multilevel quantizer below 32 kbits/sec was not employed.

At the receivers of figure 6.43(a), 6.43(b) and 6.43(c), the binary sequence is decoded into PAM samples and then bandpass filtered to suppress any out-of-band noise components. The bandwidth of the filter is dictated by the bandwidth of the input signal used for each process. The sampling frequency used to store the processed data was 16 kHz in all cases. This is an adopted standard for the storage of all the input and processed speech data. When necessary, the processed data needed to be converted back to this sampling frequency by applying the reverse process to the procedures in table 6.4, i.e. to convert from 8 kHz to 16 kHz sampling frequency, 1:2 linear interpolation was performed, and so on. The exception being that for the case of converting 24 kHz wideband output speech down

to 16 kHz, 1:2 linear interpolation was first used followed by lowpass filtering to 7200 Hz and afterwards a 3:1 sample decimation was employed. This was to avoid the aliasing that would be caused by decimating the sample sequence before applying the interpolation procedure.

Having now presented the three experiments shown in figure 6.43 and gathered together all the processed material from the different coders and bit rates, we then went on to compare the recovered signals and considered the overall performance of these three systems.

6.6.2 Results and Discussion

6.6.2.1 Informal subjective testing

Comparisons were made for the three digital transmission systems in figure 6.43(a), (b) and (c) by informal subjective listening tests performed on the output speech signal; the 10-word sequence described in section 4.2 showed that in all cases, better output signal quality resulted when using the higher channel transmission rates as the quantization noise reduces for all the coders that were tested. As we have seen in section 5.5, the quantization noise reduces as the number of bits per sample is increased for the multilevel coders. For the differential coders, when the sampling rate is increased, this increases the sample to sample correlation and hence reduces the variance of the error sequence.

When the bandlimited speech transmission system and the wideband transmission system were compared, it was found that the wideband system offered a better subjective speech quality than the bandlimited speech system at transmission bit rates of 32 and 24 kbits/sec. When the transmission bit rate was lowered to 16 kbits/sec, the speech signal from the wideband CVSD system was still preferable to that obtained from the bandlimited speech plus the CVSD system. When the CVSD was replaced with the CFDM codec the two systems of figure 6.43(a) and (b) showed similar performances. This is possibly because the 16 kbits/s-CFDM coder was unable to track the rapidly varying unvoiced signals and therefore distorted and attenuated these speech components thereby rendering the wideband performance comparable with that of the bandlimited speech system. The CVSD codec appeared able to track the unvoiced /s/ waveform (see next section) although it seemed unable to preserve the original shape of the waveform. Generally it may be suggested that the wideband systems offered better performances in quality than the bandlimited systems when the transmission bit rate was high enough to transmit the unvoiced sounds.

When the wideband transmission systems were compared with the bandwidth compression systems (figures 6.43(b) and (c)), it was found that the subjective quality of the system using bandwidth compression was superior to the quality discerned from the wideband system at the same transmission bit rate. The bandwidth compression system also produced a better impression of quality and output signal bandwidth than when bandlimited input signals were used at similar transmission rates. The most marked improvement in

subjective performance was noted when the 16kbits/sec delta modulators were employed. The high level of quantization noise produced was very noticeable for the bandlimited and wideband systems, but this appeared to be significantly masked by the presence of the high frequency unvoiced sounds from the bandwidth compression system. The bandwidth compression systems were more successful in recovering the higher frequency unvoiced sounds than the wideband systems as the rapidly varying unvoiced signals were effectively slowed down by the frequency compression system prior to encoding. This eased the task of the coders when attempting to track the unvoiced sound.

Therefore the better subjective performances were discerned from the bandwidth compression systems rather than for any of the wideband or the bandlimited systems used. We may investigate the findings more closely by referring now to the signal waveforms obtained en-route from one of the digital coders.

6.6.2.2 Waveforms

Some waveform comparisons were made for the performance of the 16 kbits/sec CVSD coder operating under the three conditions of figure 6.43(a), (b) and (c); these waveform plots are shown in figure 6.44. The original waveform for the /sI/ in sister is shown in figure 6.44(a) while the output waveform from the bandlimited system using the 16kbits/sec CVSD coder is shown in figure 6.44(b). The output signal from the fixed frequency mapping pre-processor is shown in figure 6.44(c) and when this signal is encoded and decoded by the

16kbits/sec CVSD system the resultant waveform is illustrated in figure 6.44(d). Finally this signal from figure 6.44(d) is passed through the fixed frequency de-mapping procedure from section 6.4.3 to yield the final output signal of figure 6.44(e). The signal may be compared with the output from the wideband digital system of figure 6.43(b), which is shown in figure 6.44(f).

Conventionally, the 16 kbits/sec CVSD coder is operated on a 3.4 kHz signal and so the quantization noise appearing in the voiced /I/ waveform is going to be a dominant feature in the listening quality of this sound. This was indeed found to be the case in our subjective tests of the output signal. When the wideband signal of figure 6.44(a) was processed by the 16 kbits/sec CVSD coder, then the presence of the output signal due to the unvoiced /s/ tended to mask the effect of the coder noise as shown in figure 6.44(f). The shape of the processed /s/ was still heavily distorted which was again supported by the results of the informal subjective listening tests. When the bandwidth compressed signal of figure 6.43(c) was processed by the same coder, the resultant waveform of figure 6.44(d) showed similar traits to that of figure 6.44(f) perhaps indicating that the bandwidth compressed /s/ signal may well be the fastest changing signal that this particular coder is able to cope with. The signal of figure 6.44(d) when further processed by the inverse frequency mapping configuration, yielded the final output signal of figure 6.44(e). We found that the processed /s/ signal still appears heavily distorted as compared with the original signal of figure 6.44(a) but it did seem to be its best representation when compared with the waveforms of figure 6.44(b) and 6.44(f).

The voiced waveform of figure 6.44(e) appears to be relatively free from the quantization noise due to the coder; this was achieved by the 300 to 3400 Hz bandpass filter in the frequency mapping receiver of figure 6.3(b). The voiced waveform of figure 6.44(b) had not been bandlimited by a similar filter in error. If this had been done then more of the quantization noise would have been removed prior to plotting this waveform. The filter had however been inserted before conducting the subjective listening tests and so these tests are still comparable.

A final test which was conducted on the digital coding systems was the segmented signal to noise ratio which posed as an objective measurement to compare the input and output waveforms of the systems. It was however significant from the waveforms of figure 6.44 that no system preserved the actual shape of the original waveform particularly the unvoiced /s/ and so the SNR measurements tended to give rather pessimistic readings in all cases. In future then, it may well be more indicative to use the spectral SNR measurements, as was used for the random phase processing arrangement in section 6.4.4. These measurements should assist in discerning how well the spectral distribution is preserved as the measurement does not rely upon processed time waveform integrity. This is an obvious task for further investigation into the performance of these digital coding systems.

We have concentrated on presenting our results for the delta modulators. Since these did indicate the most contrasting performance of the systems in figure 6.43 (especially at low

transmission bit rates), it is however fair to say that according to our subjective impressions, the best form of digital coding employed was the ADPCM coder. This was found to be true at all transmission bit rates and as we have stated in section 5.5, it is because this type of coder combines the advantages of multilevel quantizing and differential signal coding.

We have mentioned that the four coders had also been tested using the adaptive frequency mapping bandwidth compression scheme of section 6.4.3.5. The work was carried out by V T Tse[ref 91] at the Department of Electrical Engineering, Loughborough University of Technology. The results were, indeed, very similar to those obtained using the fixed frequency mapping procedure discussed here and so they do not really warrant any further description in this thesis.

The bandwidth compression procedure of section 6.3, i.e. the VUBS system, was also digitised by an ADPCM encoder employing a novel technique to control the adaptation of the predictor stage. This investigation was undertaken by C C Evci[refs 87,88] at Loughbourough University of Technology, and the interested reader is highly recommended to consult these references for further details relating to the study of encoding pre-processed signals to reduce the transmission bit rate.

Our original intention was to compress a 7kHz speech signal into a 3.1 kHz signal and to convey it via an analogue telephone-bandwidth channel. Subsequently, we used conventional codecs to digitise this signal. An alternative strategy is to encode the 7kHz speech signal directly and apply bit-rate compression techniques using novel coding schemes; however, time did not permit us to pursue this approach.

6.7 Overall Discussion and Conclusions

In this chapter we have looked at different techniques with the objective of conveying wideband 300 to 7600 Hz speech via the telephonic bandwidth channel of 300 to 3400 Hz. Most of the methods developed had fitted into the basic structure of figure 6.3, i.e. the bandwidth modification was carried out only upon the wideband constituents of the input signal, namely the unvoiced speech utterances. The operations in the time domain, i.e. the VUBS and TDHS methods, were the simplest to implement but had the penalty of yielding slightly more unnatural processed sounds in the reconstructed signal than their frequency domain counterparts. This was notably true for the /s/ utterances. A slight modification from the basic system of figure 6.3 was imposed by the adaptive frequency mapping scheme and this seemed to offer the best quality of reconstructed signal than the other methods discussed in this chapter. The outcome was matched by the associated complexity of the above scheme which was evident in the amount of computer execution time required to simulate the method.

From our informal subjective listening tests based upon the 10-word sequence, we rank the order of preference, as far as the listening quality of the processed signal is concerned, in descending order as follows:

(i) Adaptive frequency mapping scheme

(ii) Fixed frequency mapping scheme

(iii) Random phase processing

(iv) Time domain harmonic compression

(v) Voiced/unvoiced band switching system

(vi) Second order Fourier transform method.

Perhaps it is unfair to include the last method as it was by no means fully developed to its possible best potential. Leaving item (vi) aside, we find that the more refined frequency domain approaches appear to yield the best overall speech quality. All the methods of (i) to (v) tended to produce speech that gave a subjective impression that a wider than 300 to 3400Hz channel was used to convey the signal and all of the signals appeared to show an improvement over the conventional bandlimited speech of 300 to 3400Hz.

If we compare the quality of speech produced by the latest development of the quality enhancement scheme detailed in section 5.4.3, we find that the enhancement scheme is perhaps adequately placed between (ii) and (iii) of the above "quality-scale". This is probably because the enhancement scheme is more speech-specific than the random phase processor of section 6.4.4 but the frequency mapping schemes appear to convey more of the necessary information regarding the spectral reproduction of the signal which in this case seems to correlate well with a better listening performance.

When applying digital coding to the transmitted signal, we found that the enhancement scheme of the previous chapter was the simplest to implement; this was due to the fact that the transmitted signal was indeed a conventional bandlimited signal. The frequency mapping

and VUBS systems, on the other hand, produced a different form of 300 to 3400Hz signal specifically for the compressed unvoiced sounds. This meant that the coders required re-optimisation for the bandwidth compression schemes only and not for the quality enhancement schemes. Apart from this obvious difference, the process of channel digitisation produced similar results as far as coder performance and channel transmission rate comparisons were concerned.

Finally, it was noted for all the systems, listed (ii) to (v) on the above scale, that the reproduced /s/ sound sometimes did not blend quite naturally with the bandlimited voiced speech from each respective system. This was perhaps due to the dichotomous switching of the voiced/unvoiced detector used to decide whether the input signal should be bandwidth compressed or not. The subjective effect of listening to a signal that changes abruptly from wideband to narrow band was sometimes disturbing and impaired the overall quality of the processed signal. This outcome can be obviated by the use of some form of gradual switching to blend temporally the wideband and narrow band signals more smoothly. The form of switching used and the processing of the resultant signals (thereby modifying the model in figure 6.3) is a question posed for further consideration.

One such system which did provide a possible solution to the above problem was the adaptive frequency mapping scheme, where different mapping laws were used to suit the incoming spectrum. During sustained sounds, the mapping law used was not found to change

appreciably. It was notably at the transitions between sounds that differing laws had been selected enabling the best spectral fit to the original spectrum to be established. This form of adaptation appeared to demonstrate a better quality output signal as far as the informal subjective listening experiences were concerned. A simplification of the above scheme would be to use a set of template spectra and transmit a code relating to the one that gives the best spectral fit to the input spectrum (in a similar manner to the pattern matching vocoder described in section 3.3.1.1). As well as enabling a smooth transition between sustained voiced and unvoiced sounds, this could also be used to model the more rapidly changing spectra in the input speech, particularly the stop consonants. This again is a good candidate for future consideration.

6.8 Note on Publications (refs 88,90, 92-94)

A paper entitled "Voiced/Unvoiced Band-Switching System for Transmission of 6kHz Speech over 3.4kHz Telephone Channels", in co-authorship with Dr R Steele and Dr C S Xydeas has been published in The Radio and Electronic Engineer Journal of the IERE, Vol 51, No 5, May 1981. The paper is a brief transcription of the description in section 6.3 and the results of 6.3.1.

A paper has also been published and presented in an international conference on Acoustics, Speech and Signal Processing by the IEEE at Atlanta, Georgia, USA in April 1981. The paper is entitled "Wideband Quality Speech Encoders with Bit Rates of 16-32kbits/s" and presents a summarised version of the fixed frequency mapping

strategy of section 6.4.3 in conjunction with the digital coding schemes of section 6.6. Tape recorded demonstrations were also included in the conference.

An article entitled "Frequency Compression of 7.6kHz Speech into 3.3 kHz Bandwidth"(ref 94) in co-authorship with Dr R Steele and Dr C S Xydeas has been published in the IEEE Transactions on Communications, Vol 31, No 5, May 1983. The paper includes the main results from section 6.4.3.5 regarding the adaptive frequency mapping algorithm. An abridged version of this paper was presented in an international conference on Acoustics, Speech and Signal Processing by the IEEE at Boston Mass., USA in April 1983.(ref 93) Again taped demonstrations were included in the conference.

Chapter 7

Recapitulation

7.1 Introduction

In this thesis a number of methods to overcome the effects of bandlimiting speech signals have been proposed and investigated. The work presented has been broadly categorised into:

(i) quality enhancement of bandlimited speech signals, where an attempt was made to produce a speech signal that subjectively appears to have a wider bandwidth than that available at the receiver input;

(ii) where a wideband speech signal is spectrally compressed into a bandlimited channel and subsequently re-expanded at the receiver to yield a wideband signal which as closely as possible resembles the original signal.

Both of the categories were tackled by spectral extension and scaling techniques in the time and frequency domains. In the first case, the frequency domain techniques proved to be simpler than the time domain methods but the converse was true for the latter case. Most of the techniques discussed were first applied with equal processing to the entire input speech signal (i.e. voiced and voiceless speech plus silence pauses). During the course of the research, it was realised that the telephone bandwidth of 300 to

3400Hz was quite adequate for conveying voiced sounds and silence intervals. Unvoiced speech, on the other hand, contains significant spectral components well outside the telephone range and therefore these sounds were found to suffer the most distortion and attenuation. It was these types of sound that our investigation had been focussed upon.

In dealing with the separation of voiced, unvoiced and silence pauses, various decision strategies were examined. These ranged from signal waveform inspection to power level and zero-crossing measurements plus spectral distribution techniques. It was found a relatively straightforward task to detect the /s/ sounds and stop consonants from wideband speech but certainly it was more difficult to detect reliably /s/ sounds from bandlimited speech. Some of the spectral enhancement techniques used required only the selection of voiced speech which was found to be adequately achieved by measuring the first shift autocorrelation coefficient and applying an empirically determined threshold to detect the voiced speech. This latter method performed equally well for wideband and bandlimited speech.

The voiced/unvoiced decision techniques discussed were by no means exhaustive, but they did serve as useful 'front-ends' to our spectral extension and wideband compression algorithms now summarised.

7.2 The Spectral Extension Techniques

These were generally concerned with the replacement of the out of band components that have been suppressed by telephone line filtering. The objective was to apply this process without any prior knowledge of the original wideband signal since only the received bandlimited sound was used. One of the earlier schemes which was used in an attempt to meet this objective was the spectral duplication method in which the received in-band spectrum was translated, scaled and added to the received signal. This indeed produced a signal that subjectively appeared to have a wider bandwidth than the received signal but was heavily distorted and of very poor quality.

High frequency regeneration was then applied in a more signal specific manner by adding an extra formant to the received bandlimited voiced speech. This was performed by producing a wideband excitation signal from the received signal whereby the received signal was passed through a 'W' non-linear transfer function network. The excitation signal was then fed into a fixed resonant filter before being added to the received voiced speech. The added signal was controlled by a voicing decision based upon the first shift autocorrelation coefficient. The results of this configuration unfortunately proved to be disappointing in that the "added formant" sounded more like a spurious whistle during sustained voiced speech. It was considered that the system could well have been improved by including some form of adaptation to control the gain and the frequency of the formant. It is suggested that the adaptive algorithm could be based upon the in-band formant frequencies and their residues.

Our focus of attention was then directed towards high frequency regeneration applied to bandlimited unvoiced speech. Initially, bandlimited noise was added to all the unvoiced sounds and, as expected, this proved to show very poor results in output speech quality. Various refinements were then used in order to attempt to reproduce the high frequency spectrum of some of the unvoiced speech sounds by employing a combination of different non-linear transfer functions to suit some of the different unvoiced sounds in the data. The method that yielded the best results was that which applied spectral extension only for /s/ sounds. This used the W-transfer function in conjunction with the modulation of the wideband waveform envelope of the /s/ signal. The method was further improved when spectral shaping of the regenerated signal was included by using the resonant filter network previously employed for voiced speech. We of course acknowledge the fact that processing of the speech signal at the transmitter was now necessary to extract the waveform envelope and to decide upon the duration of the /s/ sounds from the original wideband speech, but the side information may be conveyed at the expense of a slight reduction in the 300 to 3400Hz bandlimited speech signal using a small spectral slot within that band.

In connection with this framework, we replaced the ideal 300 to 3400Hz channel with a digital medium and considered the available techniques for source coding the base band signal plus the waveform envelope of the /s/ sound, including the necessary signalling information. Generally, the DPCM systems were found to perform better than the PCM and adaptive delta modulation systems for

bandlimited speech. The digital codecs were also compared with and without the spectral enhancement schemes whereupon using the informal subjective tests, it was found that when the enhancement schemes were employed, a better quality speech signal was produced. This was found to be true for all the coding systems and more noticeably so for the low transmission rate delta modulators. This was because the quantization noise was masked by subjective impression of a broader band of frequencies in the output signal from the coding system plus the post-processor.

Our next course of action was directed towards the lower end of the spectrum. In this case, the spectrum was extended not by non-linear transfer functions but by low frequency oscillators driven at the pitch frequency and its second harmonic component. The pitch was discerned using the cepstrum method and another technique known as data reduction. The regenerated low frequencies were then spectrally shaped by a gain factor derived from the 300 to 1000Hz in-band spectrum and then by an analytical representation of the vocal tract response. Unfortunately, none of these methods produced an overall improvement in quality compared to the subjective impression of the 300 to 3400Hz bandlimited speech. There was indeed a perceptual presence of the base frequencies which are of course absent from the bandlimited signal. However, it was the 'blending' of these base frequencies that did not appear appropriate. Obviously what seems to be required is an improved regulating algorithm which controls the amplitudes of the synthesised low frequency components, assuming that the pitch frequency is estimated fairly accurately. The procedure of pitch detection warrants an exclusive research programme by itself due to the nature of the problem.

7.3 Spectral Compression Techniques

In Chapter VI we considered to invoke fully the transmitter terminal for processing the wideband input speech signal to enable spectral compression to be performed. In the spectral extension techniques of Chapter V, the transmitter was used only to extract a minimum of information from the wideband speech signal and to transmit this alongside the telephone bandwidth speech signal. Here we have permitted ourselves to convert the wideband signal into a new signal which occupies the telephone band. We then applied an inverse process at the receiver to reconstruct the wideband signal.

The first of these compression schemes was the voiced/unvoiced switching system in which the voiced speech was left unprocessed and allowed to be transmitted directly via the 300 to 3400Hz channel. The unvoiced speech was bandlimited to 3000 to 6000Hz and heterodyned down to 300 to 3400Hz before being transmitted via the channel. At the receiver, the corresponding inverse operations were applied. The technique was the simplest comparatively under this category and the output signal did show a subjective impression of a wider bandwidth than the 300 to 3400Hz actually used for the transmission medium. However, the unnaturalness associated with the unvoiced sounds prompted us to elaborate upon the process; this required us to include the 300 to 3400Hz band of frequencies of the unvoiced sounds as well as the upper band and still conform to the spectral compression requirements. In order to meet the objective, we turned our attention to processing in the frequency domain using the FFT algorithm. As block analysis was now performed, we reviewed

the consequences of applying the DFT process to the data using different window weighting functions and then selected the windows that were most suited to our application.

The first technique in the frequency domain that we examined was the second-order spectrum process where the forward Fourier Transform was applied twice successively to each block of data. It was argued that if this technique had produced a concentration of the signal components in the lower region of the second order spectral domain then one could apply a smooth truncation procedure to economise upon the number of symbols transmitted. Again the reverse operations were carried out at the receiver terminal to reproduce the wideband speech. When the process was applied to voiced speech as well as unvoiced speech, it was found that the intelligibility and quality were severely degraded when half the second order spectral coefficients were discarded (corresponding to the rejection of the upper range). This was because the arguments upon which the technique was founded are not valid for real speech signals. This perhaps was due to the spectral fine structure of voiced and unvoiced speech giving situations in the first order spectral domain which inhibited a concentration of the components in the second order spectrum domain. If this process is to be adopted for future research then we suggest using homomorphic techniques to separate the speech excitation parameters and apply the second order spectrum compression method to the vocal tract response only.

The next frequency domain notion to be developed was a system termed 'frequency mapping'. The principle was to select the frequency components from the signal spectrum that were perceptually the more

significant i.e. those components which constitute the formants. The frequency components corresponding to spectral troughs were decimated whereby say, only every second or third component was selected, the others being discarded. The selected components were then mapped or translated into a lower frequency band such that the highest location of the components corresponds to 3400Hz. The inverse DFT was then applied and the time domain signal was transmitted to the receiver via the bandlimited channel.

When the receiver recovered and relocated the frequency components to their original positions, the spaces between the decimated samples were filled with linearly interpolated frequency components between adjacent transmitted components. The inverse DFT was then applied to generate the wideband output speech signal. The frequency mapping process was applied to unvoiced sounds only, with a mapping characteristic that was particularly suited to the /s/ sounds. This produced a moderate quality output signal for all sounds since unlike the VUBS method, the lower frequency range of the unvoiced sounds were now represented although more coarsely than the upper band. The frequency mapping system was further refined by incorporating an adaptive procedure in which the compression algorithm was rendered more signal specific. Four different mapping characteristics were employed to compress the unvoiced speech sounds and each mapping law was formulated to suit one or more of the unvoiced spectral characteristics from the data used. During the operation of the algorithm, the mapping law used for application on a particular block of unvoiced speech samples was selected by testing all of the mapping laws available and finding the one that

yields the minimum spectral error (the best spectral-SNR). A code representing this choice was signalled to the receiver where the inverse frequency mapping procedure was applied to reconstitute the wideband block of unvoiced speech samples.

The results of this process showed a better performance than that gained from the fixed frequency mapping case particularly for the onset and decay of sustained unvoiced signals such as the /s/ sounds. This gave a subjective impression of a smoother transition between bandlimited voiced speech and the wideband unvoiced speech and thus the processed signal was informally considered to be of good quality compared to our previously discussed compression systems. The other unvoiced speech sounds, such as the stop consonants, also seemed to sound subjectively better when processed by the adaptive mapping strategy. This supports the notion that applying various mapping laws that are tailored to suit these unvoiced sounds offers some subjective improvement in output speech quality.

Our next frequency domain compression technique involved processing the phase spectrum rather than the modification of the magnitude spectrum. For /s/ unvoiced sounds, the phase spectrum was set to zero at the transmitter and replaced with a uniformly distributed random phase at the receiver. Spectral compression was attained at the transmitter by transferring the upper half of the magnitude spectrum into the phase array then applying the inverse DFT to transmit the time domain signal. At the receiver, the phase array samples were transferred back into the upper half of the magnitude

spectrum; the magnitude spectrum now having random phase values. The signal was then converted back into the time domain by the IDFT. When the processed signal was combined with the bandlimited non-/s/ speech to yield the output speech signal, the subjective quality appeared to be fairly good and gave an impression that a wider than telephonic bandwidth transmission channel had been used to convey the signal. As expected, the quality of the processed signal did not appear to be as good as either of the frequency mapping schemes; this is probably the penalty paid for the fact that the phase processing scheme is slightly simpler insofar as the number of operations required is less than that for any of the frequency mapping algorithms. The frequency mapping schemes also preserve the phase of the spectral energy peaks where the perceptual effect is the greatest.

After these investigations into frequency domain compression techniques, we then revisited the time domain by considering a process termed "Time Domain Harmonic Scaling" proposed by Malah.[ref 2] This involved a relatively simple interpolation operation between speech segments of one pitch interval apart which imposed a linear spectral compression to the speech signal. At the receiver, linear spectral expansion was obtained by formulating two pitch intervals of output speech for every three pitch intervals of input speech. The process was repeated by shifting the input sequence at the receiver by one pitch interval so as to formulate the next two pitch intervals. The net result was to generate two pitch intervals of output speech for each pitch interval of received speech. In our experiment, an arbitrary block length was used

instead of the pitch intervals for unvoiced speech and we simply bandlimited the voiced speech from 300 to 3400Hz. The result was to compress linearly the spectrum of wideband unvoiced speech and combine this with the bandlimited voiced speech. The effect of this scheme on our informal subjective listening experiences indicated that an impression of received wideband speech existed the same as the previous systems but there was a characteristic unnaturalness associated with the processed unvoiced sounds in the form of a "buzzy noise". We had already postulated that this might be due to the fact that the derivation of the time domain scaling algorithm depended upon the regularly spaced harmonic structure of the input signal. Voiced speech conforms to this nature, at least in the short term sense, but unvoiced speech generally does not. The resultant effect of the algorithm is possibly to produce a more regular harmonic structure into the processed unvoiced speech which apparently generates the buzz sound. We did not, however, examine this artifact any further (experimentally or analytically) but it certainly would be considered a worthwhile suggestion for future investigation with the intention of deriving a suitable remedy for improving upon this situation.

Finally, as with the quality enhancement systems, the ideal 300 to 3400Hz bandwidth channel filter was replaced by a digitally coded channel and again different source coding techniques were considered that might apply to the spectrally compressed signal. The input signal to the digital coder was different from the bandlimited transmitted signal in Chapter V as in this case spectral compression was applied to the unvoiced sounds. They therefore possessed

different characteristics to normal bandlimited speech; and it was because of this that the coders needed to be re-optimised from their nominal operating conditions. Once this had been done for the four types of digital encoding methods (i.e. APCM, ADPCM, CVSD and CFDM), the output speech from the decoder was post-processed by one of the respective algorithm, viz VUBS, fixed frequency mapping and adaptive frequency mapping. The final output speech from these systems was compared to speech digitally encoded at the same transmission bit-rate. The input speech signal to the encoder was in one case bandlimited from 300 to 7200Hz, and in the other case it was limited from 300 to 3400Hz. Using the transmission standards of 16, 24 and 32 kbits per second, we found that for all types of input signal, the quality of the output speech worsened as the transmission rate decreased. Generally, the output speech from the coders using the telephone bandwidth input signal was found to be less noisy than the coders using the wideband input signal. This was because the wideband speech required a higher sampling rate, hence a reduced number of bits per sample were available to quantize the signal. This gave rise to a greater amount of granular noise. The problem of using a higher sampling rate for any of the bandwidth compressed signals did not arise when allowing the same quantizing accuracy as used for the bandlimited input signals. For this reason, there was less granular noise in the post processed signals than in the wideband signals; also the apparent bandwidth of the post processed signals was greater than that of the bandlimited signals. It was found then that the quality of the output speech from the coding systems using bandwidth compression was better than that of the output speech from the coding systems without using any bandwidth

compression. This was most notable for the case of the 16kbit/s CVSD system. When this digitizer was used to encode bandlimited speech, the resultant quality was comparatively noisy; if wideband speech was used then the coder failed to track the high frequency unvoiced sounds. When bandwidth compression was used, however, the subjective impression of the presence of the unvoiced sounds tended to mask the impression of the granular noise during the silence intervals. This result was true for all three of the aforementioned bandwidth compression algorithms used in conjunction with the encoding systems. This gives a positive indication towards the worthwhile application of the bandwidth compression systems discussed in this thesis.

7.4 Methods of Appraisal

7.4.1 Spectral-SNR Measurements

The segmented spectral-SNR has already been introduced and discussed in detail in section 6.4 when this was applied to the phase processing bandwidth compression system. There is of course no reason why we cannot apply this comparison procedure to all of the bandwidth compression and expansion algorithms in this summing up stage. A 256 mS specimen of the input speech corresponding to the sound /s/-/ka/ from "S K Harvey" was processed by the eleven schemes listed in table 7.1. The input speech comprises of a fricative, a stop consonant, a silence pause and a vowel and was sufficiently short enough so as not to consume an exhaustive amount of computer execution time. In all of the eleven cases, the transmitted signal

has been bandlimited at some stage from 300 to 3400Hz but the output signal in each case is compared with the 300 to 7600Hz original signal. We have also included the telephonic bandlimited speech in row (4). The method used to compute the segmented spectral-SNR is precisely the same as that detailed in section 6.4.4.4(c) and the values obtained for each of the systems are ranked in increasing magnitude in column 3 of table 7.1.

The first point to note is that without exception, the bandwidth compression systems out-performed the bandwidth extension systems by this measurement criterion. The fidelity of the reproduced wideband signal from the bandwidth compression systems is most possibly due to the fact that the system has an à priori knowledge of the original wideband signal, whereas in most cases, the bandwidth extension schemes do not. The price paid for this is of course the requirement of the transmitter processing for the compression systems as well as receiver processing.

The second point is that if the SP-SNRSEG value for the 300 to 3400Hz bandlimited speech is taken as a "datum level" then three of the systems listed above that are shown actually degrade the reproduced speech below the datum whereas the other seven schemes show some improvement in performance beyond the fidelity afforded by the bandlimited speech.

Table 7.1 Comparison of Segmented Spectral-SNR Values

PROCESS	COMPONENTS PROCESSED	SP-SNRSEG (dB)
(1) HFR by adding bandpass filtered noise controlled by in-band signal (section 5.1)	All speech plus silence intervals	12.26
(2) HFR by adding bandpass filtered noise at fixed level (section 5.4.1.1)	Unvoiced speech only	15.19
(3) Low frequency synthesis by local oscillator driven at pitch frequency (section 5.6.1)	Voiced speech only	16.02
(4) 300 to 3400Hz bandlimited speech	All the signal	16.16
(5) Extra formant synthesis at fixed frequency and amplitude (section 5.3)	Voiced speech only	16.60
(6) HFR for /s/ sounds using W-function plus spectral shaping (section 5.4.1)	/s/ sounds only	16.83
(7) Time domain harmonic scaling (section 6.5)	Unvoiced sounds only	17.57
(8) Random phase processing (section 6.4.4)	/s / sounds only	17.58
(9) Fixed frequency mapping (section 6.4.3)	Unvoiced speech	18.11
(10) Adaptive frequency mapping (section 6.4.3.5)	Unvoiced speech	18.58
(11) Voiced/unvoiced band switching system (section 6.3)	Unvoiced speech	18.92

A final point to note here is that although we have endeavoured to include a variety of different speech sounds in our short specimen of data, it may well be argued that there will be an inconsistency inherent in comparing the SP-SNRSEG measurements as not all of the methods process the same segments of speech (indicated in column 2). However, one reason for applying such a strategy is to attempt to relate the comparative SP-SNRSEG values with our own valuations based upon informal subjective listening experiences of the entire 10-word sequence obtained in section 4.2.

7.4.2 Informal Subjective Listening Tests

We have already discussed our impressions of the listening tests conducted on recordings of the processed material throughout Chapters V and VI for the individual algorithms and also we have presented comparisons whilst reviewing the development of the algorithms in section 7.2 and 7.3. Here we wish to briefly summarise these comparisons against the ranking of the system performances from the SP-SNRSEG values and the time waveform SNRSEG values.

Based upon informal listening experiences by a small number of colleagues, the processes are listed on table 7.2 in descending order of preference. Our impressions were discerned more by listening quality performances rather than the intelligibility comparisons of the processed signals. Both the listings of the time waveform segmented SNR and the spectral SNR values show a decreasing trend, a trend which coincided with our subjective listening preferences.

Table 7.2 Comparison of Informal Subjective Listening Tests

Order of Preference	Time waveform Segmented-SNR (dB)	SP-SNRSEG (dB)	Process
1	15.68	18.58	Adaptive frequency mapping (section 6.4.3.5)
2	15.74	18.11	Fixed frequency mapping (section 6.4.3)
3	14.35	16.83	HFR for /s/ sounds using the W-transfer function plus spectral shaping (section 5.4.1)
4	13.99	17.58	Random phase processing (section 6.4.4)
5	13.93	17.57	Time domain Harmonic scaling (section 6.5)
6	17.69	18.92	Voiced/unvoiced band switching system (section 6.3)
7	14.68	16.16	300 to 3400 bandlimited speech
8	14.42	16.60	Low frequency synthesis by local oscillator driven at pitch frequency (section 5.6.1)
9	14.28	16.02	Extra formant synthesis at fixed frequency and amplitude (section 5.3)
10	9.59	12.26	HFR by adding bandpass filtered noise controlled by in-band signal (section 5.1)
11	13.44	15.19	HFR by adding bandpass filtered noise at fixed level (section 5.4.1.1)

However, in neither case are the trends monotonically decreasing, the most notable exception being that of the VUBS system (item 6) which produced the highest segmented SNR and spectral-SNR values over all the other systems listed. This may be because although the VUBS system preserves the most significant region of the spectrum (i.e. where the bulk of the signal energy resides), the spectral region with the lower signal energy is discarded and this could well be perceptually more detrimental than the simple VUBS algorithm would suggest. In retaining the spectral region with the highest signal energy, the time waveform is correspondingly well preserved as noted by the time waveform segmented-SNR, therefore, neither measurement bears a close relation to the relative subjective performance of the system.

Another example worth mentioning is item 10, i.e. the high frequency regeneration by adding bandpass filtered noise controlled by the 300 to 3400Hz received signal. In this case, the process is applied for the entire received signal during silence pauses and voiced speech as well as during unvoiced speech. By doing this we have impaired the signal time waveform and the spectrum to a greater degree than for any other scheme in table 7.2. However it was not shown to provide the least subjective impression. This was likely because varying the level of the added noise was found to produce a signal that was subjectively more "listenable" than that obtained by keeping the added bandlimited noise level fixed (item 11). The intelligibility of item (10) appeared to be very high in contrast to the degraded signal from process (11).

Having presented the SNR values and a brief indication of the informal subjective listening scores for the systems investigated, we cannot at this stage conclude whether the signal to noise ratio measurements based on the time waveform or on the short-time spectrum give a clearer indication to the perceptual impression of the processed signal. Obviously to arrive at any firm conclusion, what really needs to be done is to perform measurements upon a vast quantity of processed data and to conduct formal subjective listening tests on that data. One suggestion which may well be worthwhile is to weight the signal spectrum in accordance with a perceptual criterion before taking the spectral-SNR measurements. This could give the spectral readings the "edge" over the time domain segmented SNR measurements when attempting to correlate them to the subjective listening performance of the processes involved. If this is so then it would go a long way in justifying the additional processing involved in forming the SP-SNRSEG values over that involved in measuring the time waveform segmented SNR.

7.4.3 Spectrographic Displays

Another means of discerning the relative performances of the systems discussed is to inspect the temporal variation of the short-time power spectrum. This can be achieved by the utility of a spectrographic display (or spectrogram). To print out a spectrogram of a section of recorded speech material usually requires some specialised hardware; since no such device was available in the Department of Electrical Engineering at Loughborough, it was decided to attempt to produce a fairly approximate form of spectrogram display using the ICL 1900 computer lineprinter employing the character overprinting facility.

The procedure adopted for this process was to use a sliding block strategy and then to calculate the short-time power spectrum with the DFT. This is an alternative, although less flexible, approach to that of using a frequency sweeping digital filter.(refs 95-97) The method used to process our displays is shown in the flow chart of figure 7.1. The data is first segmented into a block of samples, the blocksize determining the approximate frequency resolution of the final display. In order to improve upon the accuracy of the spectral estimation by the DFT, a Hanning weighting window function is applied to the data as explained in section 6.4.1. The magnitude coefficients are converted into a dB scale in order to enhance the values of the less significant frequency components (e.g. those residing between the formant peaks of voiced speech). The natural roll-off of the voiced speech spectrum was compensated by an equaliser having a linear frequency response with a positive gradient. The equaliser was implemented by applying a 6dB per octave lift to the frequency components of the spectrum above 1kHz and leaving the components below 1kHz unaltered. This was found to bring the formants of the data roughly to the same height.

When using spectrogram machines, one of the options that is available is the type of analysis required, i.e. whether the analysing filter should be of a wideband or narrowband limits; the typical specification might be 200Hz and 50Hz respectively. The advantage of using a wide bandwidth analysis filter is that the time resolution is increased compared to the narrowband filter, conversely, a narrowband filter would display a finer frequency resolution than its wideband counterpart. In order to observe the

pitch striations of voiced speech, a narrowband analysis would almost certainly be required. In our algorithm, we have attempted to duplicate crudely this facility of the wideband option by forming the spectral envelope of the frequency components. For the narrowband case, the spectral components are left unaltered. The spectral envelope is formed by selecting the local peaks in the spectrum (i.e. those points connecting a positive gradient on the left hand side and a negative gradient on the right hand side). The frequency samples between the peaks are then linearly interpolated. After this process is performed for the wideband analysis (or not performed in the case of narrowband analysis) the lower levels of the spectral values are bottom clipped by a preset marking depth (usually around 30-60dB). This is such that very low levels of frequency values, having large negative dB values will be ignored.

Now that we have a finite range of values, the separation between these levels is emphasised before displaying the spectral line. This is achieved by raising these values to the third power, i.e. a cubic transfer function is applied. In order to plot the line representing the spectrum, the cubed values are divided into ten amplitude regions using eleven equally spaced thresholds. This includes the upper and lower limits of the spectral values to be plotted. If a particular value happens to fall within the lowest amplitude region, then it will be represented as a space character (or blank) on the line printer. If a value falls within the next region then this is represented as a '.' character. For a spectral value residing in the third region, this is shown as a '.' overprinted by a '"' character, and so on. For a value occupying

the uppermost region, this is written on the line printer by overprinting nine characters. In this way, a two dimensional plot is formed to represent the spectral section on the line printer. The vertical direction represents the frequency axis and the relative blackness of the line represents the intensity of the frequency coefficient.

In order to incorporate the third dimension, i.e. the time axis, a sliding block strategy is used. The whole process is repeated on a block of data that is formed by shifting the new block along the signal by a preset time interval with respect to the first block. This time shift is termed the time resolution of the spectrogram. Incidentally, the frequency resolution is constrained by the displayed frequency range and the number of line printer characters per line, this being an inherent limitation of the DFT process. The minimum time resolution is set by the sampling interval of the data which corresponds to having a sliding block shift of one sample per displayed line.

After the next line has been printed, the process is repeated until all the data is processed. In this way, a three dimensional image is built up of the frequency/time/amplitude distribution of the input signal. Figure 7.2(a) shows a displayed version of the utterance /s/-/ka/ from "S K Harvey" using a wideband analysis to display the formant trajectories, and figure 7.2(b) shows the distribution of the same segment using narrow-band analysis to roughly display the pitch striations. The input signal was taken from our original speech data of 300 to 7600Hz bandwidth.

Apart from the problem of the limitations imposed upon the frequency resolution displayed, it was found awkward to indicate a frequency range much greater than from dc to 4kHz. This was because our equalisation strategy did not seem adequate for frequencies above this value for our data. If one considers the spectrum, for example of a voiced /I/ sound from the data (shown in figure 5.14(b)(i)) then it can be seen that the spectral energy comprising the voiced speech is mostly confined to below 4kHz. Its trend has a steady roll-off which is tolerated by the equaliser. Unfortunately, above 4kHz the spectral tilt flattens, which may be due to amplifier and ADC noise. If our equalisation strategy were used on the entire spectrum then this noise in the frequencies between 7 to 8kHz would be prominent in the final display due to the equaliser lift. This would give a completely unnatural distribution. The remedy to this would be to use perhaps some form of adaptive equalisation which would not destroy the original spectral distribution of the wideband speech (for both voiced and unvoiced sounds) and also prevent the hf noise from swamping the display.

Our final method for evaluating the results of the systems investigated was by plotting spectrographic displays from a custom machine designed to produce spectrograms. The displays were obtained by a SONAGRAPH spectrogram machine(ref 98) by courtesy of the Speech Pathology Department at Scraptoft Polytechnic, Leicester. These were used to give a clearer and more decisive output for comparison than the line-printer plots of figures 7.2(a) and (b). The Sonagraph spectrograms are all shown in figure 7.3; it is pointed out that since these are only photographic copies, the inherent high contrast gain and the inconsistency of reproduction must be taken into account when comparing the plates.

The original signal is shown in sub-figure (a) for the utterance "fist". The signal is displayed from 0 to 8 kHz and is linearly scaled in frequency by the calibration bar on the left hand side. It can be seen that the /f/ signal appears very feint due to its comparatively low amplitude.

The 0.3 to 3.4 kHz version of the signal "fist" is shown in sub-figure (b) where it can be seen to be absent from frequency components above 4kHz. There is some overload distortion at the centre of the /I/ sound due to the incorrect setting of the recording level of the signal.

Sub-figure (c) shows the 0.3 to 3.4 kHz transmitted signal of the Voiced/Unvoiced Bank Switching system (VUBS), here the heterodyned /s/ signal now replaces the bandlimited /s/ signal. In sub-figure (d), the /s/ sound re-occupies the 3 to 6 kHz range in the output signal from the VUBS receiver.

The output from the random phase processor is shown in sub-figure (e) where some of the lf components of the /s/ signal appear to be present. The signals from both the random phase processor and the VUBS method appear to show that the /t/ is bandlimited in comparison to the original signal in figure 7.3(a).

The transmitted signal from the fixed frequency mapping algorithm is shown in sub-figure (f) where unfortunately the recording level perhaps is too low to offer a clear picture to represent this signal. When the marking depth of the spectrogram was increased then this produced a noisy background.

The output signal from the FMAP receiver is shown in sub-figure (g) where it seems that the hf components of the /s/ signal are again produced well.

For the output signals of the systems shown so far, the abrupt switching between the hf components of the /s/ signals and the bandlimited voiced signals seems evident. This effect is seen to be alleviated by the adaptive frequency mapping algorithm, the output signal from this latter method is shown in sub-figure (h). Here the signal seems to compare more favourably with the original signal than do any of the other signals as it shows a gradual blending when the signal changes from voiced to unvoiced speech and vice versa. It also seems to reproduce the hf components of the /t/ sound.

The next two plates show the original signal and the output signal for the word "sister" processed by the high frequency regeneration system of section 5.4.3. It can be seen that the hf components of the /s/ sounds appear to be synthesized with a relative amplitude distribution similar to the original except for an attenuation factor. The added hf components were attenuated in order to make the output signal were "listenable".

Finally, the results for the baseband synthesis system using a single oscillator driven at the pitch frequency are shown in sub-figures (k), (l) and (m) respectively for the original, bandlimited and processed signals. The frequency scale displayed is now 0 to 4 kHz which has been linearly divided into 500Hz steps by the calibration bar on the left hand side of each picture. The sounds represented are the utterances "sister" and "father" where

the original signal is lowpass filtered from 0 to 7.6 kHz, the bandlimited signal is filtered between 0.3 and 3.4 kHz and the processed signal occupies a bandwidth of 0 to 3.4 kHz. Although there is some spurious lf components shown on all the three plates, it should be possible to discern that the attenuated fundamental component in the bandlimited example of sub-figure (ℓ) seems to be replaced by a low frequency tone in the processed signal of sub-figure (m). This was indeed supported by our informal subjective listening tests in section 7.4.2.

7.5 Closing Remarks

In this thesis we have discussed techniques to enhance the quality of speech that is required to be conveyed via a telephonic bandwidth channel. The techniques described have been categorised into bandwidth compression/expansion methods and bandwidth extension processes. All of the methods described are based upon source and/or sink coding, rather than using bit rate reduction techniques at the channel encoding stage. Some of the techniques have in fact used the readily available digitiser strategies such as APCM, ADPCM and ADM.

We have developed the systems described in an attempt to overcome one of the limiting factors imposed upon telephone speech. Although none of our methods have yet been tested upon "real" telephone lines using extensive formal conversational speech, we would like to believe that our research has laid down some foundations for applications in this area, particularly with the audio-teleconferencing systems and the "phone-in" speech post-processing described in Chapter II.

ERRATA

E.1 Section 5.3.3.

It is acknowledged that the type of filter used to synthesise an additional formant for voiced speech was not implemented as intended. This is shown by the filter response curve in figure 5.12(a) which displays a very narrow peak. The effect of the sharpness of the peak was evident in the informal subjective listening tests of the processed speech; i.e. the added formant sounded like a spurious whistle. The reason for this was possibly due to parameter error in the program. The intended filter response should have been two proximate but resolvable poles with a greater bandwidth than that in figure 5.12(a). It is believed that this would have produced a better quality processed signal, however, time did not permit us to investigate this.

E.2 Section 5.4.3.

The filter program was also used in section 5.4.3 to shape the spectrum of the regenerated high frequencies of the /s/ spectrum. Again the filter response curve in figure 5.30(a) shows that it was not implemented as intended. The spectral peak at about 3kHz should not have been present, however, the effect upon the processed signal did not appear to be too detrimental. The required filter response curve was expected to have a single broadband peak at 4.8kHz with monotonically decreasing slopes. The effect of this filter upon informal listening experiences of the processed speech was not investigated.

Appendices

In this section, we propose to itemise some of the subroutines that have been frequently used throughout the experiments performed in this project. We will briefly review the background of the subroutine followed by a listing of the same. We will not discuss the details regarding the full computer program of each of the experiments nor the methods used for the computer job control and file management; we only wish to present a 'flavour' of our programming structure. It is presumed that the interested reader who wishes to continue with any of the experimentation will be able to originate his own computer programs from the thesis description to suit the machine that is available to him.

A.1 Digital Filtering (ref 99)

The filtering operations that are performed by the algorithms discussed in this thesis are implemented as programmed digital filters (except for the anti-aliasing filters applied before and after analogue to digital, and digital to analogue conversion respectively). Some of the advantages of digital filters are: (a) accuracy, (b) flexibility and (c) freedom from drift.

Two types of digital filter have been used in this thesis i.e. the recursive and non-recursive types. The non-recursive filter was generally used to model an ideal telephone channel where we preferred to know the delay characteristics of the signal exactly. When this consideration was not important, the recursive filter was applied which is more economical in terms of computing time.

A.1.1 Recursive Filter Design

The recursive filters used in our simulations were designed with a Butterworth characteristic (Rader and Gold 1967). The gain characteristic of an N_B^{th} order Butterworth filter is given by the expression

$$|H(j\,2\pi ft)| = 1/\left\{1 + \left[\frac{\mathrm{Tan}(\pi fT)}{\mathrm{Tan}(\pi f_c T)}\right]^{2N_B}\right\}^{1/2} \qquad (A.1)$$

where T is the sampling interval

For frequencies less than the cut-off frequency f_c, $[\mathrm{Tan}\,(\pi fT)/\mathrm{Tan}(\pi f_c T)]$ is fractional and the gain is approximately unity. For frequencies exceeding f_c, where $(\tan \pi f\, T/\tan \pi f_c T)$ becomes large, the gain is close to zero. The higher the value of N_B, the order of the filter, the better is the approximation to the ideal brickwall lowpass characteristic. The gain of a Butterworth filter falls uniformly as the frequency is increased from zero to one half the sampling frequency. At the cut-off frequency, the gain lies at precisely 3dB below its value at zero frequency.

A Butterworth filter of N_B^{th} order has N_B poles which lie on a circle in the z-plane. The poles are given by the value of β_m which fall within the unit circle where the real and imaginary parts, U_m and V_m are given by

$$U_m = (1 - \tan^2 \pi f_c T)/d$$

$$V_m = 2 \tan \pi f_c T \sin(m\pi N_B)/d$$

where $$d = 1 - 2 \tan \pi f_c T \cos(m\pi/N_B) + \tan \pi f_c T \qquad (A.2)$$

If N_B is even, $m\pi/N_B$ should be replaced by $\pi(2m + 1)/2N_B$ in (A.2). An N_B^{th} order Butterworth filter has N_B zeros which are all situated at $z = -1$, i.e. at half the sampling frequency.

The Fortran subroutine BUTTER(ref 99) which is given in List 1, computes the poles and zeros of a Butterworth filter of specified order N and cut-off frequency FC. After execution of the Fortran statement CALL BUTTER (N, FC, ALPHA, BETA), the complex arrays ALPHA and BETA contain the zeros and the poles of the required transfer function. The arrays COEFF in List 3 which will then yield the coefficients of cascaded second order transfer functions in a serial realisation of the Butterworth lowpass filter.

A serial form of cascaded second order sections is used such that the order of the denominator of the transfer function of each section never exceeds two. The reason for this is that for higher values of N_B the digital filter which results can be excessively sensitive to the effects of arithmetic rounding error. Small variations in the denominator coefficients produced by the finite word length of the computer can produce surprisingly drastic changes in the transfer function poles, and can even result in filter instability.

The general structure of a stage in the serial formation of the filter is shown in figure A.1. The section has a transfer function of:

$$H(z) = a_0 \left[\frac{1 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}} \right] \qquad (A.3)$$

which produces two zeros and two poles in the z-domain. The gain coefficient, a_0, is determined by applying a normalising boundary condition which will be discussed shortly.

A.1.1.1 Frequency Transformations

A frequency transformation (Constantinides 1970), enables a filter of one type to be transformed into some other type. In our case, we design a prototype lowpass filter and then apply a lowpass-to-bandpass frequency transformation from the lowpass filter to produce a bandpass design. If a bandpass filter with lower and upper cut-off frequencies f_1 and f_2 is required then the design is produced by taking a lowpass filter whose cut-off frequency f_c is $f_2 - f_1$ and replacing z^{-1} in its transfer function by the expression $-z^{-1}(z^{-1}-a)/(1-az^{-1})$. The parameter a is given by the formula

$$a = \cos\{\pi(f_1 + f_2)T\}/\cos\{\pi(f_2 - f_1)T\} \qquad (A.4)$$

The transformation results in a filter whose order is double that of the prototype. If the lowpass-to-bandpass transformation is applied directly to a second order section of a lowpass prototype, a fourth

order second results. To avoid the problem of numerical accuracy, a final design consisting of second order sections is required. This difficulty can be resolved by applying the transformation to the poles and zeros of the lowpass prototype. If the lowpass prototype has a zero at $z = \alpha$, then the bandpass filter must have zeros at values of z which satisfy the equation

$$\alpha^{-1} = -z^{-1}(z^{-1}-a)/(1-az^{-1}) \qquad (A.5)$$

This is a quadratic equation in z which can be solved by the usual quadratic equation solution. The solution shows that the bandpass filter must have zeros at

$$z = \frac{1}{2}(1+\alpha)a \pm \left\{\frac{1}{4}(1-\alpha)^2 a^2-\alpha\right\}^{1/2} \qquad (A.6)$$

A similar formula gives the poles of the bandpass filter in terms of the poles of the lowpass prototype. Subroutine BPASS (List 2) accepts as its input the poles and zeros of a lowpass prototype in arrays as delivered by the subroutine BUTTER. It computes the poles and zeros of the resulting bandpass filter by the use of expression (A.6). The bandpass filter is then realised as a cascade of second order sections by the use of subroutine COEFF (List 3) operating on the output of subroutine BPASS.

A.1.1.2 Gain Normalisation

In order to specify the filter gain at a particular frequency, it is necessary to fix the value of a_0 in expression (A.3) according to the required boundary condition. For a multiple stage serial filter

(i.e. a filter whose order N_B is greater than 2) the gain coefficients, a_o, are combined into a single gain coefficient which preceeds the whole filter network.

Initially the gain coefficient is set to unity and then the boundary condition is applied whereby it is required to set the mid-band gain of the filter to zero-dB (or unity gain) by the use of the expression

$$H(z)\Big|_{z=\exp(j\omega_m T)} = G \qquad (A.7)$$

where $H(z)$ is the transfer function of the complete filter

$$\omega_m = 2\pi (f_2 - f_1)$$

and T is the sampling interval.

The value of a_o is then given by $1/G$ such that the filter now conforms to the specified boundary condition. If it is required to set the d.c. gain of a lowpass design to zero-dB then the above expression is applied by allowing $\omega_m = 0$, i.e. $z = 1$.

Now that the filter has been designed, the gain and phase response of the filter may be computed by a call to subroutine FREQ (List 4) which evaluates the transfer function $H(j\omega_k T)$ for a number of frequency values of ω_k; 129 in our case.

The magnitude and phase characteristic for a 300 to 3400Hz filter employing 15 cascaded second order stages (i.e. a 30^{th} order filter) is shown in figures A.2(a) and (b). The sampling frequency is 16kHz. It can be seen that the mid-band gain is in fact zero dB with transition bands that rapidly roll-off monotonically from the passband. The phase response is plotted for values between $\pm 180^{o}$ and as expected, it is non-linear especially at the transition bands.

So, to summarise, the design of the filter is implemented by specifying the cut-off frequencies, the sampling frequency and the order of the filter which is equal to the order of the filter for a low-pass design and equal to half the order of the final bandpass design. Subroutine BUTTER computes the poles and zeros of the lowpass prototype; subroutine BPASS applies the necessary frequency transformation to those poles and zeros followed by a call to COEFF to yield the coefficients of the realisable filter. Finally a call to FREQ checks the gain and phase response of the filter design. The complete design of the filter is performed by a call to subroutine SETUP (List 5) which applies all of the above procedure.

Once the filter coefficients have been computed they are stored in function HIIR (List 6) and subsequent inclusions of the function in the main program perform the filtering process required. The variable SIGNAL represents the input sample on entry and the filtered sample is represented by the value of HIIR on exit.

A.1.2 Non-Recursive Filter Design (ref 100)

Recursive filters are adequate in most purposes for filtering speech signals. However, if it is important to minimise the waveform distortion produced by the filter then the non-linear phase characteristics of recursive filters can result in an unacceptable modification of the waveform. A non-recursive filter can easily be designed to give completely linear phase characteristics and hence produce the minimum modification of the signal waveform compatible with a given gain characteristic. An essential feature of a non-recursive filter is that its weighting sequence is of finite length. The filter, therefore, has a finite memory so that the effect of a large spurious input such as a switching transient disappears completely after a known length of time. Also the delay of the input sequence can be exactly compensated if the filter is used in one branch of a systen and not another. The other advantage of non-recursive filters is that, because there is no feedback involved, there is no possibility whatsoever of instability occurring.

The principal disadvantage of non-recursive filteres as compared with recursive designs is that to achieve a given specification they generally require more computer memory and more arithmetic operations per clock pulse than recursive filters. However, for filters of large orders, fast convolution techniques using the fast Fourier Transform (Gold and Rader 1969) can speed up the filtering process.

A.1.2.1 Design by the Window Method

This method enables non-recursive filters to be designed to produce a frequency characteristic as close as required to an arbitrary specification. To start with, the method choses the ideal frequency response function to which an approximation is to be realised by considering N equispaced frequency points from dc to the folding frequency. The passband frequency samples are then set to unity and the stop band samples are set to zero. Since this ideal brickwall filter has abrupt discontinuities between the pass and stop bands, the corresponding impulse response is of infinite duration. If the impulse response is truncated to produce a non-recursive realisation of the filter then there will be an error or overshoot in the frequency response in the vicinity of the discontinuity. This is usually referred to as the Gibbs phenomenon. If the ideal frequency response is smoothed by the multiplication of a window function then careful choice of the window can result in a frequency response function with appreciably less in-band and out-of-band ripple.

A.1.2.2 Design by the Sampling Method

Design of non-recursive filters from frequency response specifications has been considered by Martin, noted in ref.100, who specified initial values of the frequency response at selected frequencies, leaving unspecified values of the frequency response in pre-selected transition bands. He then used a minimisation procedure to solve for final values of the frequency response at equally spaced frequencies. The criterion used for the minimisation

was that the maximum deviation of the continuous frequency response from the ideal frequency response be minimised for both in-band and out-of-band frequencies.

When using the frequency sampling method for filter design, the designer need never be concerned with the impulse response. This is intuitively appealing for filters with long impulse responses where the fast convolution techniques can be used (using the fast Fourier Transform). The sampling procedure is capable of being exploited to yield an "optimum" filter as discussed above which results in a trade off between ripple height and transition bandwidth. Once the frequency samples, H_k, have been specified for $k = 0, 1, \ldots N/2$ then the interpolated (continuous) frequency response may be formulated(ref 100) as

$$H(j\omega T) = \frac{\exp\left[-\frac{j\omega NT}{2}\left(1 - \frac{1}{N}\right)\right]}{N} \cdot$$

$$\sum_{n=0}^{N-1} \left\{ \frac{H_k \exp\left(-j\,\frac{\pi k}{N}\right) \sin\left(\frac{\omega NT}{2}\right)}{\sin\left(\frac{\omega T}{2} - \frac{\pi k}{N}\right)} \right\} \qquad (A.8)$$

It is seen that $H(j\omega T)$ consists of a sum of elementary functions of the form $\sin(\frac{\omega NT}{2})/\sin(\frac{\omega T}{2}-\theta)$. In the design of a lowpass or bandpass filter, one needs to choose the value of the frequency samples in the transition bands according to some criterion to minimise the ripple caused by these elementary functions. It is possible that the ripples caused by the frequency samples within the transition bands can be made to cancel the ripples caused by the fixed samples. As the number of transition values is increased, one can produce a finer cancellation.

(a) The Minimisation Algorithm

To start with, most of the H_k values are preset and the remaining few (transition coefficients) will be varied until the maximum sidelobe is a minimum. Figure A.3 shows a typical specification for a lowpass filter. In this example, there are N_B samples preset to 1.0, M transition samples, and the remaining samples are preset to 0.0. The transition samples are denoted by T_1,T_2 and T_3. The following search procedure is then adopted

1. First, a one-dimensional search is conducted by setting say $T_3 = T_2 = 1$ and the value of T_1 in the range 0.0 to 1.0 is iteratively determined which gives the minimum value of the first sidelobe in equation (A.8). This value is labelled as point A in figure (A.4).

2. Two dimensions are now used in which T_2 is slightly perturbed from its preset value of unity (to a slightly smaller value) and the one dimensional search is repeated, varying T_1, as before. The two dimensional point (point B) along with point A determines the path of steepest descent along which to do the two dimensional search.

3. A search is made along the line found in step 2, yielding a minimum of $H(j\omega T)$ corresponding to point C. A new path of steepest descent is obtained by varying T_1 and keeping T_2 fixed at the value of point C, yielding point D; then perturbing T_2 slightly and again varying yielding point E. A further search is made to yield point F. If the difference between the values of $H(j\omega T)$ at the minimum of the searches along the lines of steepest descent (points C and F) is less than some prescribed threshold, the search is ended and point F is the two dimensional solution. Otherwise, the procedure of step 3 are re-iterated to yield refinements of the path of steepest descent until two consecutive searches yield minima whose difference satisfies the threshold condition. Practically it has been found(ref 100) that a two dimensional search always terminated within three iterations when the threshold was set to 0.1dB.

4. When three dimensions are invoked, $T_3 = 1.0$ and then is perturbed to a slightly lower value and steps 1 to 3 are repeated. This defines two points on a three dimensional line along which a search for a minimum is conducted. The search procedure is terminated when the difference in minima between two consecutive three-dimensional searches is less than a prescribed threshold.

The results of the method explained are published as a set of tables in the reference (100) and may be applicable as a "cook book" for designing the finite impulse response filters. The type of data selected for our use is taken from the table III page 93(ref 100) and is applied for the case where there are 256 impulse response coefficients (129 frequency samples) and 3 transition points. The transition values are presented here as table A.1 which shows that the highest sidelobe (minimax) can be made as low as between 85.5 and 113.1 dB below the passband gain. It may be noted that not all of the possible number of bandwidth points are tabulated and if other corresponding transition points are required then approximate values of these transition coefficients may be obtained by linear interpolation of the tabulated values. It is found(ref 100) that the deviation of the result obtained by linear interpolation will be less than 6dB from the optimum response.

(b) Bandpass Filters

The procedure discussed so far describes an optimum selection of transition coefficients for lowpass filters only. If similar optimum transition coefficients are required for bandpass filters then one may re-apply the optimisation procedure for such a case, bearing in mind that twice as many transition coefficients are required as there are two transition bands in the filter response. It will generally be found that the coefficients in each transition band will not be symmetrical, although one may run the optimisation process by choosing symmetrical coefficients and perhaps obtaining a slightly sub-optimum result.

A second approach in the design of bandpass filters is to define sub-optimum bandpass filters, which are derived very simply from the lowpass prototypes by appropriately rotating the lowpass frequency samples (including the optimised transition coefficients) to the desired centre frequency. The sampled passbands of the derived filter are identical with those of the lowpass, but at different locations. Experimentally(ref 100) the results show that a 3dB loss of sub-optimum relative to the optimum is the usual case.

(c) Programming Considerations

The subroutines FILTER, TR and INTERP (Lists 7, 8 and 9 respectively) are used together to compute the impulse response of the required lowpass, bandpass or highpass filter according to the following strategy.

(1) The upper and lower cut-off frequencies are specified by the variables FL and FU respectively and the sampling interval is stored as the variable DT.

(2) The passband and stop band(s) are specified by the real array 'A' according to the filter bandwidth and the three transition points T1, T2 and T3 are selected from subroutine TR which stores all of the transition values that are tabulated in table A.1. Those values which are not tabulated are linearly interpolated by subroutine INTERP.

(3) The imaginary array 'B' is set to zero such that the cartesian frequency arrays A and B can be inverse Fourier transformed by a call to FFT (not listed) to yield the impulse response of the filter which is again stored in array A. This array is cyclically rotated so that the peak of the impulse response now falls at the centre of the array. This produces the linear phase characteristic of the filter corresponding to a fixed delay equal to one half of the filter length.

(4) Finally, the impulse response, h(i), is normalised to set a unity gain at the mid-band frequency of the filter. This is achieved in a similar manner to the Butterworth design, where

$$G = H(z)\Big|_{z = \exp(j\omega_m T)}$$

where $\omega_m = 2\pi(FU - FL)$

and T is the sampling interval.

$$\therefore \quad h'(i) = \frac{h(i)}{G}, \qquad i = 1,2,\ldots,N \qquad (A.9)$$

N is the order of the filter which is 256 in this case.

The time waveform of the array 'A' corresponding to $\{h'(i)\}$ is shown in figure (A.5) for a bandpass filter set to 300 to 3400Hz and a sampling frequency of 16kHz.

The interpolated frequency response is computed by multiplying $\{h'(i)\}$ by a Hanning window weighting sequence

$$a_i = w_i \, h'(i) \quad \text{for} \quad i = 1,2,\ldots,N \tag{A.10}$$

where $w_i = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi i}{N}\right)\right]$

The sequence $\{a_i\}$ is expanded to 2N samples by the inclusion of a further N zero valued samples, i.e.

$$a_i = 0 \quad \text{for } i = N+1,\ N+2,\ldots,2N$$

and then the whole array is cyclically rotated by N/2 samples to render it a zero phase sequence.

$$b_{i-N/2} = a_i \quad ; \quad i = N/2+1,\ N/2+2,\ldots,2N$$

and

$$b_{i+3N/2} = a_i \quad ; \quad i = 1,2,\ldots,N/2 \tag{A.11}$$

The whole sequence $\{b_i\}$ is transformed by a 2N-point DFT to yield a representation of its interpolated frequency response.

$$B_k = \frac{1}{N} \sum_{i=1}^{2N} b_i \exp\left(-j \frac{\pi k i}{N}\right) \tag{A.12}$$

The magnitude and phase spectrum of the complex sequence $\{B_k\}$ is illustrated in figure A.6(a) and A.6(b) respectively. Notice that for a comparatively narrow transition band, the minimum attenuation

is 90dB or greater. The phase response shows that the sequence $\{B_k\}$ is wholly real which has positive valued elements when the phase $\angle B_k$ is zero and negative valued elements when $\angle B_k$ is $\pm 180^o$.

A.1.2.3 Sectioned Convolution

Once the impulse response has been determined by the above strategy, it now becomes appropriate to implement the filter by convolving this impulse response with the input signal to yield the filtered output signal. If the input sequence has a length of N_1 samples and the filter impulse response has N_2 samples, then the convolved sequence will be comprised of $L = N_1+N_2-1$ samples. In many of our filtering situations, we are interested in computing the convolution of these two finite duration sequences where one sequence is much longer than the other sequence, i.e. $N_1 >> N_2$ in the above case. Of course, we can always use a value $L = N_1+N_2-1$ but this is generally inefficient and impractical for several reasons. First, the entire longer sequence must be available before the convolution can be carried out, and furthermore since no processing occurs before the entire sequence is available, this implies long delays before the output is obtained. To alleviate these problems, a technique can be used to section the larger sequence and compute partial results that can be pieced together to form the desired output sequence.

The method used is called the overlap-add segmentation method; for simplicity we assume that the input sequence $\{x(n)\}$ is effectively of infinite duration and the duration of $h(n)$ is N_2 samples. The

sequence $x(n)$ is sectioned into pieces, each of duration N_3 samples. The value of N_3 is generally chosen to be of the order of N_2 (ref 101). Thus the input sequence can be represented as

$$x(n) = \sum_{k=0}^{\infty} x_k(n) \tag{A.13}$$

$$\text{where} \quad x_k(n) = \begin{cases} x(n) & ; \quad kN_3 \leq n \leq (k+1)N_3-1 \\ 0 & \text{otherwise} \end{cases} \tag{A.14}$$

The linear convolution of $x(n)$ and $h(n)$ thus can be written as

$$y(n) = \sum_{m=0}^{n} h(m)\, x(n-m)$$

$$= \sum_{m=0}^{n} h(m) \sum_{k=0}^{\infty} x_k(n-m) \tag{A.15}$$

$$= \sum_{k=0}^{\infty} h(n) * x_k(n)$$

$$= \sum_{k=0}^{\infty} y_k(n) \tag{A.16}$$

where $y_k(n) = h(n) * x_k(n)$

The durations of each of the convolutions of equation (A.16) comprises of (N_2+N_3-1) samples. There is a region of (N_2-1) samples over which the k^{th} and the $(k+1)^{th}$ convolution overlap and the appropriate sequences must be added to produce the output sequence $y(n)$.

The convolution operation is performed by a call to the subroutine FCONV (X, N1, A, N2, L) in List 10, where

'X' contains the input sequence of N1 samples on entry, and contains the output (delayed) sequence on exit.

'A' contains the filter impulse response (or any finite length sequence) of N2 samples.

Since it is required to store the overlapping samples in the two-dimensional array 'Z', L refers to the row of 'Z' in which these samples are stored. This is such that the same subroutine may be used for 'parallel' filtering by calling FCONV more than once in the same program for different filtering operations, without causing confusion of the data. The array 'Y' is a working array in which the convolution process of computing $y_k(n)$ in equation (A.16) is performed. The result of successive calls to FCONV subroutine is that the output sequence 'X' ($= y_k(n)$) may be concatenated to form the entire sequence $y(n)$ according to the equation (A.16).

Examples of the results of this type of filtering process can be found throughout Chapters 5 and 6 of this thesis. Specifically there are some illustrations shown in figures 5.27(b) and 6.5(b).

A.2 High Frequency Regeneration

The high frequency regeneration operations, described in section 5.3, for speech quality enhancement are performed by calling the subroutines LOGCOM and 'W' from Lists 11 and 12 respectively. The

LOGCOM routine applies logarithmic compression and the 'W' routine simulates the W-transfer function. These are illustrated in figure 5.6(a) and (b) respectively.

A.2.1. Logarithmic Compression

The subroutine LOGCOM (List 11) employs the equation (5.14) to apply logarithmic compression, i.e.

$$x' = 4\alpha \, \mathrm{sgn}(x) \log_e [c|x| + 1] \qquad (5.14)$$

where $\alpha = 1/\log_e [c|x|_{max} + 1]$

$\mathrm{sgn}(x) = x/|x|$

and c is the compression factor (and equal to 5 in our case).

The transfer function is set up in array 'X' of 1000 points which is then used as a look-up table for the input block of samples contained in array 'A' of dimension N. ZM is the maximum absolute value that array 'A' is likely to have and this value is initially assigned from the main program segment. The maximum value of G * A(I) in line 17 is equal to 500, thus the value of J ranges from 1 to 1000. This corresponds to the number of points in array 'X'. Thus line 18 provide the resultant value of x' for the input sample x according to the equation (5.14).

A.2.3 The 'W' Transfer Function

The routine 'W' (List 12) is called immediately after the LOGCOM routine where it is now known that the values of the input array 'A' range between -1.0 and +1.0 only. The W-transfer function is stored in array 'Y' of 1000 points according to equation (5.13):

$$\begin{aligned} y &= 1 - 4|x'| \quad , |x'| \leq 0.5 \\ &= 4|x'| - 3 \quad , 0.5 < |x'| \leq 2.0 \\ &= 0 \quad \text{elsewhere} \end{aligned} \tag{5.13}$$

The values of 'Y' are then used as a look-up table for processing the input array 'A'. Again, the range of 'J' in line 22 varies between 1 and 1000 which corresponds to the number of points in array 'Y'. Line 23 yields an output value of y, corresponding to the 'Y' array, for every logarithmically compressed sample x', denoted by the 'A' array.

A.3 Spectral Peak Synthesis

The routine 'FORMANT' (List 13) simulates a 4-pole resonant filter and is used to synthesise an additive formant to bandlimited voiced speech, or to generate a spectral peak in a regenerated /s/ spectrum. The detailed description of these two operations can be found in sections 5.3.3 and 5.4.1.5 respectively. The recursive network is shown in figure 5.10 and the z-plane representation is shown in figure 5.11. From these we had formulated the following expressions i.e.

the filter transfer function $H(z) = H_1(z).\ H_2(z)$

where $$H_1(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

and $$H_2(z) = \frac{1}{1 + b_1 z^{-1} + b_2 z^{-2}} \qquad (5.18)$$

The filter coefficients were determined from the centre frequency, the pole separation and the damping factors of the poles in the s-plane representation in figure 5.10.

$$\begin{aligned} a_1 &= -2 \exp(\alpha_1 T) \cos(\omega_1 T) \\ a_2 &= \exp(2\sigma_1 T) \\ b_1 &= -2 \exp(\sigma_2 T) \cos(\omega_2 T) \\ b_2 &= \exp(2\sigma_2 T) \end{aligned} \qquad (5.21)$$

where $T = 1/F_s$, the reciprocal of the sampling frequency.

To generate the extra formant for addition to bandlimited voiced speech, the parameters were selected such that

$$\sigma_1 = \sigma_2 = -100\ s^{-1}$$

$$f_c = 4\text{kHz}$$

and $\Delta f = 200\text{Hz} = 0.2\text{kHz}$.

Such that in equation (5.21), $\omega_1 = 2\pi(f_c - \Delta f)$ and $\omega_2 = 2\pi(f_c + \Delta f)$.

To regenerate the spectral peak for the /s/ utterance, the parameters were adjusted to

$$\sigma_1 = \sigma_2 = -1{,}100s^{-1}$$

$$f_c = 4.8kHz$$

and $$\Delta f = 400Hz$$

The filter in both cases was implemented as two cascaded second order stages as shown in figure A.7(a) and (b) respectively. The variable 'INPUT' in routine FORMANT is a sample generated from the high frequency regeneration subroutines 'LOGCOM' AND 'W'. The variable 'OUTPUT1' from the network in figure A.7(a) is precisely the input variable applied to the network in figure A.7(b) at any one time instant. The 'OUTPUT2' samples emerging from the second network comprise the filtered signal whose spectrum is modified by the resonant network. The method used for checking the spectral transfer function of the designed resonant circuit is to evaluate the transfer function H(z) around the unit circle in the z-domain. The gain and phase response can be computed from $|H(j\omega T)|$ and $\angle H(j\omega T)$ respectively for a number of frequency points. Subroutine 'RESPONSE' (List 14) is used to carry out this task using 100 values of ω between dc and πF_s . The gain and phase response of the two types of resonant filter discussed are shown in figures 5.12 and 5.30 respectively.

A.4

LIST 1

```
      SUBROUTINE BUTTER (N, FC, ALPHA, BETA)
C
C     THIS SUBROUTINE COMPUTES THE POLES AND ZEROS OF A
C     BUTTERWORTH RESPONSE LOWPASS DIGITAL FILTER
C
C     INPUTS ARE:     N =      ORDER OF FILTER REQUIRED
C                     FC =     REQUIRED CUTOFF FREQUENCY
C                              AS A FUNCTION OF THE CLOCK
C                              FREQUENCY
C     OUTPUTS ARE:    ALPHA =  COMPLEX ARRAY CONTAINING
C                              THE TRANSFER FUNCTION ZEROS
C                              IN ITS FIRST N LOCATIONS
C                              (ALL N ZEROS LIE AT Z=-1.0)
C                     BETA =   COMPLEX ARRAY CONTAINING
C                              THE TRANSFER FUNCTION POLES
C                              IN ITS FIRST N LOCATIONS.
C                              COMPLEX POLES OCCUPY ADJACENT
C                              LOCATIONS; IF N IS ODD THE
C                              REAL POLE IS IN LOCATION 1
C
      COMPLEX ALPHA(30),BETA(30)
      WC=3.141592654*FC
      TAN2=2.0*SIN(WC)/COS(WC)
      TANSQ=0.25*TAN2**2.0
      IF(N.EQ.1) GOTO 2
      IN=MOD(N,2)
      N1=N+IN
      N2-(3*N+IN)/2-1
      DO 1 M=N1, N2
      A=3.141592654*FLOAT(2*M+1-IN)/FLOAT(2*N)
      ANUM=1.0-TAN2*COS(A)+TANSQ
      U=(1.0-TANSQ)/ANUM
      V=TAN2*SIN(A)/ANUM
      I=(N2-M)*2+1
      BETA(I+IN)=CMPLX(U,V)
```

LIST 1 (Cont'd)

```
1     BETA(I+IN+1)=CMPLX(U,-V) LPF Poles
      IF(IN) 3,3,2
2     BETA(I)=CMPLX(((1.0-TANSQ)/(1.0+TAN2+TANSQ))),0.0)
3     DO 4 I=1,N
      ALPHA(I)=(-1.0,0.0)       Position of LPF Zeros
4     CONTINUE
      N1=N+1
      DO 5 I=N1, 30
      ALPHA(I)=(0.0,0.0)
      BETA(I)=(0.0,0.0)
5     CONTINUE
      RETURN
      END
```

LIST 2

```
      SUBROUTINE BPASS (N.F1,F2,ALPHA,BETA)
C
C     THIS SUBROUTINE COMPUTES THE 2N POLES AND ZEROS OF A
C     BANDPASS DIGITAL FILTER FROM THE N POLES AND ZEROS
C     OF A LOW PASS PROTOTYPE.
C
C     INPUTS ARE          N =   ORDER OF LOW PASS PROTOTYPE
C                        F1 =   LOWER CUTOFF FREQUENCY
C                        F2 =   UPPER CUTOFF FREQUENCY
C                     ALPHA =   ARRAY CONTAINING THE ZEROS
C                               OF LOW PASS PROTOTYPE
C                      BETA =   ARRAY CONTAINING POLES OF
C                               LOW PASS PROTOTYPE
C     (THE LOWPASS SHOULD HAVE A CUTOFF FREQUENCY AT
C     FC=F2-F1.  ALL FREQUENCIES ARE EXPRESSED AS A FRACTION
C     OF THE CLOCK RATE.)
C
C     OUTPUTS ARE     ALPHA = ARRAY CONTAINING THE ZEROS OF
C                             THE BANDPASS FILTER
C                      BETA = ARRAY CONTAINING THE POLES OF
C                             THE BANDPASS FILTER
      COMPLEX ALPHA(30),BETA(30),Z,S
      P=3.14159265359
      A=COS(P*(F1+F2))/COS(P*(F2-F1))
      IM=MOD(N,2)
      N1=N-1
      N2=N
   10 DO 20 IA=1,2
      S=CMPLX((FLOAT(2*IA-3)),0.0)
      IB=2-IA
      DO 20 I=1,N1
      J=N2-I+1
      K=N2(I+IB)-I+1+IM*(IA-1)
      Z=(0.5,0.0)*((1.0,0.0)+ALPHA(J))*CMPLX(A,0.0)
      ALPHA(K)=Z+S*CSQRT(Z*Z-ALPHA(J))
```

LIST 2 (Cont'd)

```
      Z=(0.5,0.0)*((1.0,0.0)+BETA(J)*CMPLX(A,0.0)
      BETA(K)=Z+S*CSQRT(Z*-BETA(J))                    Transformation of Pole
20    CONTINUE
      IF(N1.EQ.N2)  RETURN
      N1=1
      N2=1
      IM=0
      GOTO 10
      END
```

LIST 3

```
      SUBROUTINE COEFF (N,ALPHA,BETA,A1,A2,B1,B2,A0,N0,FL,FM)

      SUBROUTINE COMPUTES THE COEFFICIENTS IN A SERIAL
C     FORM REALISATION OF A DIGITAL FILTER
C
C     INPUTS ARE          N  =  NUMBER OF SECTIONS IN FILTER
C                     ALPHA  =  ARRAY HOLDING FILTER ZEROS
C                      BETA  =  ARRAY HOLDING FILTER POLES
C     (CONJUGATE PAIRS MUST BE LOADED IN ADJACENT
C     LOCATIONS 1,I+1 WHERE I IS AN ODD NUMBER)
C
C     OUTPUTS ARE A1,A2,B1,B2  ARRAYS HOLDING SECTION
C                              COEFFICIENTS
C                    A0 = GAIN COEFFICIENTS FOR UNITY
C                      GAIN AT ZERO FREQUENCY FOR LPF
C                       OR AT MID-BAND FREQUENCY FOR BPF
C
      COMPLEX ALPHA(30), BETA(30)
      DIMENSION AMOD(30), A1(15),A2(15),B1(15),B2(15)
      COMMON/BI5/DT
      COMPLEX/AN,AD
      PI=4.0*ATAN(1.0)
      FM=(FL+FU)/2.
      IF(FL.EQ.0.0) FM=0.0
      AG-2.0*PI*FM*DT
      A0=1.0
      DO 200 I=1,N
      I1=2*I-1
      I2=2*I
      A1(I)=REAL(-ALPHA(I1)-ALPHA(I2))
      A2(I)=REAL(ALPHA(1)+ALPHA(I2))      Compute Coefficients
      B1(I)=REAL(-BETA(I1)-BETA(I2))
      B2(I)=REAL(BETA(I1)*BETA(I2))
      AMOD(I1)=CABS(BETA(I1))
      AMOD(I2)=CABS(BETA(I2))
```

LIST 3 (Cont'd)

```
      AM=(1.0,0.0)+CMPLX(A1(I)*COS(AG),-A1(I)*SIN(AG))
       +CMPLX(A2(I)*COS(2.0*AG),-A2(I)*SIN(2.0*AG))        Gain
      AD =(1.0,0.0)+CMPLX(B1(I)*COS(AG),-A1(I)+SIN(AG))   Normali-
       +CMPLX(B2(I)*COS(2.0*AG),-B2(I)*SIN(2.0*AG))       zation
      RAT=CABS(AN)/CABS(AD)
      AO=AO*RAT
C     AO=AO*(1.0+A1(I)+A2(I))/(1.0+B1(I)+B2(I))
200   CONTINUE
      IF(ABS(AO).LT.1.0E-6)  AO=0.1E-5
      AO=1.0/AO
      WRITE(2,21)  NO
      CALL WRITE (AMOD, 2*N)
21    FORMAT (/,10X,' THE MAGNITUDE OF THE POLE POSITIONS',
     1 'FOR FILTER', I3, '    ARE :')
C     AO=0.1E-4
      RETURN
      END
```

LIST 4

```
      SUBROUTINE FREQ(N,NF ,A1,A2,B1,B2,A0,GAIN,PHASE,XAXIS)
C
C     THIS SUBROUTINE COMPUTES THE GAIN AND PHASE CHARACTERISTICS
C     OF A DIGITAL FILTER FROM THE COEFFICIENTS OF ITS SECOND
C     ORDER SECTION REALISATION.
C
C     INPUTS ARE      N   =    NUMBER OF SECOND ORDER SECTIONS
C                              IN THE FILTER
C                     NF  =    NUMBER OF POINT REQUIRED ON
C                              THE FREQUENCY SCALE
C                     A0  =    GAIN COEFFICIENT
C              A1,A2,B1,B2  = ARRAYS HOLDING SECOND ORDER
C                              SECTION COEFFICIENTS
C     OUTPUTS ARE,    GAIN  = ARRAY HOLDING VALUES OF FILTER
C                              GAIN IN DECIBELS
C                     PHASE  = ARRAY HOLDING VALUES OF FILTER
C                              PHASE SHIFT IN DEGREES
C                     XAXIS  = ARRAY HOLDING FREQUENCY VALUES
C
C     THE GAIN AND PHASE ARE COMPUTED AT NF POINTS EQUALLY
C     SPACED FROM ZERO TO HALF CLOCK FREQUENCY INCLUSIVE.
C
      COMPLEX WJ,WJ2,H,AD,AN
      DIMENSION A1(N),A2(N),B2(N),GAIN(NF),PHASE(NF)
      DIMENSION XAXIS(NF)
      C=180.0/3.141592654
      DO 403 J=1,NF
      OMEGA=3.141592654*(10.0**(4.0*FLOAT(J)/FLOAT(NF)))/10000.0
      XAXIS(J)-(10.0**(4.0*FLOAT(J)/FLOAT(NF)))/20000.0
      WJ-CMPLX(COS(OMEGA),SIN(OMEGA)
      WJ2=WJ*WJ
      H=CMPLX(A0,0.0)
      DO 403 I=1,N
      AN=(1.0,0.0)+CMPLX(A1(1),0.0)*WJ+CMPLX(A2(I),0.0)+WJ2
      AD=(1.0,0.0)+CMPLX(B1(1),0.0)*WJ+CMPLX(B2(I),0.0)*WJ2
```

LIST 4 (Cont'd)

```
        X=CABS(AN)
        IF(X-1.0E-6) 404,404,401
401     X=CABS(AD)
        IF(X-1.0E-6) 405,405,402
402     H=H*AN/AD                                   Gain and
        GAIN(J)=20.0*ALOG10(CABS)(H))               Phase
        PHASE(J)=C*ATAN2(AIM AG(H),REAL(H))         Response
403     CONTINUE
        RETURN
404     GAIN(J)=-120.0
        PHASE(J)=0.0                        i.e. |H(jwt)| and /H(jwt)
        GOTO 403
405     GAIN(J)=120.0
        PHASE(J)=0.0
        GOTO 403
        END
```

LIST 5

```
      SUBROUTINE SETUP(FL,FU,FS,NR,NO)
C
C     THIS SUBROUTINE COMPUTES THE COEFFICIENTS OF THE
C     SERIAL REALISATION OF A BANDPASS TIME DISCRETE
C     FILTER DERIVED FROM A LOW PASS BUTTERWORTH
C     RESPONSE PROTOTYPE.  THE GAIN AND PHASE CHARACTERISTICS
C     ARE PLOTTED FROM DC TO HALF THE CLOCK FREQUENCY.
C
C     THE INPUTS ARE      FS = SAMPLING FREQUENCY
C                         FL = LOW BAND CUTOFF FREQUENCY
C                         FU = HIGH BAND CUTOFF FREQUENCY
C                         NR = ORDER OF THE LOW PASS PROTOTYPE
C
C     THE OUTPUTS ARE A1,A2,B1,B2: ARRAYS HOLDING THE
C                    SECTION COEFFICIENTS
C                       AO  = GAIN COEFFICIENT FOR UNITY
C                             GAIN AT ZERO FREQUENCY
C
      DIMENSION AAO(6),AA1(15,6),AA2(15,6),BB1(25,6),BB2(15,6)
      INTEGER ORDER(6)
      COMMON/BL50/AAO,AA1,AA2,BB1,BB23/BL63/ORDER
      REAL LOW
      COMPLEX ALPHA(30),BETA(30)


      DIMENSION A1(15),A2(15),B1(15),B2(15)
      DIMENSION XAXIS(200),GAIN(200),PHASE(200)
      FM=(FU+FL)/2.0
      CLOCK=FS*1000.0
      LOW=FL*1000.0
      HIGH=FU*1000.0
      NF=129
      LOW=LOW/CLOCK
      HIGH=HIGH/CLOCK
      FC=HIGH-LOW
```

LIST 5 (Cont'd)

```
      N,ORDER(NO)=NR
      CALL BUTTER(ORDER(NO),FC,ALPHA,BETA)                          LPF Prototyp
      IF(LOW.NE.0.0) CALL BPASS(ORDER(NO),LOW,HIGH,ALPHA,BETA)  Transformatio
      CALL COEFF(NR,ALPHA,BETA,A1,A2,B1,B2,A0,N0,FL,FU)             Compute
      WRITE(2,100)                                                  Coefficients
  100 FORMAT(1H1, THE FILTER SECTION COEFFICIENTS ARE: -1,/)
      WRITE(2,101)(A1(I),I=1,NR)
      WRITE(2,102)(A2(I),I=1,NR)
      WRITE(2,103)(B1(I),I=1,NR)
      WRITE(2,104)(B2(I),I=1,NR)
  101 FORMAT(1H/,/,'A1:- ',5E14.4,/,5E14.4)
  102 FORMAT(1H/,/,'A2:- ',5E14.4,/,5E14.4)
  103 FORMAT(1H/,/,'B1:- ',5E14.4/,5E14,4)
  104 FORMAT(1H/,/'B2:- ',5F14.4/,5E14.4)
      CALL FREQ(NR,NF,A1,A2,B1,B3,A0,GAIN,PHASE,XAXIS)  Compute gain and pha
                                                        characteristics.
      DO 105 K=1,NF
      XAXIS(K)=XAXIS(K)*CLOCK/1000.0
C     XAXIS(K)=ALOG10(XAXIS(K))
  105 CONTINUE
      CALL PLOTFILE(XAXIS,GAIN,NF)
      CALL PLOTFILE(XAXIS,PHASE,NF)
      AA0(NO)=A0
      DO 10 I=1,ORDER(NO)
      AA1(I,NO)=A1(I)
      AA2(I,NO)=A2(I)
      BB1(I,NO)=B1(I)
      BB2(I,NO)=B2(I)
10    CONTINUE
      IF(LOW.EQ.0.0)ORDER(NO)=NB/2+1
      WRITE 2,106) I0,AA0(NO)
  106 FORMAT(1H,/,10X,'AA0(',I1,')= ',E14.4)
      RETURN
      END
```

LIST 6

```
      FUNCTION HIIR (SIGNAL,NO)
C
C     THIS FUNCTION REPRESENTS AN N STAGE CASCADED
C     FILTER WITH EACH STAGE CONTAINING 2 POLES AND
C     2 ZEROS.  THE ZEROS ARE DEFINED BY THE ARRAYS
C     A1 AND A2.  THE POLES ARE DEFINED BY THE ARRAYS
C     B1 AND B2.
C
C            SIGNAL = I/P SAMPLE
C              HIIR = O/P SAMPLE
C               NO = FILTER NO. (MAX=6)
C
      DIMENSION AO(6),A1(15,6),A2(15,6),B1(15,6),B2(15,6),Y(16)
      DIMENSION X1(15,6),X2(15,6),X3(15,6)
      INTEGER N(6)
      COMMON/BL50/AO,A1,A2,B1,B2/BL63/N
      Y(1)=SIGNAL
      DO 1 I=1,N(NO)                                          Feed the
      X1(I,NO)=Y(I)-(B1(I,NO)*X2(I,NO)+B2(I,NO)*X3(I,NO))     signal
      Y(I+1)=X1(I,NO)+A1(I,NO)*X2(I,NO)+A2(I,NO)*X3(I,NO)     through
      X3(I,NO)=X2(I,NO)                                       the filter
      X2(I,NO)=X1(I,NO)
3     NFO=NO
    1 CONTINUE
      HIIR=AO(NO)*Y(N(NO)+1)
      RETURN
      END
```

LIST 7

```
      SUBROUTINE FILTER(FL,FU,A)
C
C     THIS SUBROUTINE SETS UP BPF IR IN ARRAY 'A' WITH
C     OPTIMAL VALUES OBTAINED FROM LPF PROTOTYPE
C
      DIMENSION A(256),B(1025)
      COMMON/BL4/B/BL5/DT/BL10/FMAX   (Folding Frequency = F_s/2)
      COMPLEX HX
      LOGICAL FT
      PI=4.0*ATAN(1.0)
      FM=(FU+FL)/2.0
      AG=2.0*PI*FM*DT     Mid-band frequency.
      N=256
      FN=FLOAT(N/2+1)
      NFL=NINT(FL/FMAX)*FN)
      NFU=NINT((FU/FMAX)*FN)
      NFC=(NFL+NFU)/2.0
      NBW=NFU-NFC
      IF(NFL.EQ.0) NBW=NFU
      CALL TR(NBW,T1,T2,T3)   Select 3 transition points
C     COMPILE F-RESPONSE OF BPF
      IF(NFL.LF.4) GOTO 7
      M1=NFL-4
      A(M1+1)=T1                                  Lower
      A(M1+2)=T2                                  Transition
      A(M1+3)=T3                                  Band
      DO 1 I=1,M1
      A(I)=0.0
1     CONTINUE
    7 CONTINUE
      IF(NFL.LE.4) NFL=1
      DO 2 I=NFL,NFU
      A(I)=1.0                                    Pass Band
2     CONTINUE
```

LIST 7 (Cont'd)

```
      IF((NFU+4).GE.(N/2+1)) GOTO 8
      A(NFU+3)=T1                                   Upper
      A(NFU+1)=T3                                   Transition
      A(NFU+1)=T3                                   Band
      DO 3 I-NFU+4,N/2+1
      A(I)=0.0                                      Stop Band
3     CONTINUE
    8 CONTINUE
      DO 4 I=1,N/2+1
      B(I)=0.0          Set the imaginary components to zero
4     CONTINUE
C     IDFT
      FT=TRUE
      CALL FFT(A,B,N,FT)
C     IR NOW IN ARRAY 'A'
      DO 5 I=1,N/2
      B(I)=A(I+N/2)                                 Rotate IR
      B(I+N/2)=A(I)
    5 CONTINUE
      A(1)=0.0
      HX=(0.0,0.0)
      DO 10 K=1,N                                                    Gain
      HX=HX=CMPLX(A(K)*COS(AG*FLOAT(K)),-A(K)*SIN(AG*FLOAT(K)))      Norma-
1     CONTINUE                                                       lisation
      GM=1.0/CABS(HX)
      WRITE(2,21)  FL,FU,GM
21    FORMAT(1H  ,/,10X,'GAIN FOR ',F5.2,'TO',F5.2,'FILTER IS ',E14.4)
      DO 11 I=1,N
      A(I)=A(I)*GM
11    CONTINUE
      RETURN
      END
```

LIST 8

```
      SUBROUTINE TR(NDW,T1,T2,T3)
C
C     THIS SUBROUTINE HOLD ALL TRANSITION VALUES FOR LPF
C     PROTOTYPE FILTER TYPE 1 DATA N=256
C     C.F. TABLE III PAGE 93, RABINER ET. AL. (REF 100)
C
      DIMENSION A(130.3)
      L1=124
      L2+3
      IF(I.NE.0)GOTO 1
      I=1
      A(1,1)=0.10647949
      A(1,2)=0.19387524
      A(1,3)=0.67664281
      A(2,1)=0.01963501
      A(2,2)=0.22197911
      A(2,3)=0.70144920
      A(3,1)=0.01908569
      A(3;2)=0.22085960
      A(3,3)=0.69990539
      A(5,1)=0.02305298
      A(5,2)=0.24117076
      A(5,3)=0.71635813
      A(8,1)=0.02479248
      A(8,2)=0.24843111
      A(8,3)=0.72164702
      A(9,1)=0.02329712
      A(9,2)=0.24253562
      A(9,3)=0.71679420
      A(16,1)=0.02444458
      A(16,2)=0.24629538
      A(16,3)=0.71900030
      A(32,1)=0.02577896
      A(32,2)=0.25163493
      A(32,3)=0.72307099
```

LIST 8 (Cont'd)

```
A(48,1)=0.02421875
A(48,2)=0.24359358
A(48,3)=0.71550480
A(56,1)=0.02345581
A(56,2)=0.23957232
A(56,3)=0.71177494
A(64,1)=0.02396851
A(64,2)=0.24199281
A(64,3)=0/71380179
A(80,1)=0.02351685
A(80,2)=0.23926844
A(80,3)=0.71103085
A(96,1)=0.02435913
A(96,2)=0.24219392
A(96,3)=0.71293931
A(104,1)=0.02552490
A(104,2)=0.24590992
A(104,3)=0.71524908
A(112,1)=0.02607422
A(112,2)=0.02607622
A(112,3)=0.71857490
A(120,1)=0.02683105
A(120,2)=0.25909273
A(120,3)=0.73130690
A(121,1)=0.02561035
A(121,2)=0.25523207
A(121,3)=0.72018388
A(122,1)=0.02344360
A(122,2)=0.24778644
A(122,3)=0.72456966
A(123,1)=0.01946196
A(123,2)=0.23070314
A(123,3)=0.71099759
A(124,1)=0.01351929
```

LIST 8 (Cont'd)

```
      A(124,2)=0.20394843
      A(124,3)=0.69037794
C     ALL OTHER VALUES ARE INITIALLY 0.0
C     INTERPOLATE THE REMAINING VALUES
      CALL INTER(A,L1,L2)
    1 CONTINUE
      T1=A(NBW,1)
      T2=A(NBW,2)
      T3=A(NBW,3)
      RETURN
      END
```

LIST 9

```
      SUBROUTINE INTER(A,L1,L2)
C
C     INTERPOLATOR SUBROUTINE
C
      DIMENSION A(L1,L2)
      DO 1 J=1,L2
      DO 1 I=1,L1
      IF(A(I,J).NE.0) GOTO 1
      DO 2 K=1,100
      IF(A(I+K,J),EO.0) GOTO 2
      K1=K
      GOTO 3
    2 CONTINUE
    3 CONTINUE
      G1=(A(I+K1,J)-A(I-1,J))/K1+1)        Gradient between adjacent
      DO 4 K=1,K1                          non-zero valued samples.
      A(I-1+K,J)=A(I-1,J)+K*G1             Straight line interpolation.
    4 CONTINUE
      I=I+K1
    1 CONTINUE
      RETURN
      END
```

LIST 10

```
      SUBROUTINE FCONV(X,N1,A,N2,L)
C
C     THIS SUBROUTINE PERFORMS SEGMENTED CONVOLUTION
C     ON INPUT 'X' TO YIELD OUTPUT 'X' BY FILTER IR 'A'
C     I/P BLOCKSIZE = N1
C     I/P BLOCKSIZE = N2
C     L REFERS TO FILTER NO
C
      DIMENSION A(N2),X(N1),Y(1300),Z(6,256)
      COMMON/BL14/Y
      N3=N1+N2-1
      CALL RESET(Y,1300)
      DO 1 I=1,N2-1
      Y(I)=Z(L,I)          Load previous overlapping sequence in working
    1 CONTINUE             array.
      DO 2 I=1,N1
      DO 2 J=1,N2
      K=I+J-1
      Y(K)=Y(K)+X(I)*A(J)  Segmented Convolution
    2 CONTINUE
      DO 3 I=1,N1
      X(I)=Y(I)
    3 CONTINUE
C     STORE OVERLAPPING SEQUENCE IN Z(L,I)
      DO 4 I=1,N2-1
      Z(L,I)=Y(I+N1)       Store current overlapping sequence in working
      Y(I+N1)=0.0          array.
    4 CONTINUE
      RETURN
      END
```

LIST 11

```
      SUBROUTINE LOGCOM(A,N)
C
C     LOGARITHMIC COMPRESSOR
C
      DIMENSION A(N),X(1000)
      COMMON/BL23/ZM
      IF(M.NE.0) GOTO 1
      M=1
      CALL RAMP (X,1000)      Ramp generator; X increases linearly from -1 to
      C=5.0                   in 1000 points.
      ALPHA=1.0/(ALOG(C*500.0+1.0)
      DO 2 I=1,1000
      X(I)=SGN(X(I))+ALOG(C*ABS(X(I)+1.0)*ALPHA   Logarithmic transfer functi
    2 CONTINUE
      IF(ZM.EQ.0.0) GOTO 4
      G=1000.0/(2.0*ZM)                                Scale input samples.
      G=G*0.998
      CONTINUE
      DO 3 I=1,N
      J=NINT(G*X(I))+500
      A(I)=X(J)
    3 CONTINUE
    4 CONTINUE
      RETURN
      END
```

LIST 12

```
      SUBROUTINE W(A,N)
C
C     THIS SUBROUTINE CAUSES SPECTRAL SPREADING OF EXCITATION SIGNAL
C     BY A NON-LINEAR DISTORTION ALGORITHM:
C
C                      1 *     *    ALPH
C                        *   * *   *                    SAW-TOOTH
C                         * *   * *                     TRANSFER FN.
C                  -1      *     *
C                    -4  -2  0  2  4
C
      DIMENSION Y(1000) A(N)
      IF(M.NE.0) GOTO 5
      DO 1 I=1,1000
      Y(I)=FLOAT(I)
    1 CONTINUE
      ALPH=1.0
      G1=2.0/250.0
      G2=(1.0+ALPH)/250.0
      C1=-(ALPH+2.0)
      C2=3.0+ALPH+2.0
      DO 2 I=1,250
      Y(I)=1.0-G1*Y(1+250)          Store the 'W'
      Y(I+250)=C1+G2*Y(I+250)       function in
      Y(I+500)=C2-G2*Y(I+500)       array 'Y'
      Y(I+250)=-7.0+G1*Y(I+750)
    2 CONTINUE
      M=1
    5 CONTINUE
      Z=499.5
      DO 6 I=1,N
      J=NINT(Z*A(I))+500
      A(I)=Y(J)
    6 CONTINUE
      RETURN
      END
```

LIST 13

```
      SUBROUTINE FORMANT(X,N)
C
C     THIS SUBROUTINE SYNTHESISES AN ADDITIONAL FORMANT TO
C     THE SPEECH SIGNAL PROVIDED 'X' HAS A FLAT SPECTRUM
C     WITH PRESERVED FINE STRUCTURE
C
C     THE FORMANT IS SYNTHESISED USING 2 CASCADED
C     2ND ORDER ALL-POLE RECURSIVE RESONANT NETWORKS
C                  FC=CENTRE FREQUENCY OF POLES
C                  DELTF=POLE SEPERATION
C                  SIGMA=POLE DAMPING FACTORS
C                  G=RESONANCE GAIN FACTOR
C
      DIMENSION A(2),B(2),W(2),Y(2),X(N)
      REAL INPUT
      COMMON/BL5/DT
      IF(JC.EQ.1) GOTO 1
      JC=1
      PI=4.0+ATAN(1.0)
      FC=4.0                          Centre frequency and
      DELTF=0.2                       Separation of poles
      WC=2.0*PI*FC
      DELTAW=2.0*PI*DELTF
      SIGMA=0.1
      W1=WC+DELTAW/2
      W2=WC-DELTAW/2
      G=0.1E-2
      A(1)=-2.0*(EXP(-SIGMA*DT))*COS(W1*DT)
      B(1)=-2.0*(EXP(-SIGMA*DT)*COS(W2*D1)        Recursive filter
      A(2)=EXP(-2.0*SIGMA*DT)                     coefficients
      B(2)=EXP(-2.0*SIGMA*DT)
      WRITE(2,20) A(1),A(2),B(1),B(2)
20    FORMAT(1H ,/,10X,4F14.4)
      CALL RESPONSE(A,P,2)
1     CONTINUE
      DO 2 I=1,N
```

LIST 14

```
      SUBROUTINE RESPONSE(A,B,N)
C
C     THIS SUBROUTINE DETERMINES THE GAIN AND PHASE
C     RESPONSE OF A RECURSIVE FILTER WITH N COEFFS
C     AND 2-STAGES.
C
      DIMENSION A(N),B(I),G(100),PH(100),H(2)
      COMMON/BL5/DT
      COMPLEX H,Z,Z.J
      PI=4.0*ATAN(1.0)
      C=180.0/PI
      NFP=100
      DO 1 I=1,NFP
      OMEGA=(PI*FLOAT(I-1))/FLOAT(NFP)
      Z=CMPLX(COS(OMEGA),SIN(OMEGA))
      H(1),H(2)=CMPLX(1.0,0.0)
      DO 2 J=1,N
      AMAG=CABS(Z)**FLOAT(-J)
      ARG=ATAN2(AIMAG(Z),REAL(Z)*FLOAT(-J)
      ZJ=CMPLX(AMAG*COS(ARG),AMAG*SIN(ARG))
      H(1)=H(1)+CMPLX(A(J),0.0)*ZJ
      H(2)=H(2)+CMPLX((B(J),0.0)*ZJ
      H(1)=H(1)*H(2)
      H(1)=(CMPLX(1.0,0.0))/H(1)
2     CONTINUE
      G(I)=CABS(H(1))                          Gain and
      PH(I)=C*ATAN2(AIMAG(H(1)),REAL(H(1)))    Phase
1     CONTINUE                                 Response
      FS=1.0/DT                                i.e. |H(e^jwt)| and /H(e^jwt)
      CALL SAMPAR(FS,2*NFP)
      CALL FPLOT(G,PH,NFP)
      RETURN
      END
```

References

1) Williams, F. "Language and Speech, Introductory Perspectives" (book), Prentice-Hall, New Jersey, 1972.

2) Malah, D. "Time Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals" IEEE transactions on ASSP, Vol 27, No 2, pp 121-133, April 1979.

3) Flanagan, J.L. "Speech Analysis, Synthesis and Perception" (book), Springer-Verlag Berlin, 1972.

4) Bekesy, G.V. "Experiments in Hearing", New York, McGraw-Hill Book Co. 1960

5) Le Vay, D. "Human Anatomy and Physiology" (book) Hodder and Stroughton Ltd, 1974.

6) Mackenzie, G.W. "Acoustics" (book), Focal Press, 1964.

7) Flanagan, J.L., et al. "Speech Coding" et al, IEE Transactions on Communications, Vol 27, No 4, pp 710-737, April 1974.

8) Moye, L.S. "Study of the Effects on Speech Analysis of the types of Degradation occurring in Telephony" Report, Standard Telecommunications Laboratories Ltd, July 1979.

9) Richards, D.L. "Telecommunication by Speech" (book) Butterworth, London, 1973.

10) South, C. "Adaptive Filters to Improve Loudspeaking Telephones", Electronics Letters, Vol 15, No 21, pp 673-674, October 1979.

11) Rowlands, C.E. "Teleconferencing: A Service for the Businessman", Institution of Post Office Engineers' Journal, pp 90-94, July 1978.

12) Groves, I.S. "Orator - The Post Office Teleconferencing System", IEE Conference Publication No 184, pp 102-106, April 1980.

13) Whalley, S. "Dial-up System for Teleconferencing", British Telecom Memorandum, R13.2.2, No 77/16, Nov 1977.

14) Trueman, R. "A Communication System for Remote Conference", Systems Technology, No 24, pp 17-22, June 1976.

15) Archer, T.
Egglestone, B. "The Phone-in Phenomenon", Post Office Telecommunications Journal, pp 16-17, June 1978.

16) Radio Trent Engineering Dept Private Communication, March 1981,

17) Holmes, J.N. "A Survey of Methods for Digitally Encoding Speech Signals", IERE Int. Conf. on Digital Processing of Signals in Communications. Conference Proceedings No 39, pp 291-306, April 1981.

18) Reeves, A.M. British Patent 535,860 (1939).

19) Xydeas, C.S. "Differential Encoding Techniques Applied to Speech Signal", Ph.D. Thesis Loughborough University, November 1978.

20) Jayant, N.S. "Digital Coding of Speech Waveforms: PCM, DPCM and DM Quantizers", Proceedings of IEEE, Vol 62, No 5, May 1974.

21) Cattermole, K.W. "Principles of Pulse Code Modulation" (book) Illiffe, London, 1973.

22) Jayant, N.S. "Adaptive Quantization with a One Word Memory", Bell Systems Tech. Journal, pp 1119-1144, Sept 1973.

23) Jayant, N.S., Rosenberg, A.E. "The Preference of Slope Overload to Granularity in the Delta Modulation of Speech", J. Acoust. Soc. Am. 49, 133(A) 1971.

24) Jager, F. De "Delta Modulation, A Method of PCM Transmission using the 1 Bit Code", Philips Res, Rept. 7, pp 442-466, 1952.

25) Jayant, N.S. "Adaptive Delta Modulation with a One-bit Memory", BSTJ, Vol 49, No 3, pp 321-242, March 1970

26) Steele, R. "Delta Modulation Systems" (book), Pentech Press, London, 1975.

27) Dhadesugoor, et.al. "Delta Modulators in Packet Voice Networks", IEEE Transactions on Communications, Vol-28, Jan 1980.

28) Dudley, H. "The Vocoder", Bell Telephone Laboratories Record 17, pp 122-126, 1939.

29) Schroeder, M.R. "Vocoders: Analysis and Synthesis of Speech", IEEE. proc., Vol 54, pp 720-734, May 1966.

30) Oppenheim, A.V. Schafer, R.W. "Digital Signal Processing" (book) Prentice-Hall, New Jersey, 1975.

31) Itakura, F., Saito, S. "An Analysis-Synthesis Telephony Based on Maximum Likelihood Method" Proc. Int. Congr. Acoust. C-5-5, Tokyo, Japan, Aug 1968.

32) Markel, J.D., Gray, A.H. "Linear Prediction of Speech" (book), Springer-Verlag, New York, 1976.

33) Gabor, D. "New Possibilities in Speech Transmission", J. of IEE, Vol 94, Part III, No 32, pp 369-457, Nov 1947.

34) Fairbanks, G., Everitt, W.L. Jaeger, R.P. "Method for Time and Frequency Compression-Expansion of Speech" IRE Trans. Audio AU-2, pp 7-12, 1954.

35) Schiffman, M. "Playback Control Speech or Slow Speech Without Distortion", Electronics, pp 87-94, August 1974.

36) Silver, S.L. "How Speech can be Compressed and Expanded", Wireless World, pp 433-435, Sept 1975.

37) Bogert, B.P. "The Vobanc - A Two-to-One Speech Bandwidth Reduction System", J. Acoust, Soc. Am. 28, pp 399-404, 1956.

38) Schroeder, M.R., Flanagan, J.L., Lundry, E.A. "Bandwidth Compression of Speech by Analytic Signal Rooting", Proc. IEEE, Vol 55, pp 396-401, 1967

39) Takasugi, T., Suzuki, J. "Translation of Helium Speech by the Use of Analytic Signal", J. Radio Research Labs (Tokyo, Japan) Vol 21, No 103, pp 61-69, 1974.

40) Flanagan, J.L., Golden, R.M. "Phase Vocoder" BSTJ, Vol 45, pp 1493-1509, 1966.

41) Tribolet, J.M., Crochiere, R.E. "Frequency Domain Coding of Speech". IEEE Trans. on ASSP, Vol 27, No 5, October 1979

42) Carlson, J.P. "Digitised Phase Vocoder", Proc. Conf. on Speech Communication and Processing. A.F. Cambridge Research Labs and IEEE Audio and Electro-acoust Group, Cambridge, Mass., Nov 1967.

43) Portnoff, M.R. "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform". IEEE Trans. on ASSP, Vol 24, No 3, 1976

44) Portnoff, M.R. "Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis", IEEE Trans. on ASSP, Vol 28, No 1, pp 55-69, Feb 1980.

45) Flanagan, J.L., Christensen, S.W. "Computer Studies on Parametric Coding of Speech Spectra", JASA, Vol 68, No 2, pp 420-430, Aug 1980.

46) Flanagan, J.L., Christensen, S.W. "Technique for Frequency Division Multiplication of Speech Signals", JASA, Vol 68, No 4, pp 1061-1068, Oct 1980.

47) Schroeder, M.R., David, E.E. Jr. "A Vocoder for Transmitting 10kc/s Speech Over a 3.5kc/s Channel", Acustica, Vol 10, pp 35-42, 1960.

48) David, E.E. Jr., Schroeder, M.R., Logan, B.F., Prestigiacomo, A. "Voice-Excited Vocoders for Practical Speech Bandwidth Reduction", Proc. Stockholm Speech, Comm. Seminar, R.I.T., Stockholm, Sweden, Sept 1962.

49) Crochiere, R.E., Webber, S.A. Flanagan, J.L. "Digital Coding of Speech in Sub-Bands", BSTJ, Vol 55, No 8, pp 1069-1085, October 1976

50) Crochiere, R.E., Sambur, M.R. "A Variable-Band Coding Scheme for Speech Encoding at 4.8kb/s", BSTJ, Vol 56, No 5, May-Jun 1977.

51) Esteban, D., Galand, C., "Application of Quadrature Mirror Filters to Split Band Voice Coding Schemes", Proc. of 1977, Int. Conf. ASSP, Hartford Conn., pp 191-195, May 1977.

52) Crochiere, R.E. "On the Design of Sub-Band Coders for Low Bit Rate Speech Communication", BSTJ, Vol 56, No 5, pp 747-770, June 1977.

53) Zelinski, R., Noll, P "Adaptive Transform Coding of Speech Signals", IEEE Trans. on ASSP, Vol 25, No 4, pp 299-309, Aug 1977.

54) Shulman, J. "Speech Compression by means of the Discrete Fourier Transform", Ninth Convention of Electrical and Electronic Engineers in Israel, IEEE, April 1975.

55) Makhoul, J. "Methods of Non-Linear Distortions of Speech Signals", IEEE proc. ICASSP, pp 87-90, 1976.

56) Winckless, C.G. "Bit Rate Reduction Using Phase Redundancy of Speech", IERE, Communications Group Colloquium on Digital Coding of Speech, Feb 1977.

57) Weinstein, S.B. "Sampling-Based Techniques for Voiced Scrambling", IEEE International Conf. on Communications, Vol I, pp 16.2.1-16.2.6, Seattle, June 1980.

58) Oppenheim, A.V., Lim, J.S., Kopec, G., Pohlig, S.C. "Phase in Speech and Pictures" IEEE proc., ICASSP, pp 632-637, April 1979

59) Oppenheim, A.V., Lim, J.S. "The Importance of Phase in Signals" Proc. IEEE, Vol 69, No 5, pp 529-541 May 1981.

60) Malah, D. "Combined Time Domain Harmonic Compression and CVSD for 7.2 kbit/s Transmission of Speech Signals", I.E.E.E. proc., ICASSP, pp 504-507, 1980.

61) Malah, D. Crochiere, R.E. Cox, R.V. "Performance of Transform and Subband Coding Systems combined with Harmonic Scaling of Speech", IEEE Trans. on ASSP, Vol 29, No 2, pp 273-283, April 1981

62) King, R.A., Gosling, W. "Time Encoded Speech" Electron. Lett. Vol 14, No 15, pp 456-7, July 1978.

63) Gassmann, G.G. "Improvement of the Transmission Quality with Unchanged System Bandwidth" E.B.U. Review, No 144, April 1974.

64) Osborne, D.W., Croll, M.G. "Digital Sound Signals: Bit Rate Reduction Using an Experimental Digital Compander", BBC Research Report No RD1973/41, December 1973.

65) Croll, M.G. "The Possibility of Sending Programme Speech Contributions Digitally over the Public Telephone Network", BBC Research Report No RD1975/6, January 1975.

66) Stebbings, D.W. "An experimental Comparison of Four Methods for 64kbit/s Coding of Speech with a 7kHz Bandwidth", BBC Research Report No RD1976/11, May 1976.

67) Manson, W.I., Stebbings, D.W. "A Digital Split-band Compander for 64kbit/s Coding of Speech with a 7kHz Bandwidth", BBC Research Report No RD1977/41, November 1977.

68) Johnston, J.D., Goodman, D.J. "Digital Transmission of Commentary-Grade (7kHz) Audio at 56 or 65 kbits/s", IEEE Trans. on Communications, Vol 28, No 1, January 1980.

69) Croll, M.G. "Sound Quality Improvement of Broadcast Telephone Calls", BBC Research Report No RD1972/26, 1972.

70) Makhoul, J., Berouti, M. "High Frequency Regeneration in Speech Coding Systems", IEEE proc., ICASSP, pp 178-181, April 1979.

71) Osborne, D.W. Private Communication, November 1980, BBC Research Department, Kingswood, Surrey, UK.

72) Rabiner, L.R., Schmidt, C.E., Atal, B.S. "Evaluation of a Statistical Approach to Voiced- Unvoiced-Silence Analysis for Telephone Quality Speech", BSTJ, Vol 56, No 3, pp 455-481, March 1977.

73) Un, C.K., Hyeong, H.L. "Voiced/Unvoiced/Silence Discrimination of Speech by Delta Modulation", IEEE Trans, on ASSP, Vol 28, No 4, pp 398-407, August 1980.

74) Knorr, S.G. "Reliable Voiced/Unvoiced Decision" IEEE Trans. on ASSP, Vol 27, No 3, pp 263-267, June 1979.

75) Neuburg, N.P. "Improvement of Voicing Decisions by use of Context", IEEE proc., ICASSP, pp 5-8, April 1978.

76) Judd, M.W. Private Communication, August 1979, JSRU, Cheltenham, Glos, UK.

77) Schroeder, M.R. Private Communication, Gottingen University, W.Germany, November 1979.

78) Koh, S.N. "Digital Speech Transmission System with Quality Enhancement by High and Low Frequency Regeneration", M.Sc. Thesis, Department of Electronic and Electrical Engineering, Loughborough University, UK, September 1981.

79) Jayant, N.S. "Adaptive Quantization with One-Word Memory", BSTJ, Vol 52, September 1973.

80) Miller, N.J. "Pitch Detection by Data Reduction", IEEE Trans. on ASSP, Vol 23, No 1, February 1975.

81) Miller, R.L. "Nature of the Vocal Cord Waves", JASA, Vol 31, pp 667-677, June 1979

82) Patrick, P.J., Xydeas, C.S. "Speech Quality Enhancement by High Frequency Band Generation", IERE Conf. on Digital Processing of Signals in Communications, Proc No 49, pp 365-373, April 1981.

83) Thrane, N. "The Discrete Fourier Transform and FFT Analysers", Technical Review, Bruel & Kjaer, No 1, pp 3-25, 1979.

84) Harris, F.J. "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform", IEEE Proc., Vol 66, No 1, pp 51-83, January 1978.

85) Niederjohn, R.J., Curtis, R.A. "The Development of a Computer Processing System and its Use for the Study and Development of Processing Methods for Enhancing the Intelligibility of Speech in Noise", In-house Report No RADC-TR-77-310, Rome Air Development Centre, Griffis AFB, New York 13441, October 1977.

86) Fant, G.M. Lindquist, J. Sonnesson Hollien, H. "Speech Distortion at High Pressures" Underwater Physiology Proc., 4th Symp. on Underwater Physiology Penn., pp 293-299, 1971

87) Evci, C.C. "Prediction Techniques Applied to Differential Pulse Code Modulation Systems for Encoding Speech Signals", PhD Thesis, Department of Electronic and Electrical Engineering, Loughborough University, Leics, April 1982

88) Evci, C.C.
Xydeas, C.S.
Patrick, P.J.
"Wideband Quality ADPCM-AQF Speech Digitisers for bit rates of 16 and 32 kbits/s", Neuvieme Colloque sur le Traitement, D.V. Signal et ses Applications, GRETSI-83, Nice, France, pp 487-492, May 1983

89) Cham, C.K.
"A Computer Simulation Study of a Digital Speech Transmission System Using a Speech Bandwidth Compression Algorithm", M.Sc. Thesis, Department of Electronic and Electrical Engineering, Loughborough University, September 1980.

90) Patrick, P.J.,
Xydeas, C.S.
Steele, R.,
Cham, C.K.
"Wideband Quality Speech Encoder with Bit Rates of 16-32 kbits/s"
IEEE Proc. ICASSP, pp 844-847
Atlanta, Georgia, USA, April 1981.

91) Tse, V.K.C.
"Enhancement of Speech Signals obtained from a series combination of a 16 kbits/s Waveform Coder and a 2.4 kbits/s Vocoder", B.Sc. Thesis, Department of Electronic and Electrical Engineering, Loughborough University, Leics, April 1981

92) Patrick, P.J., Steele, R., Xydeas, C.S. "Voiced/Unvoiced Band-Switching System for Transmission of 6kHz Speech over 3.4kHz Telephone Channels", IERE, The Radio and Electronic Engineer, Vol 51, No 5, pp 233-235, May 1981.

93) Patrick, P.J. Steele, R. Xydeas, C.S. "Frequency Compression of 7.6 kHz Speech into 3.3 kHz Bandwidth" IEEE proc., ICASSP, Vol 3, Boston, Mass., USA, pp 1304-1307, April 1983

94) Patrick, P.J. Steele, R. Xydeas, C.S. "Frequency Compression of 7.6 kHz Speech into 3.3 kHz Bandwidth", IEEE Trans. Comms., Vol 31, No 5, pp 692-701, May 1983

95) Holmes, J.N., Judd, M.W., Walesby, D.H. "A High-Quality All-Digital Sound Spectrograph Developed for Speech Signal Analysis", Proc., IEEE Int. Conf. ASSP 78, Tulsa, 1978.

96) Hunt, J.H., Bridle, J.S., Holmes, J.N. "Interactive Digital Inverse Filtering and its Relation to Linear Prediction Methods", Proc., IEEE, ICASSP, pp 15-18, 1978.

97) Linggard, R., Smith, B.D.V. "A Family of Phase Complementary Filters" Proc., IEEE, ICASSP, pp 178-181, 1979.

98) Sonagram — Kay Elemetrics Co., Type B/65, Wessex Electronics Ltd., Bristol

99) Ackroyd, M.H. — "Digital Filters" (book), Butterworths, London, 1973.

100) Rabiner, L.R., Gold, B. McGonegal, C.A. — "An Approach to the Approximation Problem for Non-recursive Digital Filters" IEEE, Transactions on Audio and Electro-acoustics, Vol 18, No 2, pp 83-105, June 1970.

101) Rabiner, L.R. Gold, B. — "Theory and Applications of Digital Signal Processing" (book) Prentice-Hall, New Jersey, 1975.

102) Pirogov, A.A. — "A Harmonic System for Compressing Speech-Spectra", Elektroviaz No.3, pp 8-17, 1959. Also, Telecommunications No.3, pp 229-242, 1959.

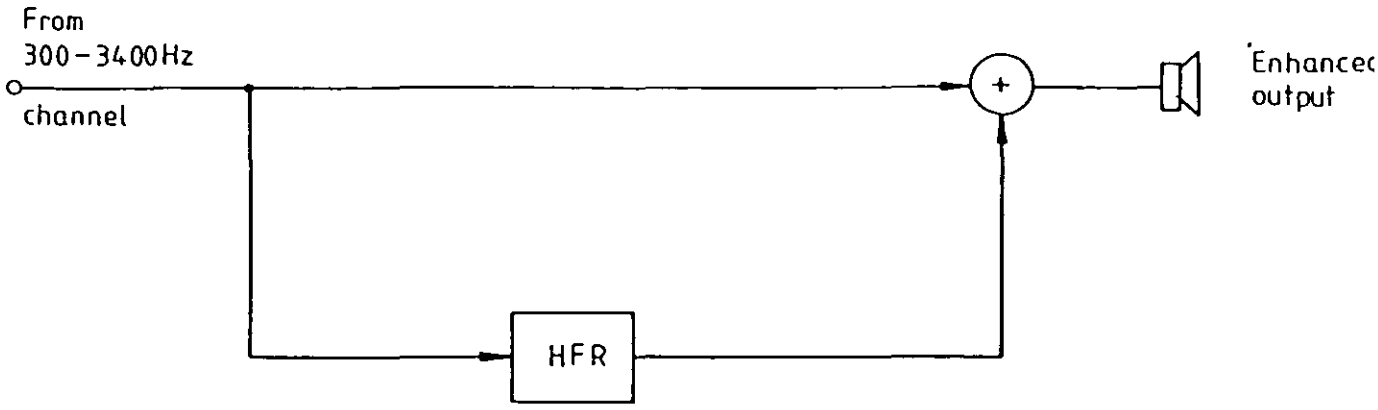


FIG. 5·1 BASIC FORM OF THE BANDWITH ENHANCEMENT PROCESS

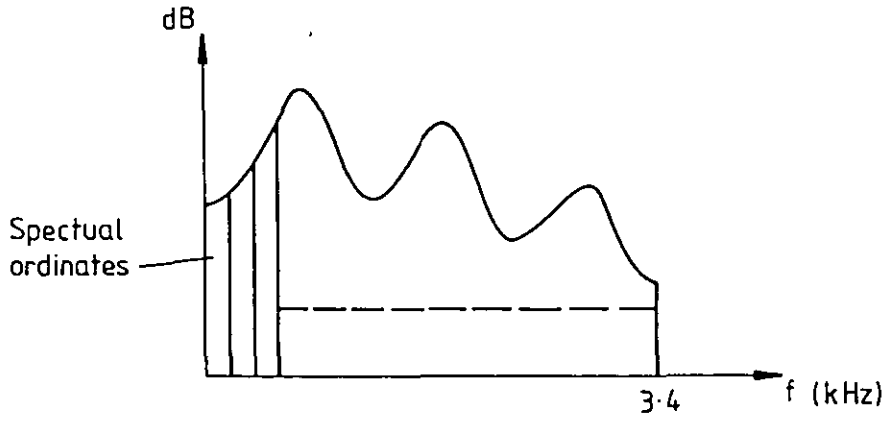


FIG. 5·2 (a) IN-BAND SPECTRUM

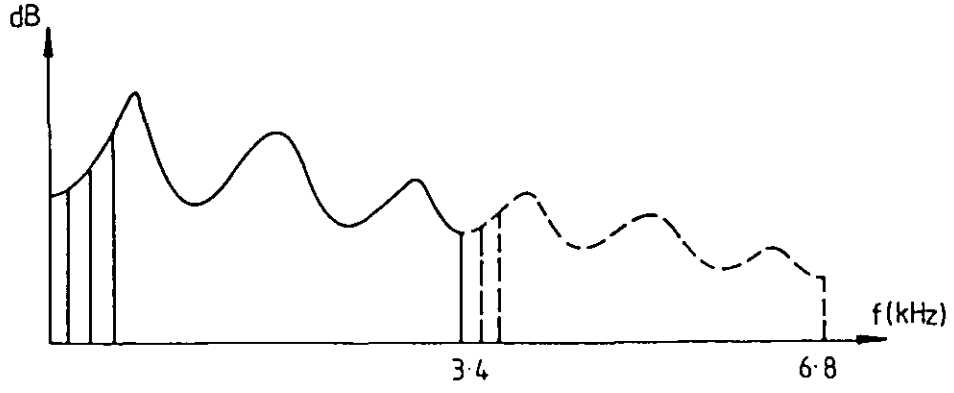


FIG. 5·2 (b) DUPLICATED SPECTRUM

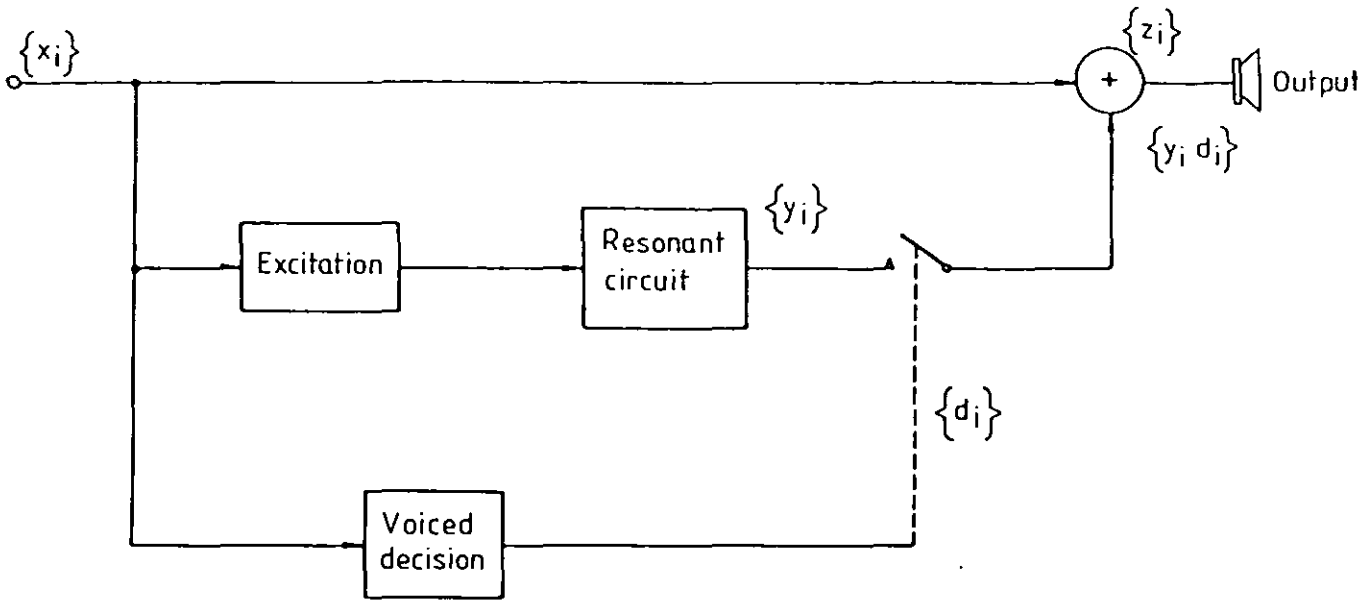


FIG. 5·3

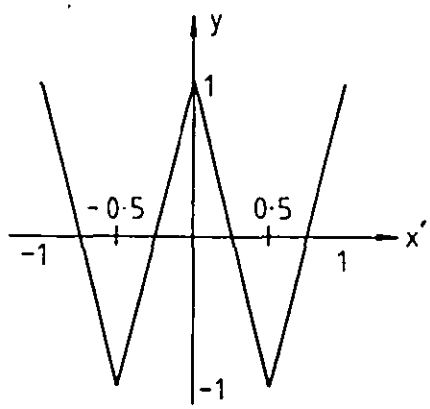


FIG. 5·4

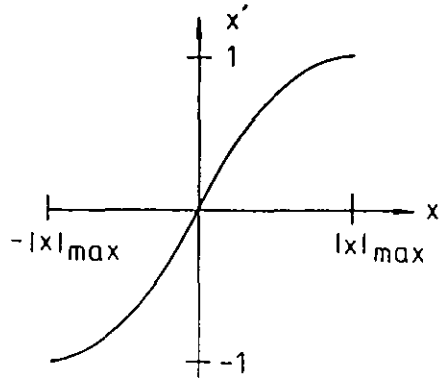


FIG. 5·5

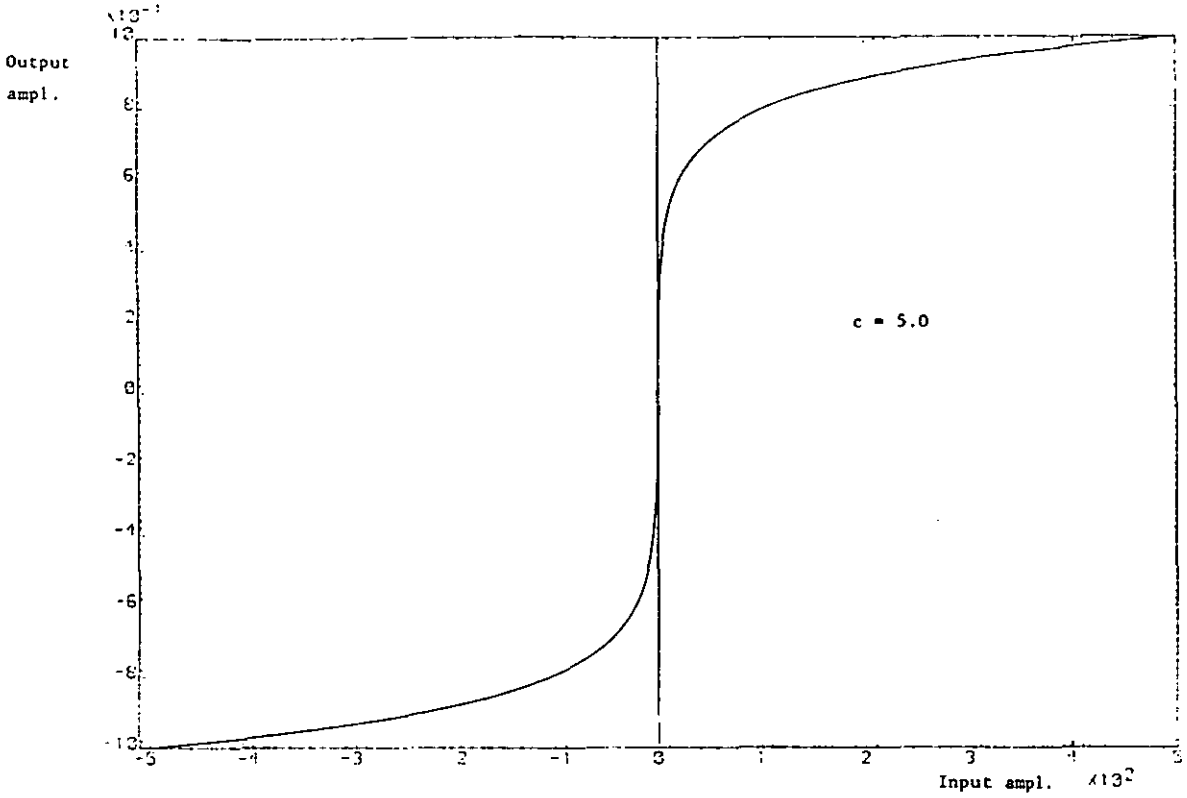


Fig5.6(a) Transfer function of logarithmic compressor.

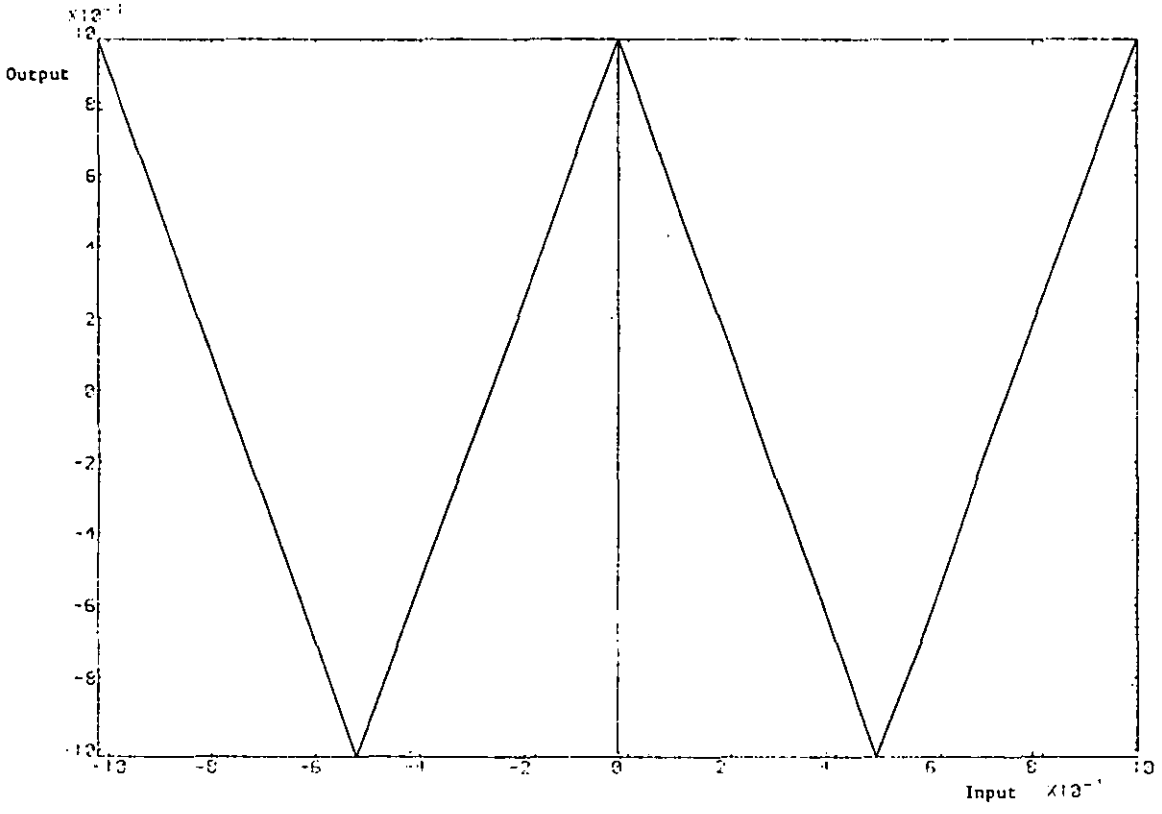


Fig 5.6(b) Input/output response of 'W' transfer function.

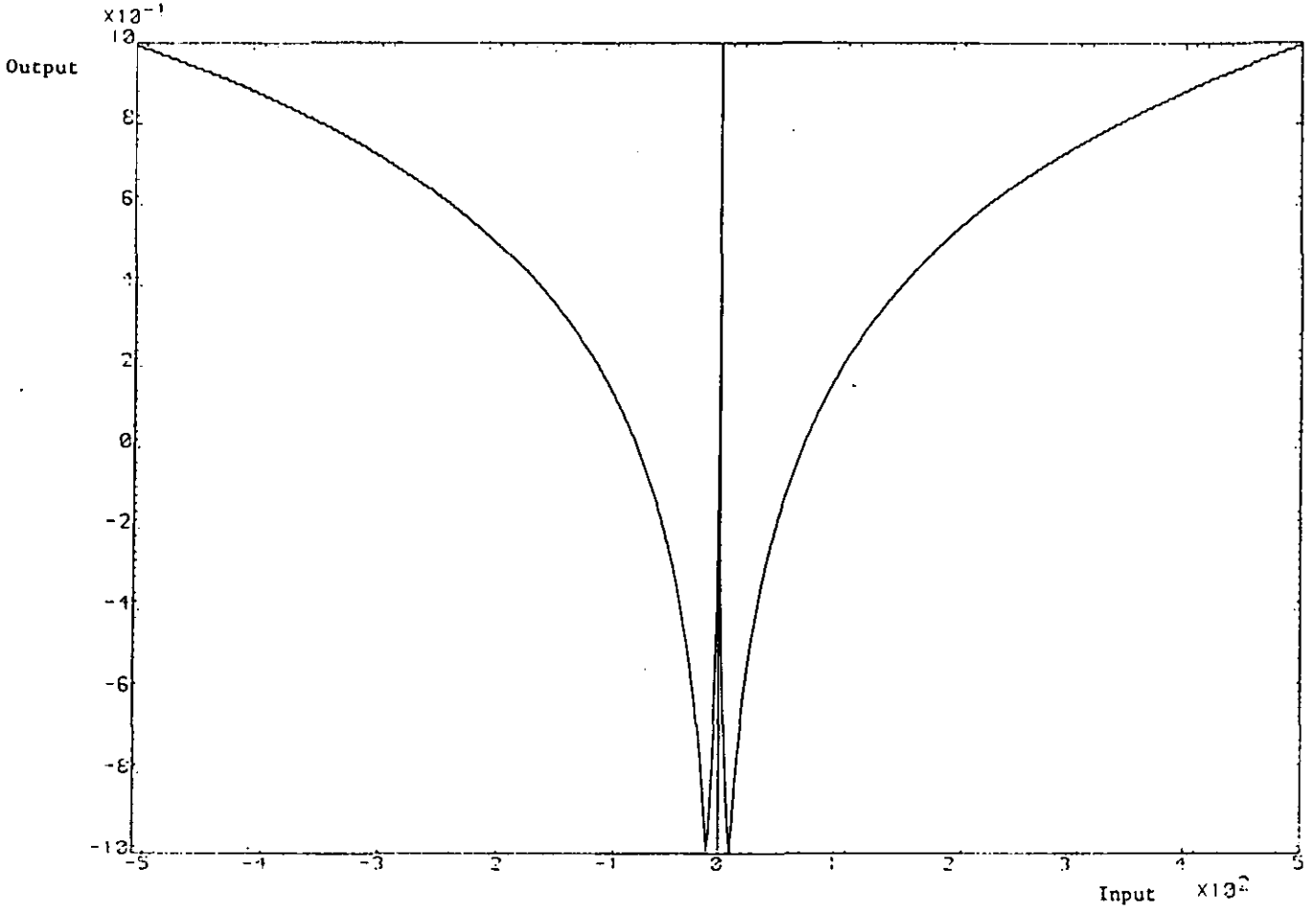


Fig 5.7 Combined response of logarithmic compressor and the 'W' transfer function.

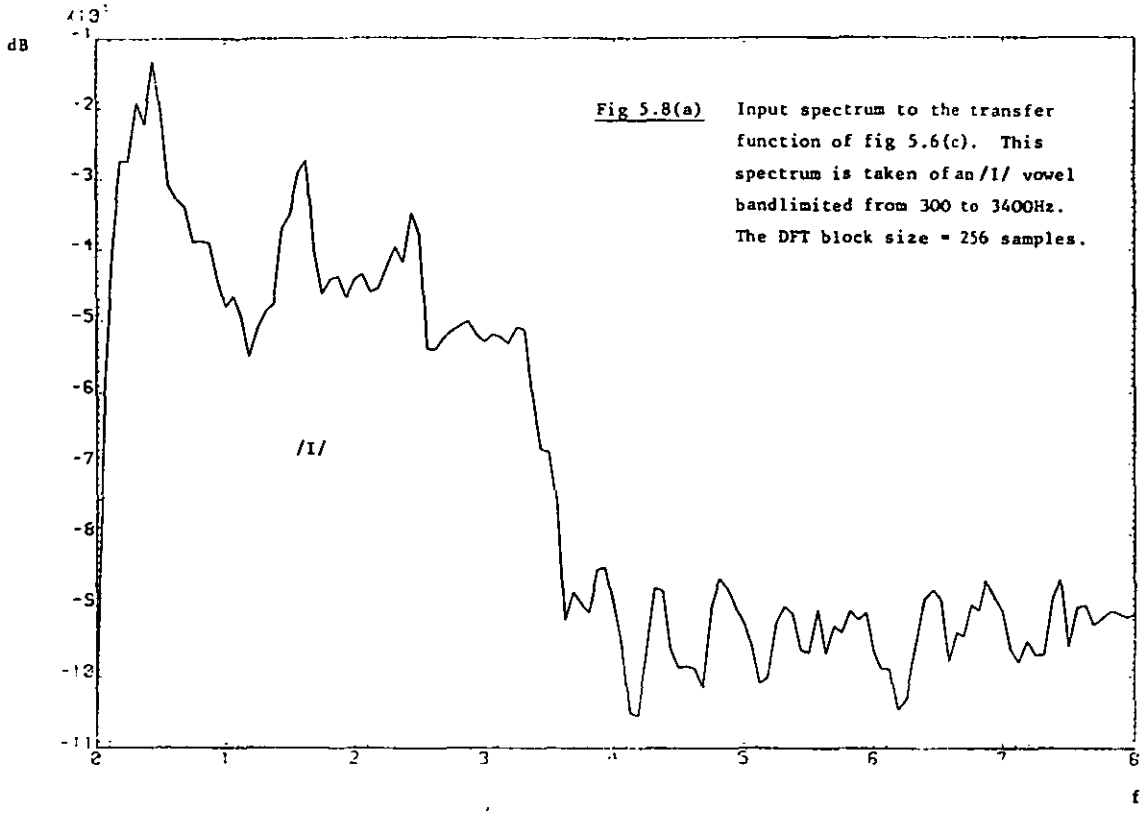


Fig 5.8(a) Input spectrum to the transfer function of fig 5.6(c). This spectrum is taken of an /I/ vowel bandlimited from 300 to 3400Hz. The DFT block size = 256 samples.

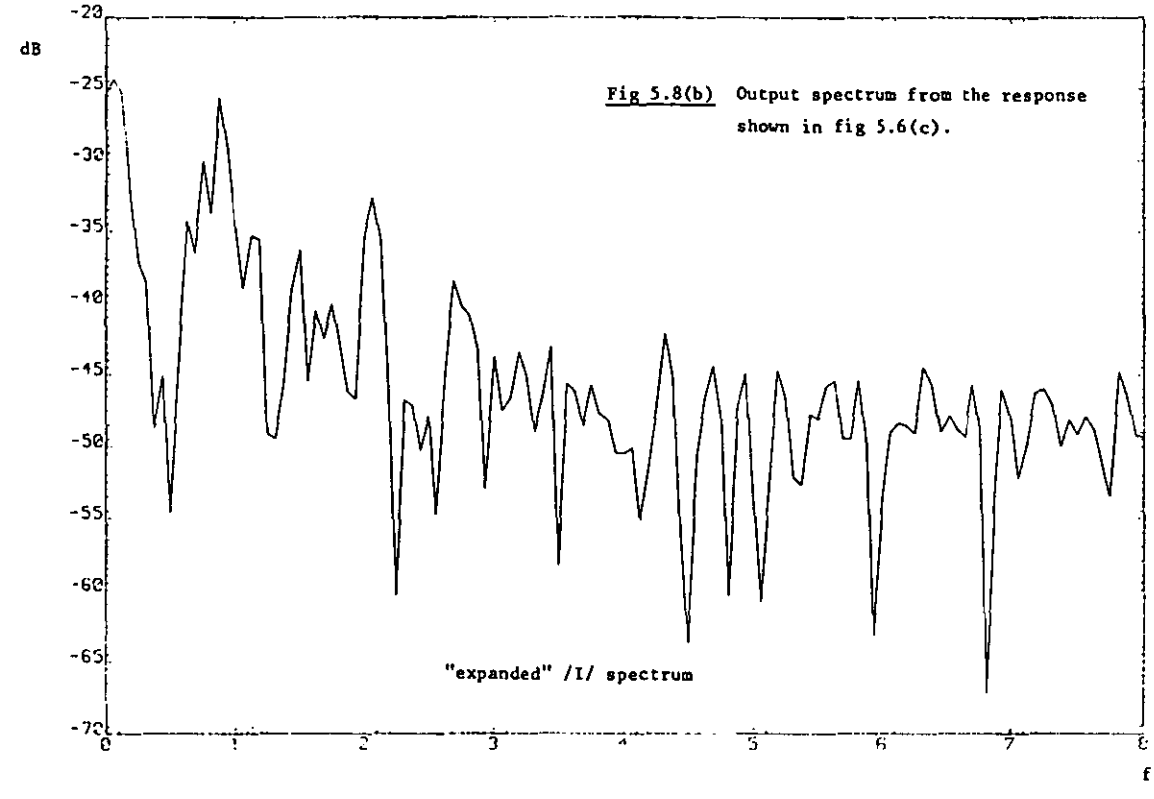


Fig 5.8(b) Output spectrum from the response shown in fig 5.6(c).

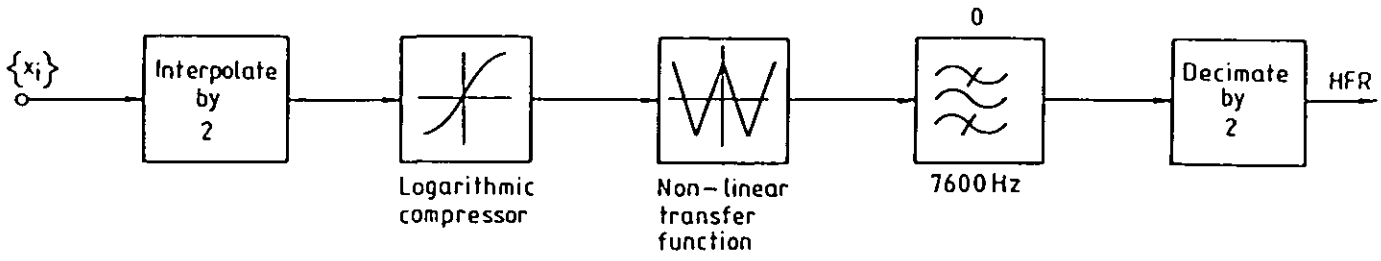


FIG. 5·9

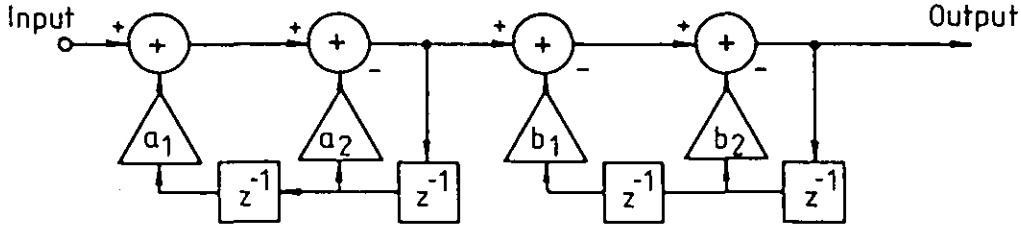
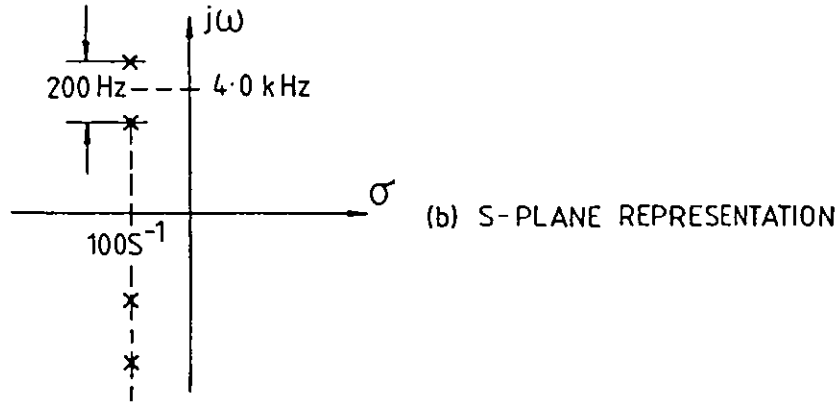


FIG. 5·10 (a) IMPLEMENTATION



(b) S-PLANE REPRESENTATION

FIG. 5·10 RESONANT NETWORK

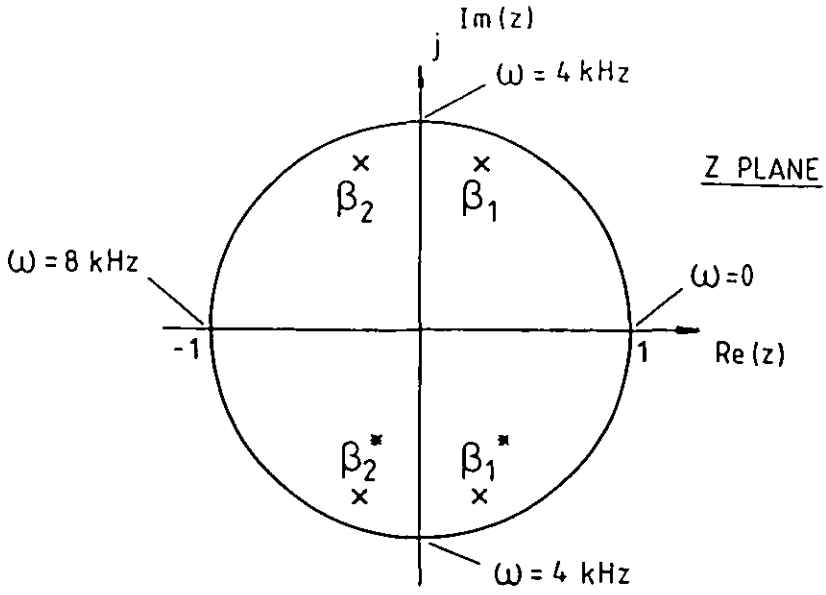


FIG. 5·11

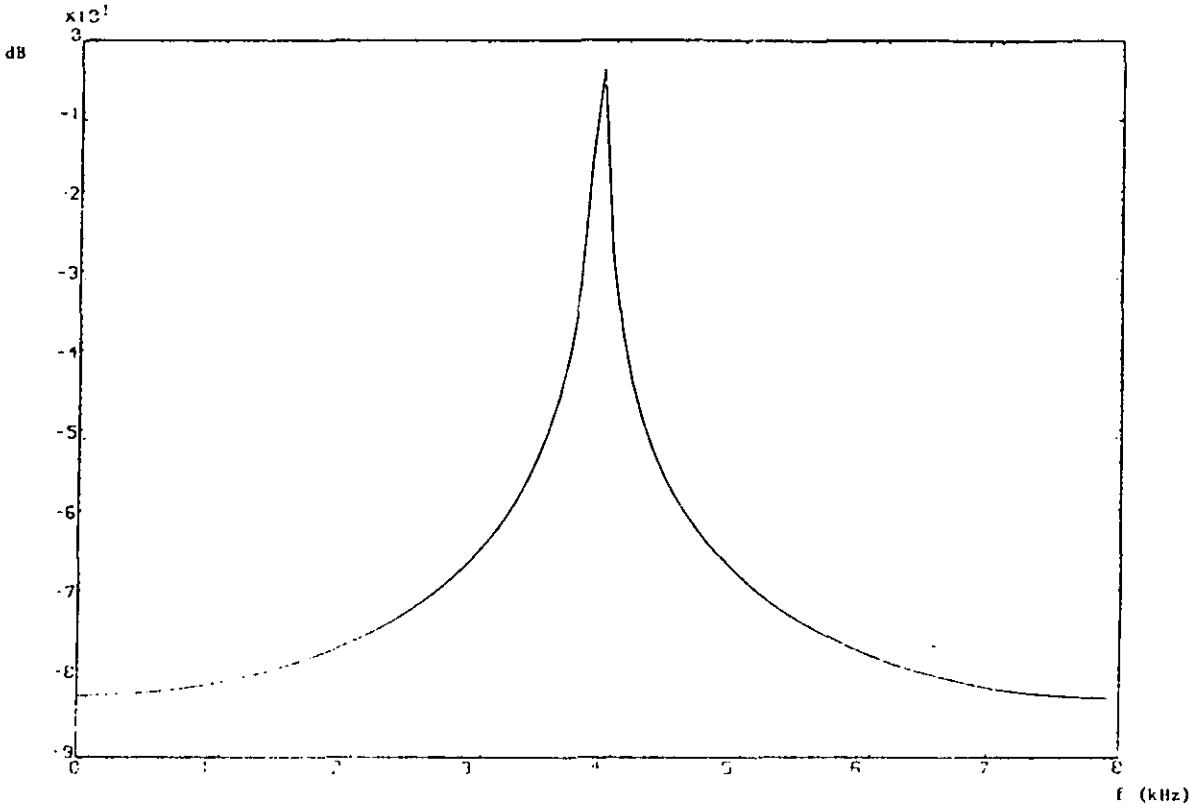


Fig 5.12(a) Spectral response of the 4-pole recursive filter.

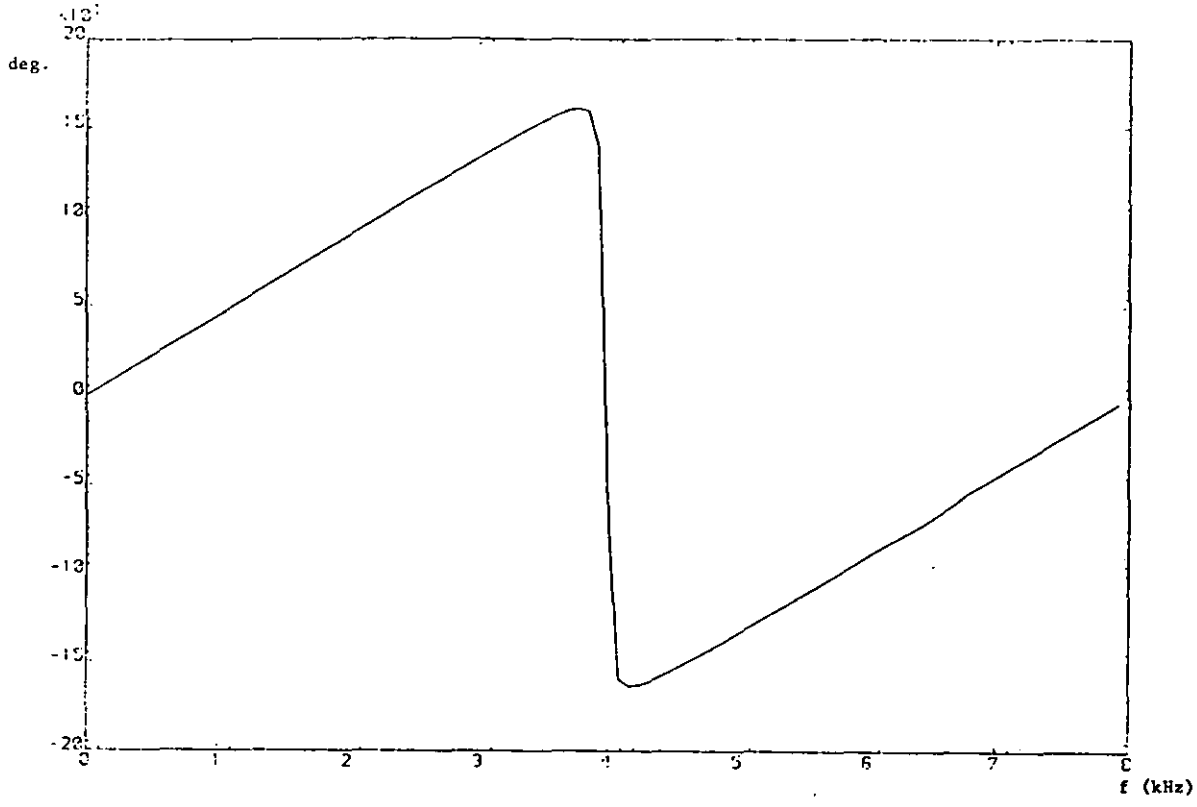


Fig 5.12(b) Phase response of the 4-pole recursive filter.

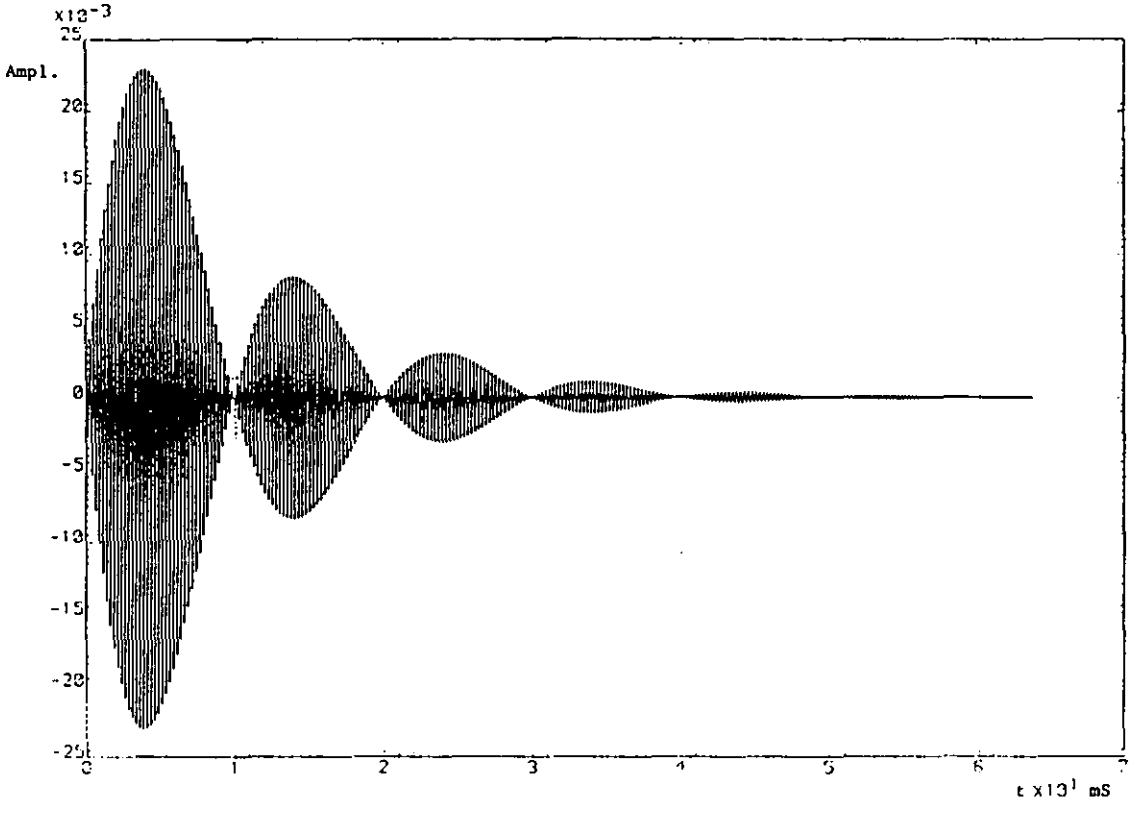


Fig 5.13 Impulse response of the 4-pole recursive filter.

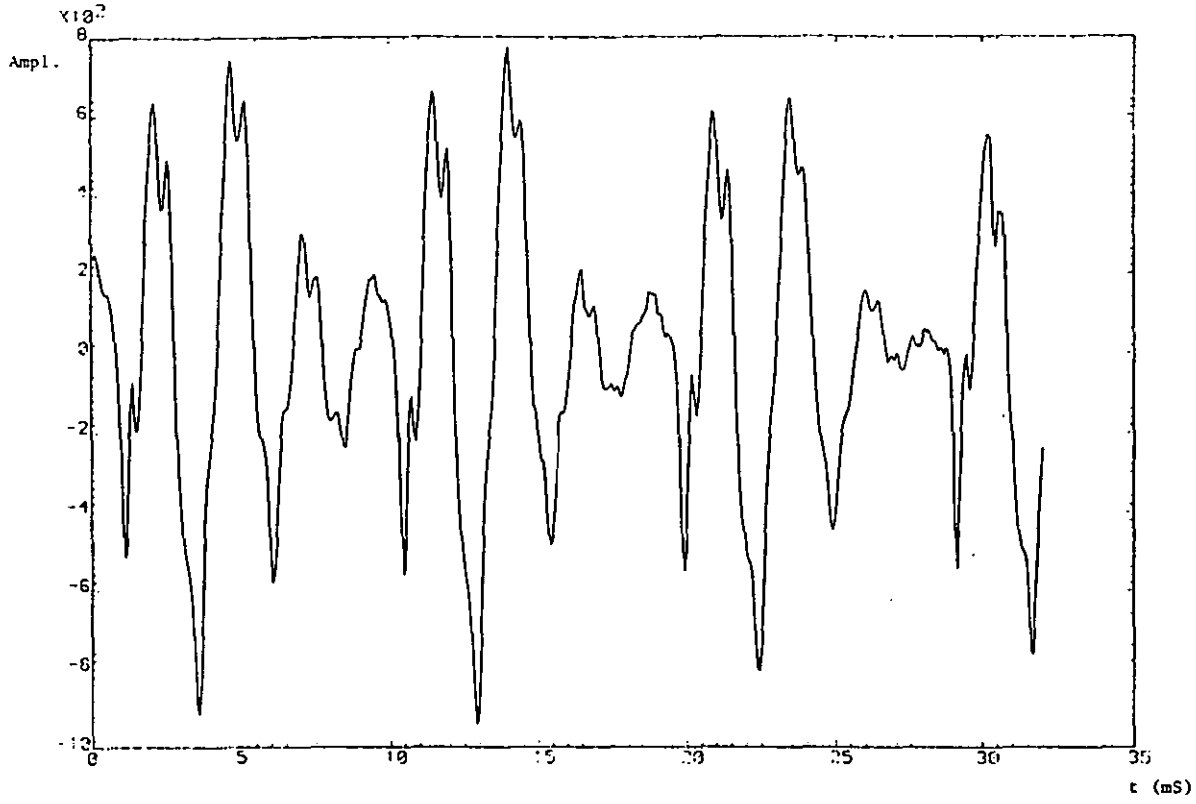


Fig 5.14(a) (i) Waveform of an /I/ vowel from "sister".
The signal is bandlimited from 300 to 7600Hz.

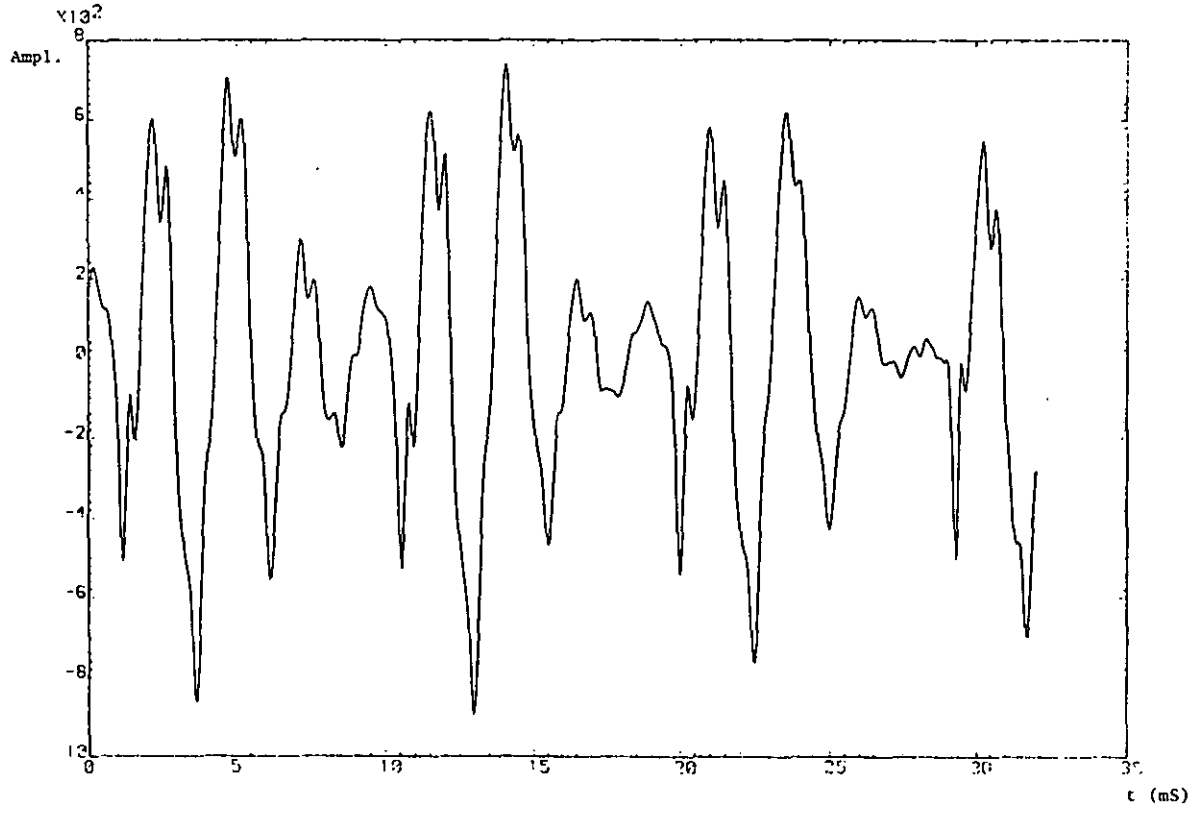


Fig 5.14(a) (ii) Waveform of an /I/ vowel bandlimited from
300 to 3400 Hz.

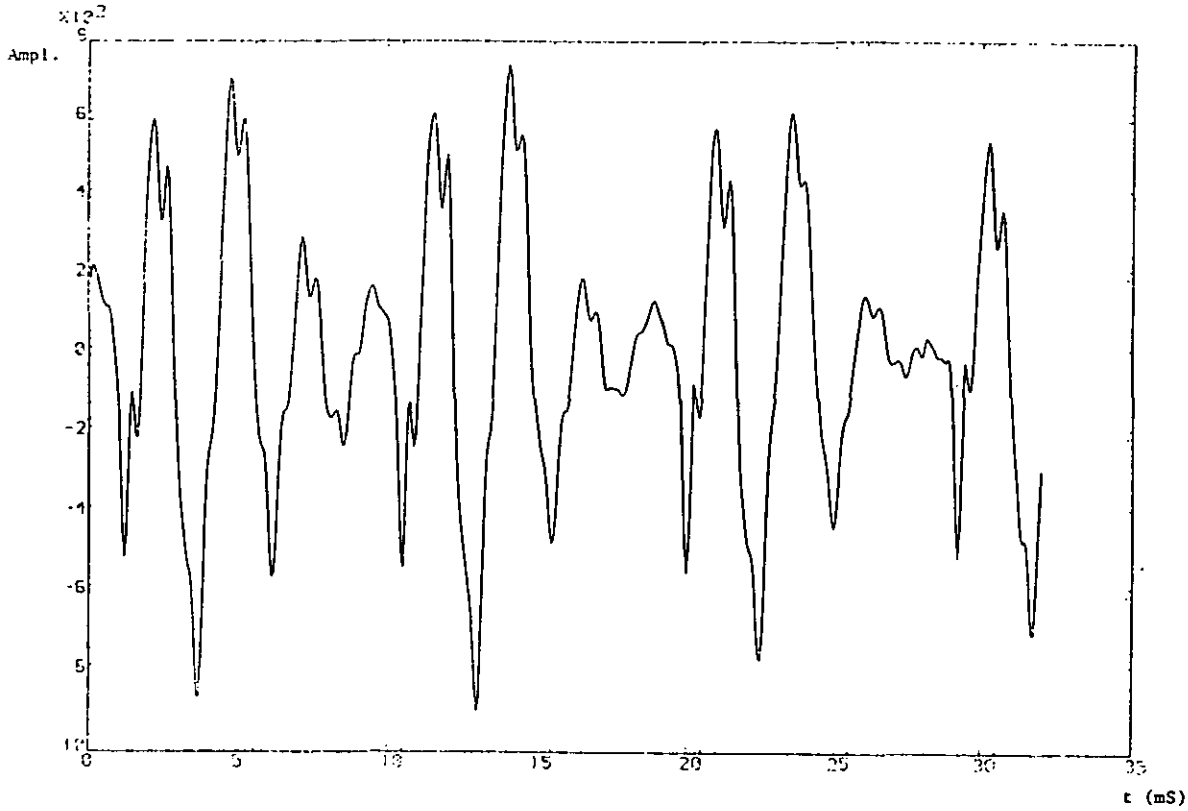


Fig 5.14(a) (iii) Waveform of a 300 to 3400Hz bandlimited vowel with an added "formant"

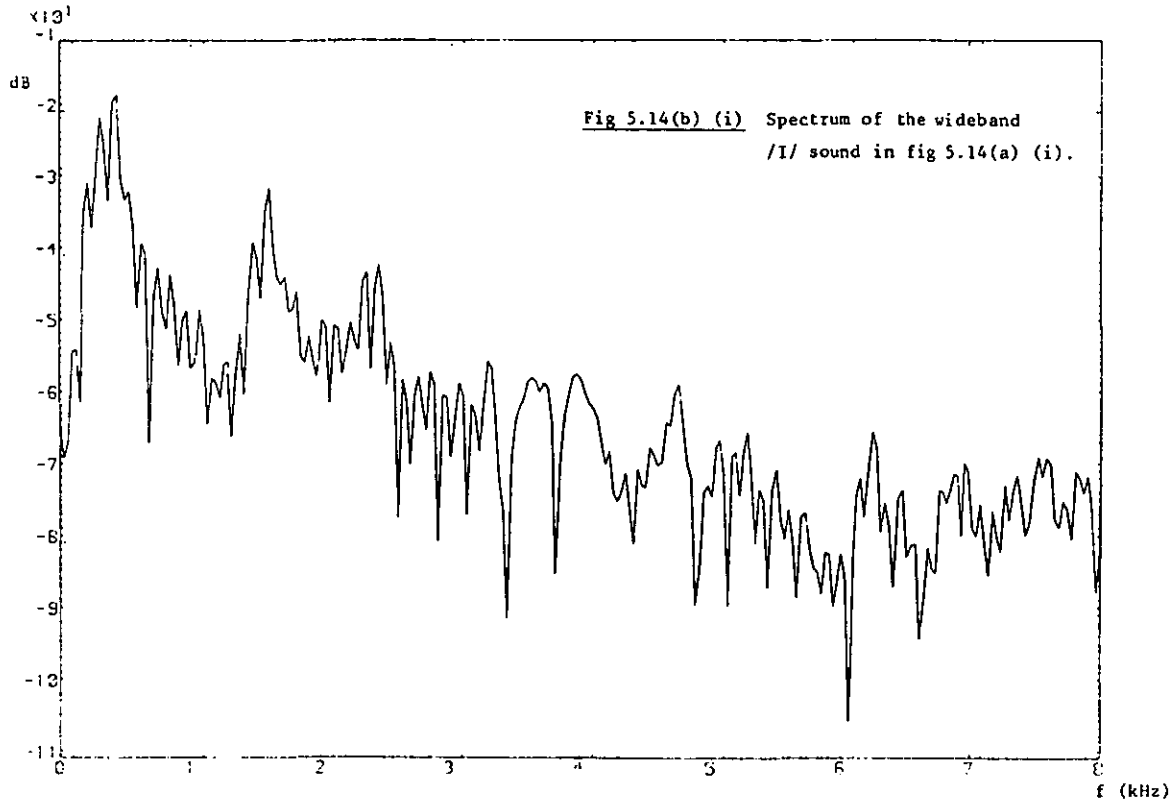


Fig 5.14(b) (i) Spectrum of the wideband /I/ sound in fig 5.14(a) (i).

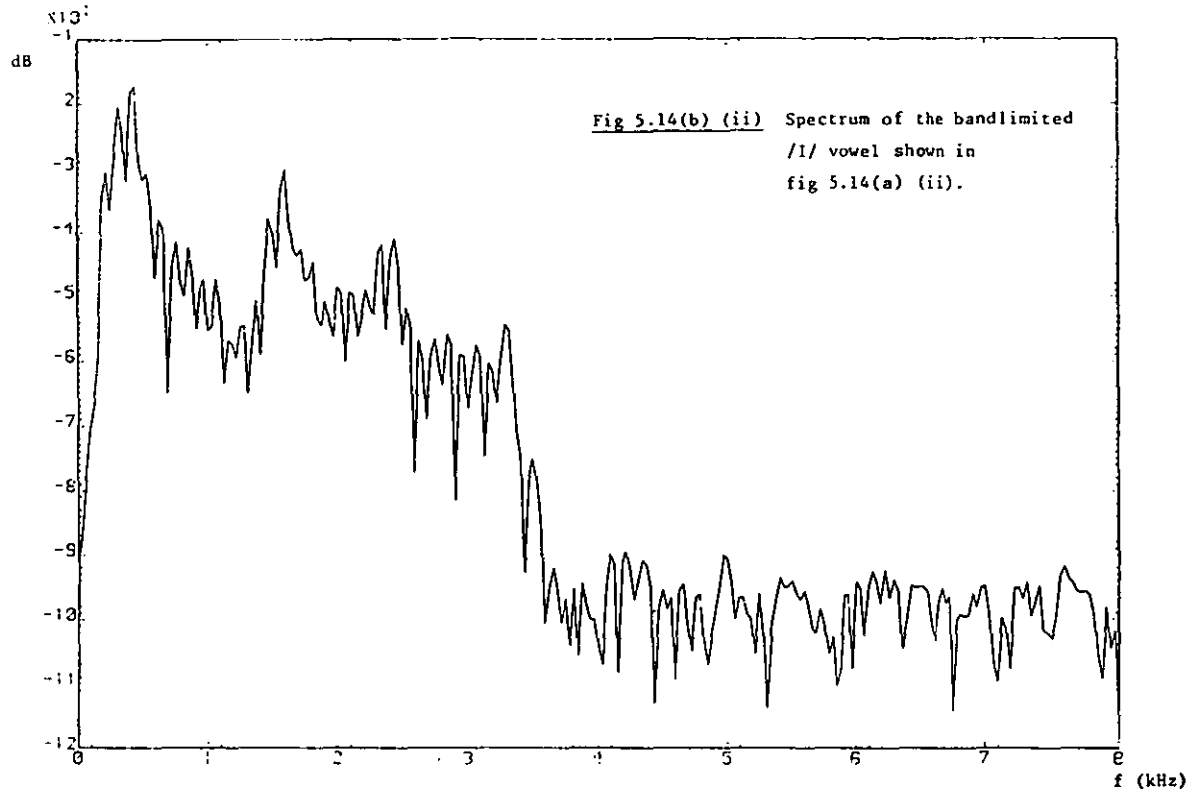


Fig 5.14(b) (ii) Spectrum of the bandlimited /I/ vowel shown in fig 5.14(a) (ii).

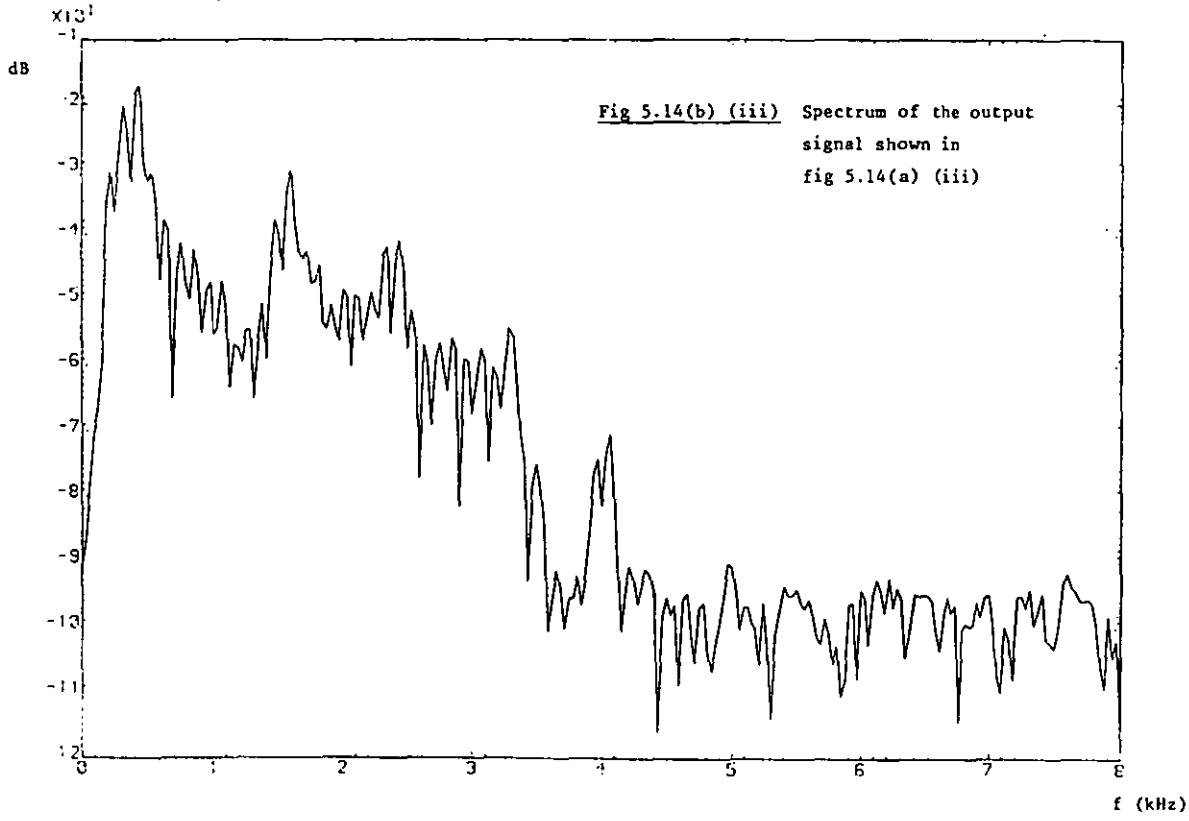


Fig 5.14(b) (iii) Spectrum of the output signal shown in fig 5.14(a) (iii)

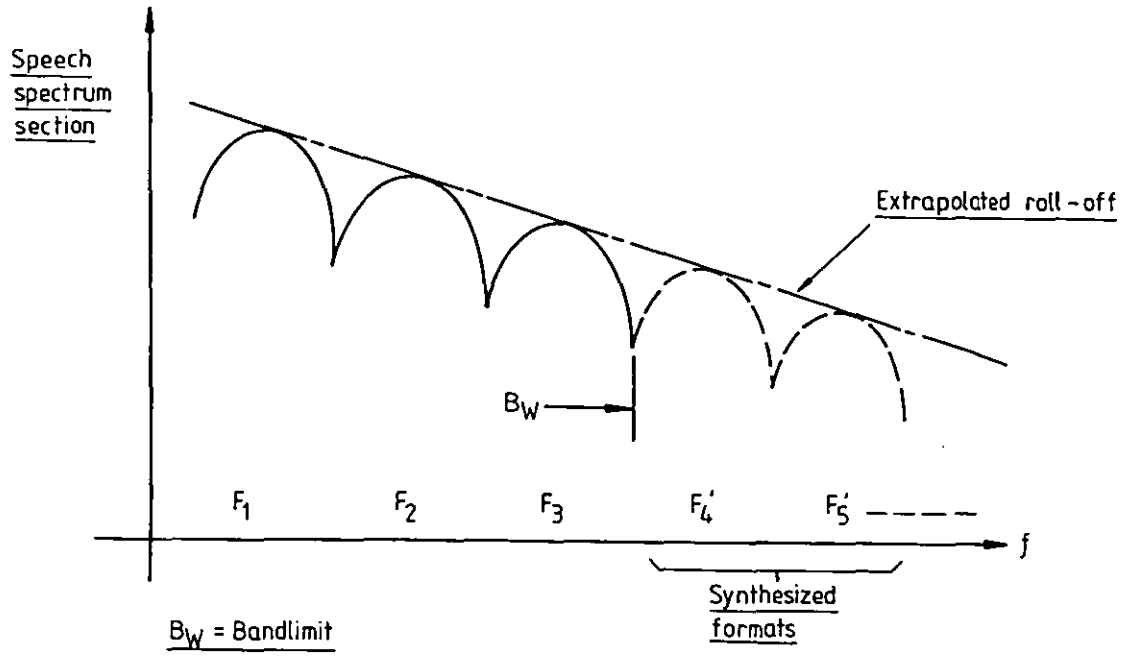


FIGURE 5·15

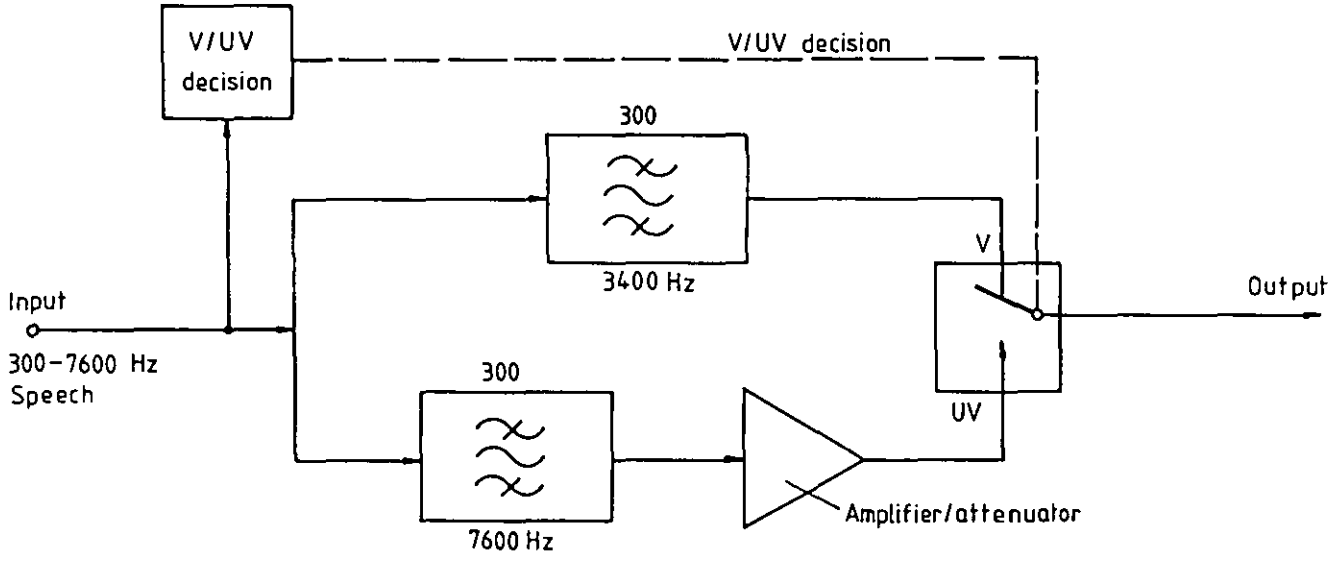


FIG. 5·16 PROTOTYPE ARRANGEMENT FOR APPLYING HFR TO INVOICED SPEECH

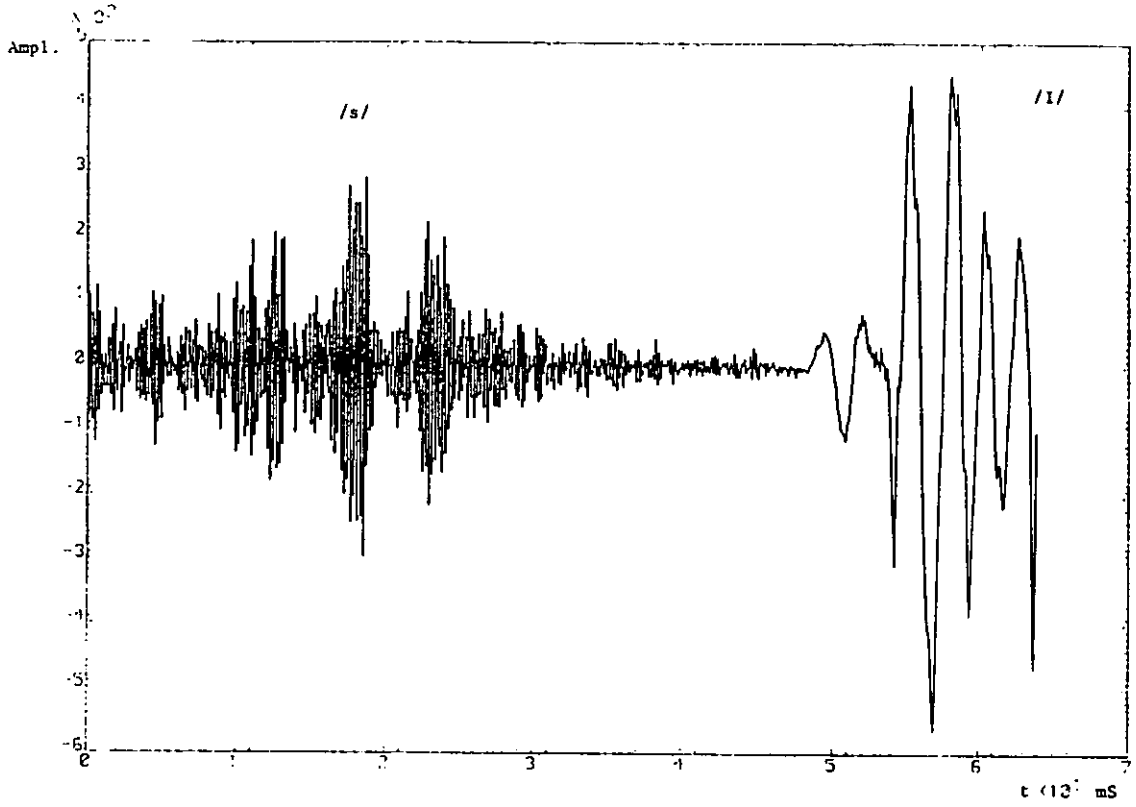


Fig 5.17(a) Original wideband signal.

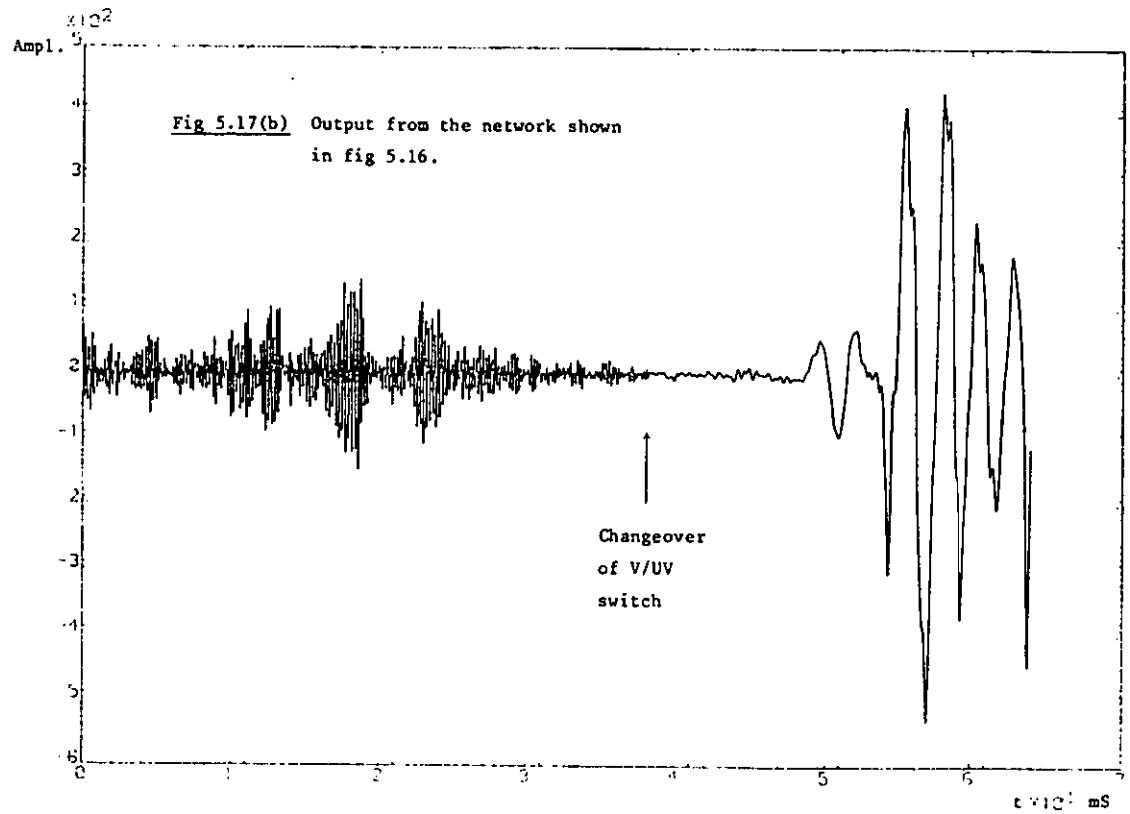
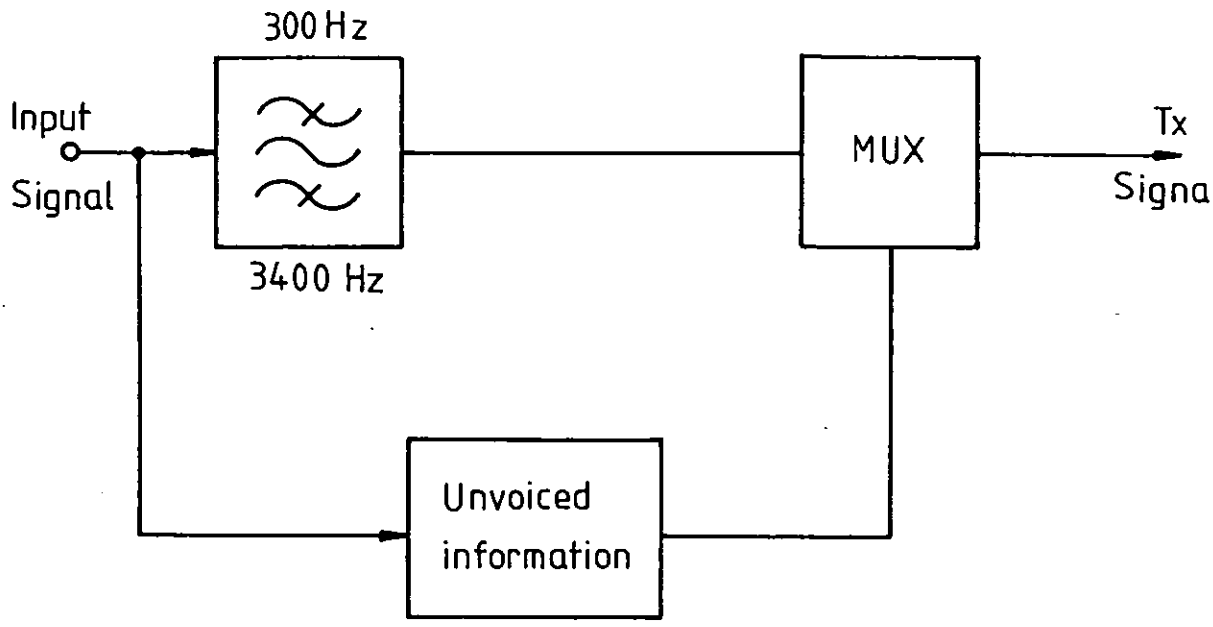
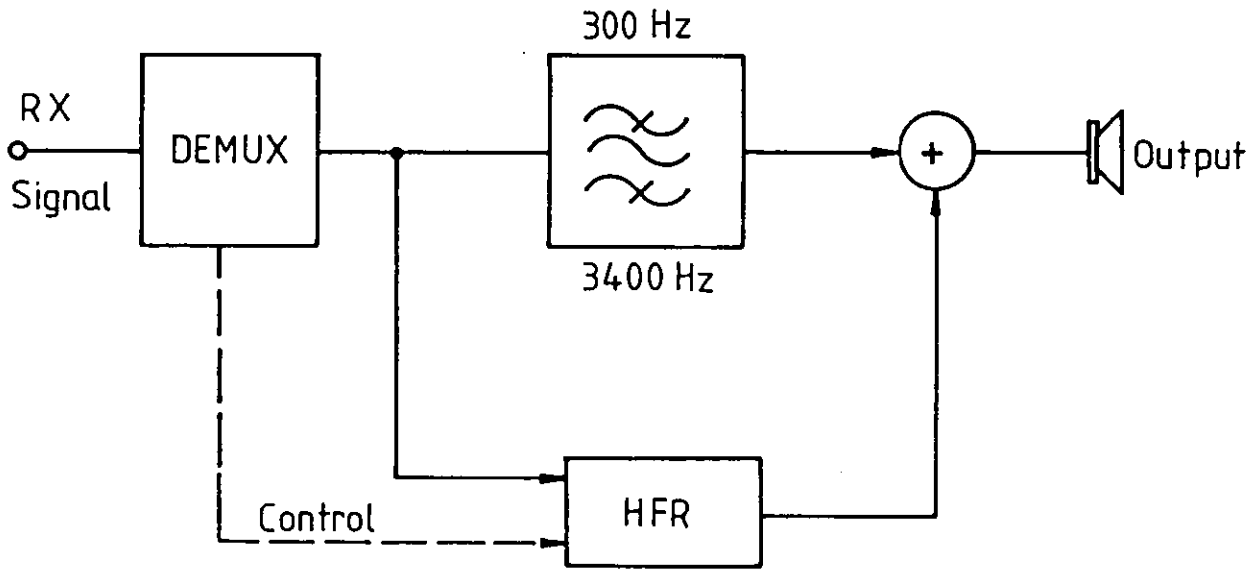


Fig 5.17(b) Output from the network shown in fig 5.16.



(a) TRANSMITTER



(b) RECEIVER

FIG. 5·18 HFR APPLIED TO UNVOICED SPEECH

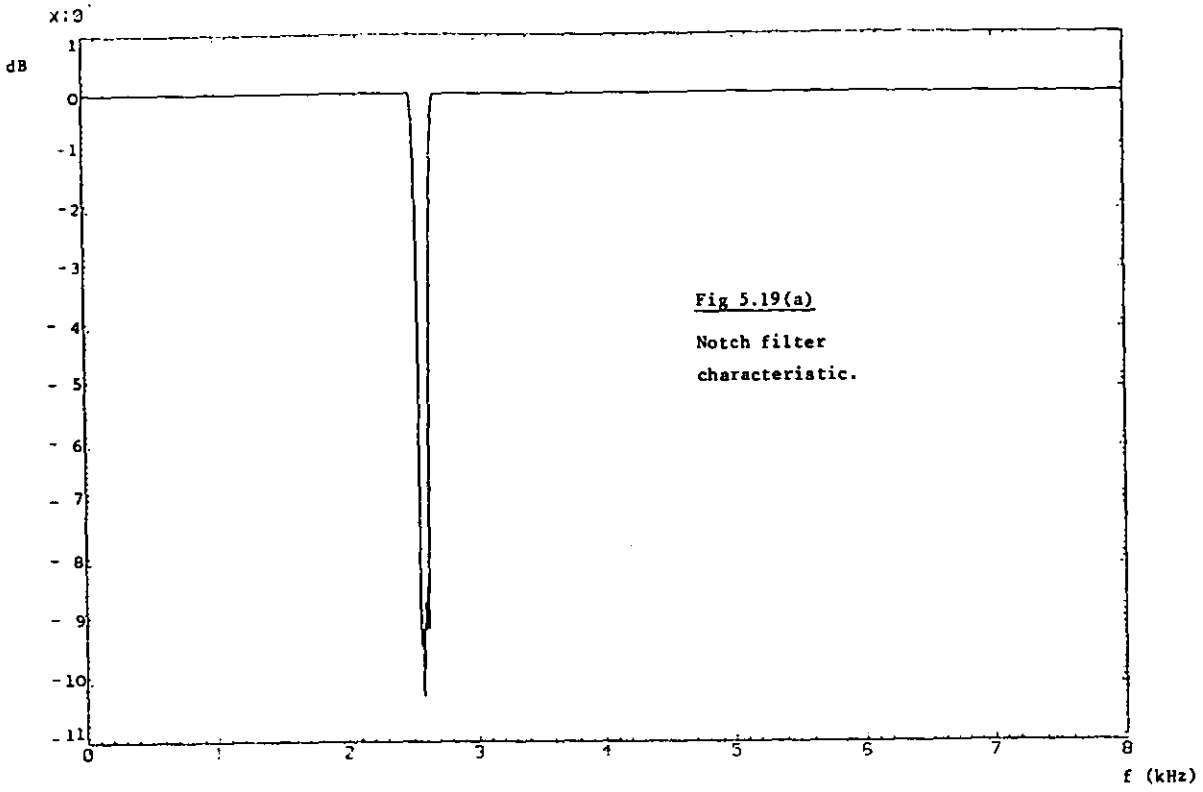


Fig 5.19(a)

Notch filter
characteristic.

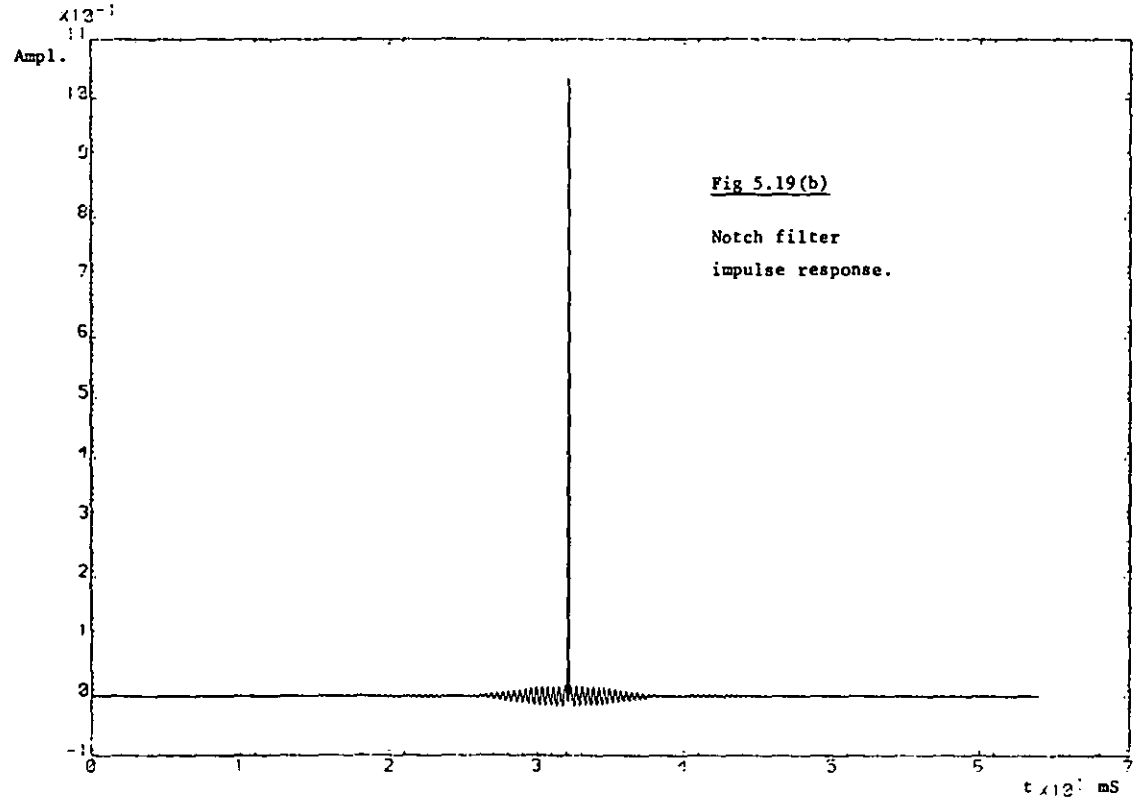


Fig 5.19(b)

Notch filter
impulse response.

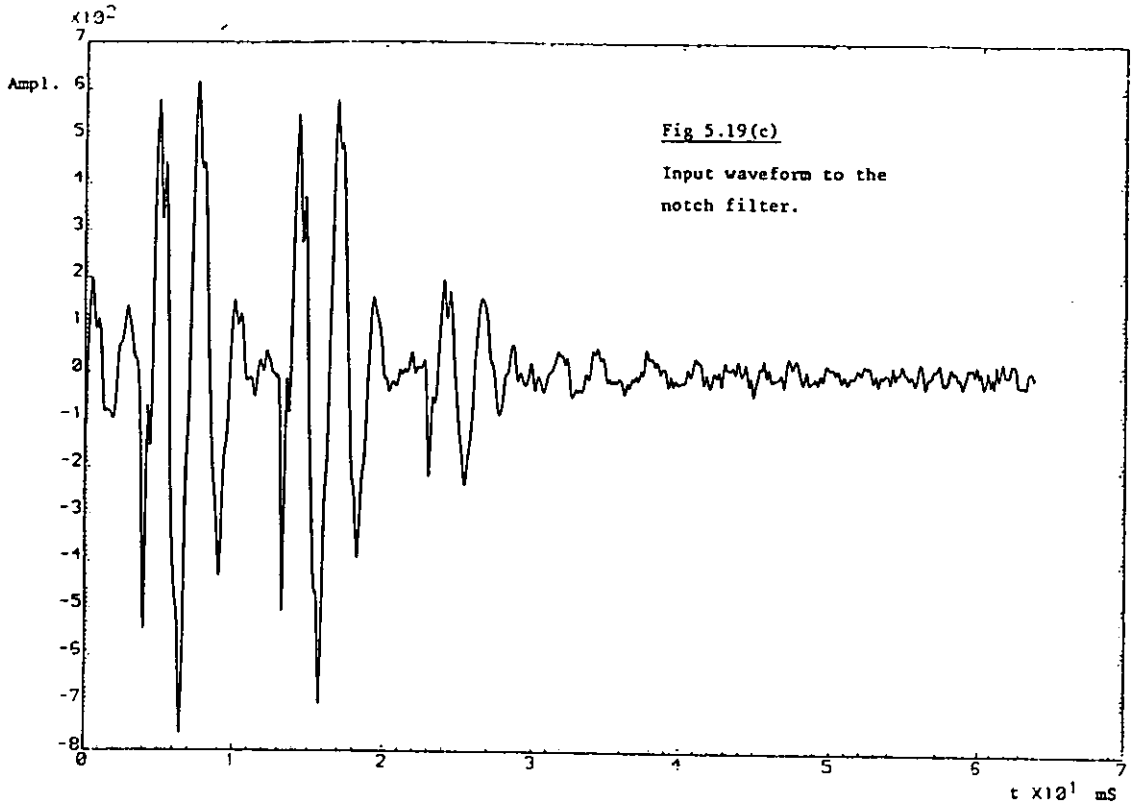


Fig 5.19(c)

Input waveform to the notch filter.

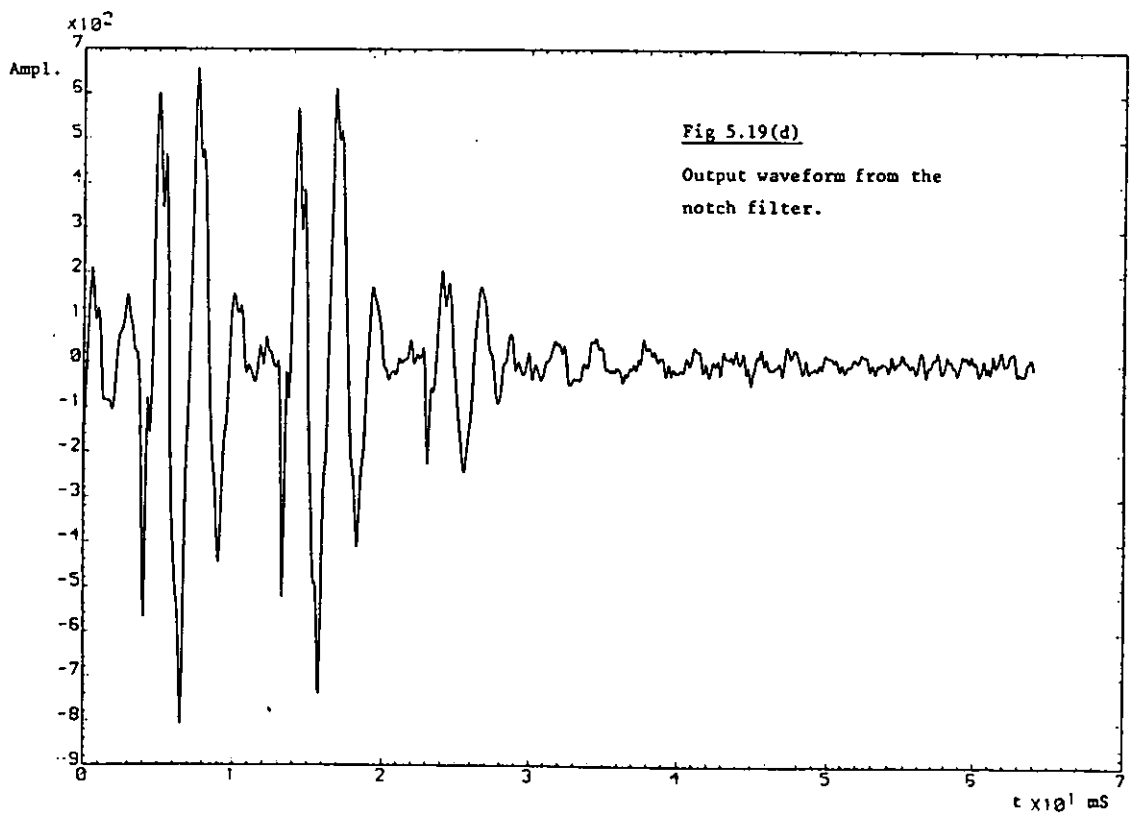


Fig 5.19(d)

Output waveform from the notch filter.

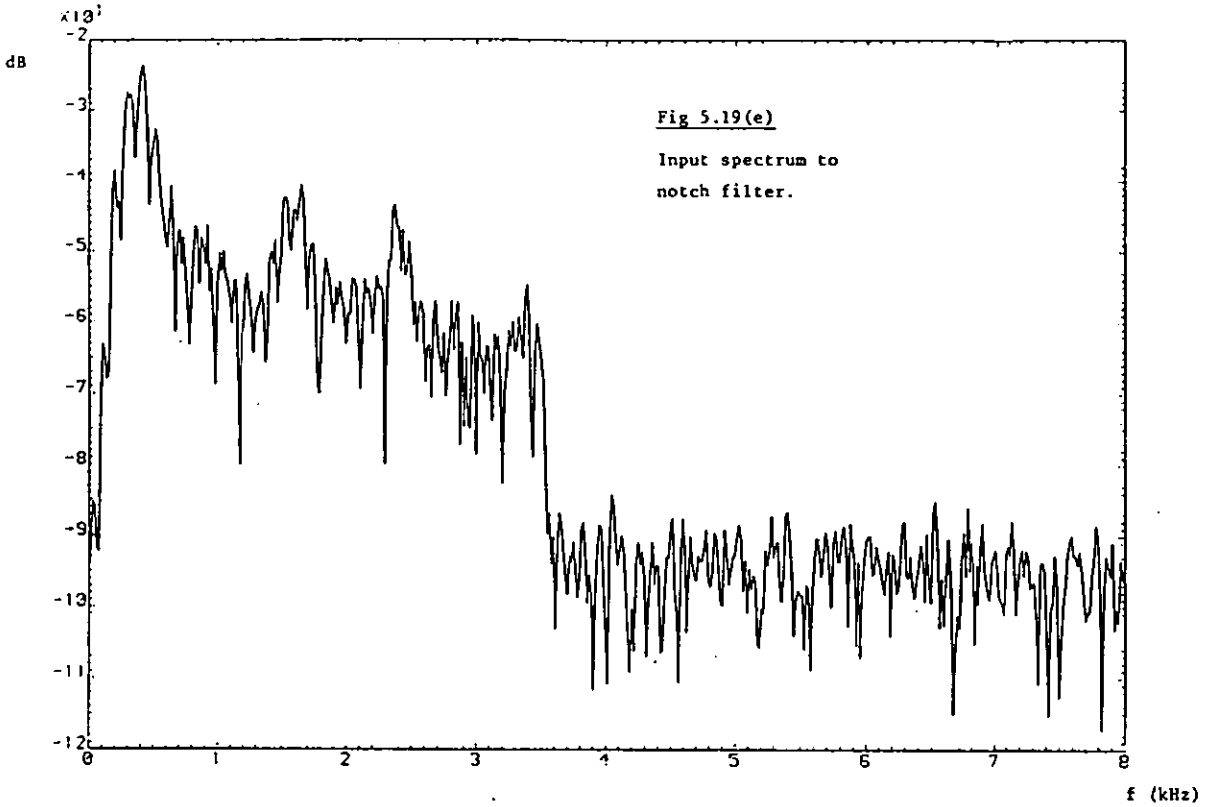


Fig 5.19(e)

Input spectrum to notch filter.

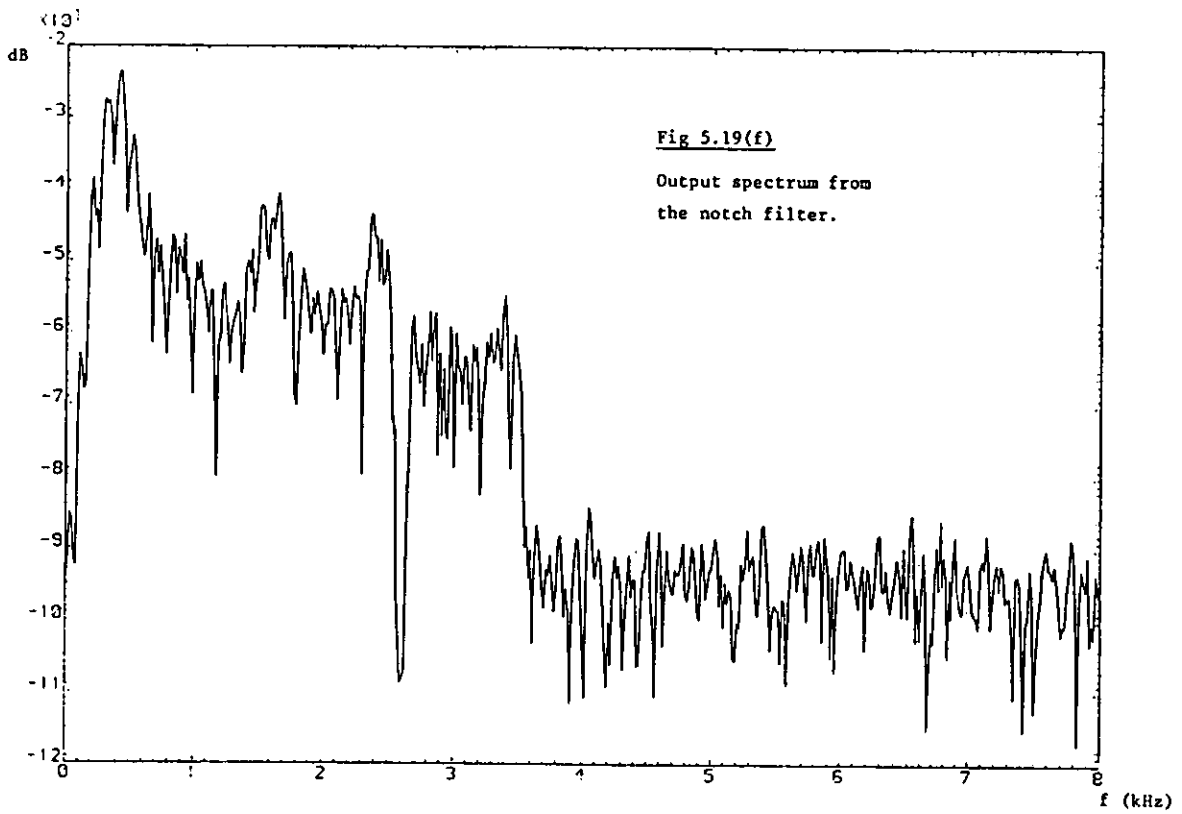


Fig 5.19(f)

Output spectrum from the notch filter.

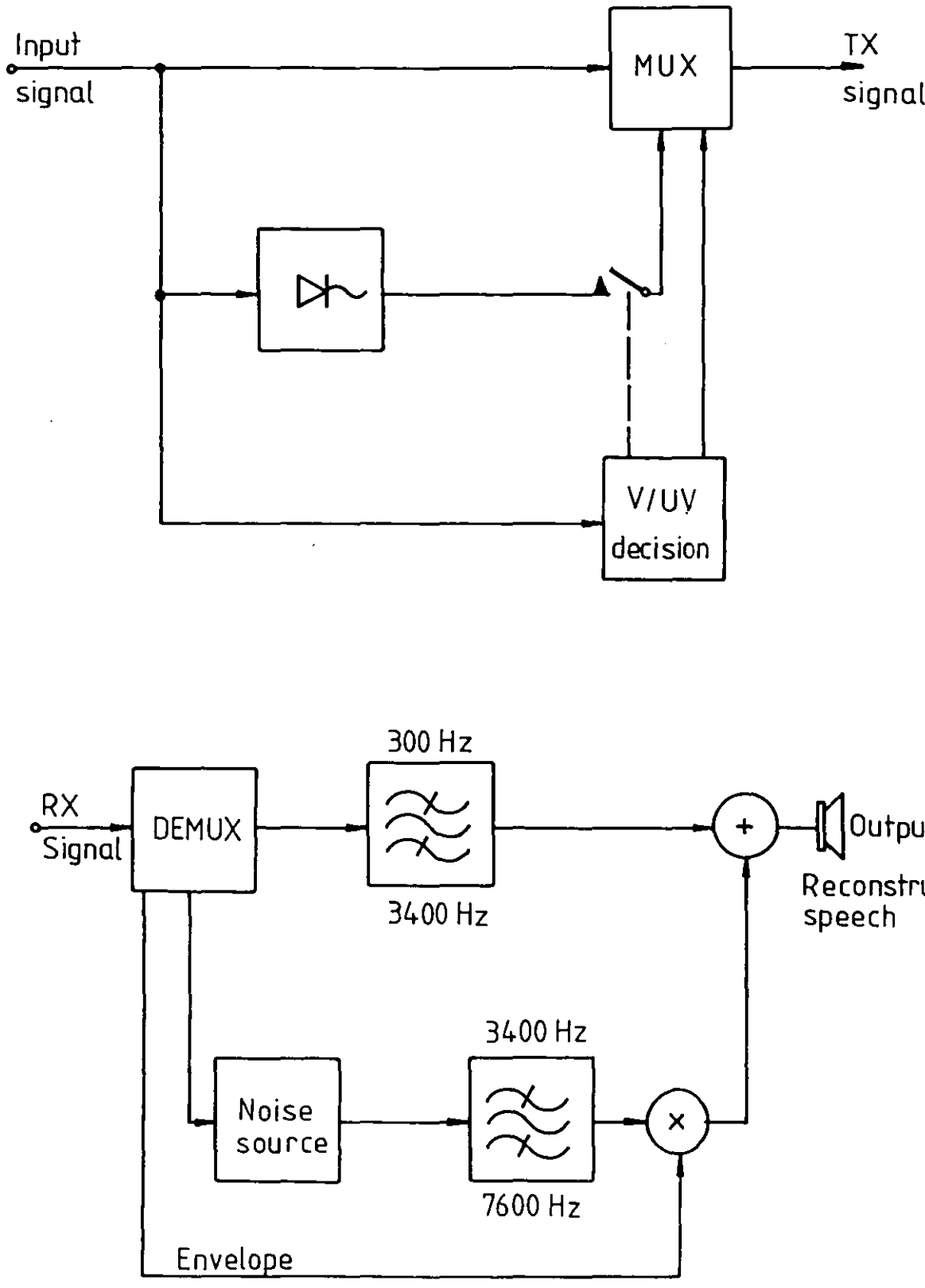


FIG. 5·20 HFR BY ADDING BANDPASS FILTERED NOISE TO BANDLIMITED UNVOICED SPEECH

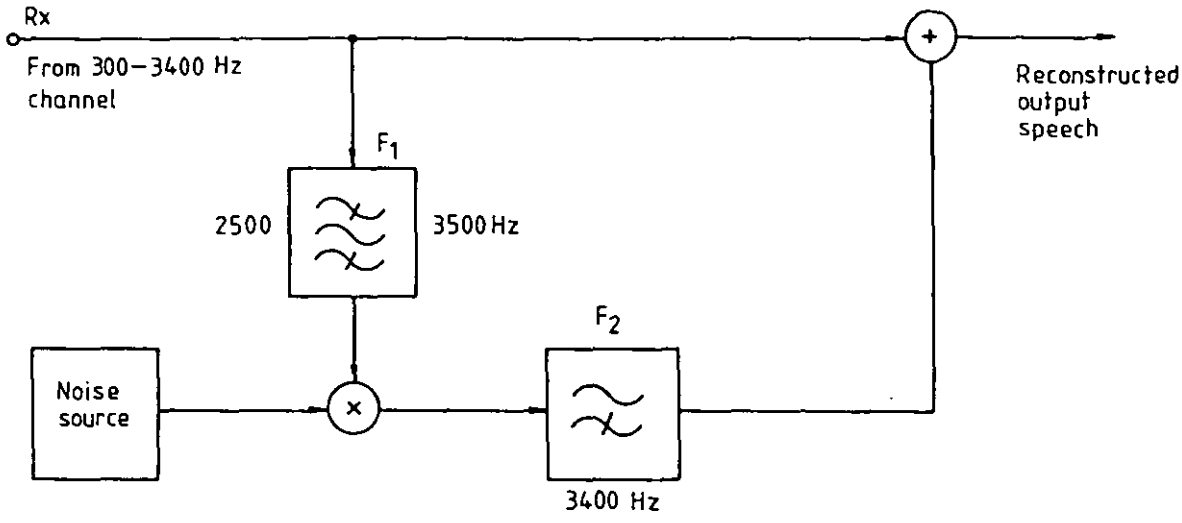


FIG. 5·21

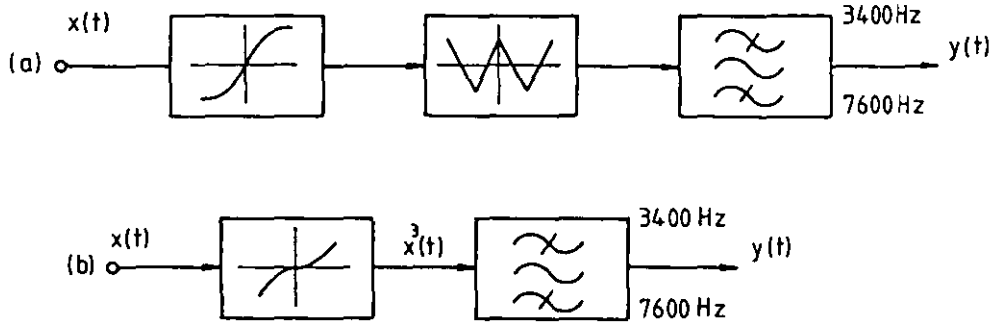
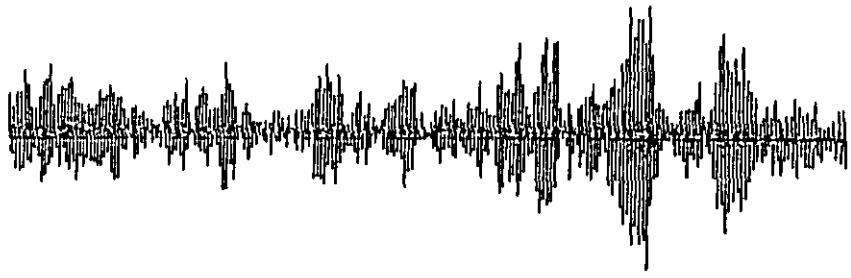


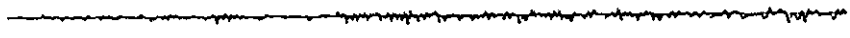
FIG. 5·22 HIGH FREQUENCY REGENERATION BY NON-LINEAR TRANSFER FUNCTIONS.
(a) LOGARITHMIC COMPRESSION AND 'W' TRANSFER FUNCTION
(b) CUBIC TRANSFER FUNCTION.
THE INPUT SIGNAL FOR BOTH CASES IS BAND LIMITED FROM 300 Hz TO 3400 Hz.

FIG. 5.23 Waveforms from the high frequency regenerator

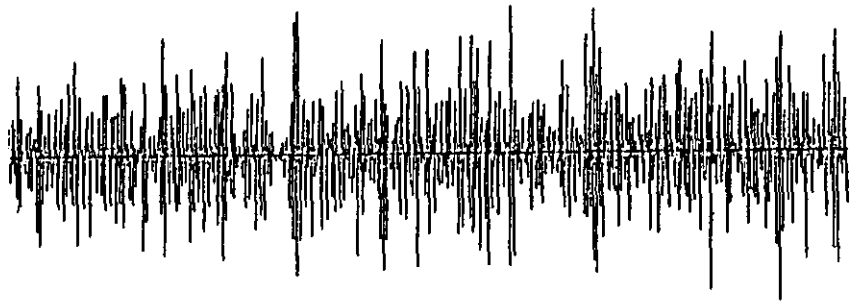
(a) original input signal /s/
300 - 7600 Hz



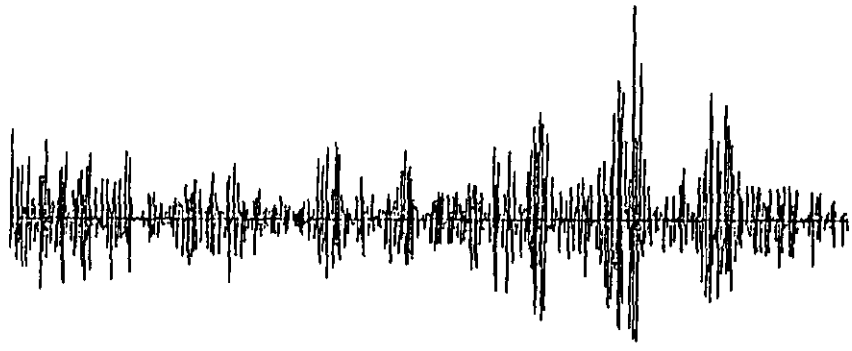
(b) BPF signal /s/
300 - 3400 Hz



(c) regeneration for /s/ by
noise source and BPF

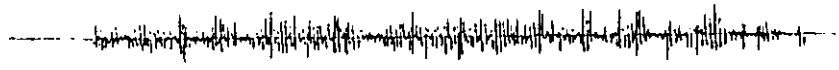


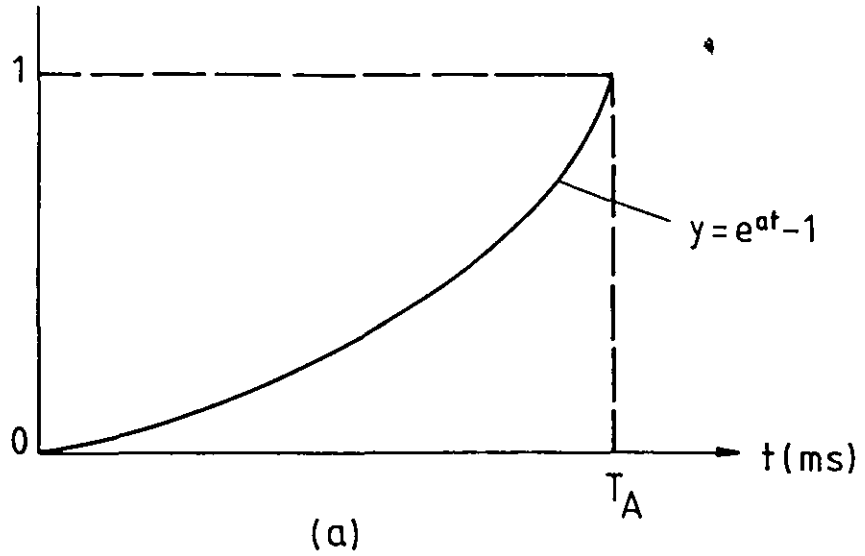
(e) output of 'W' - transfer
function and BPF, the
input signal is (b)



(d) envelope of (a) multiplied
by (c)

(f) output of cubic function
and BPF for /f/





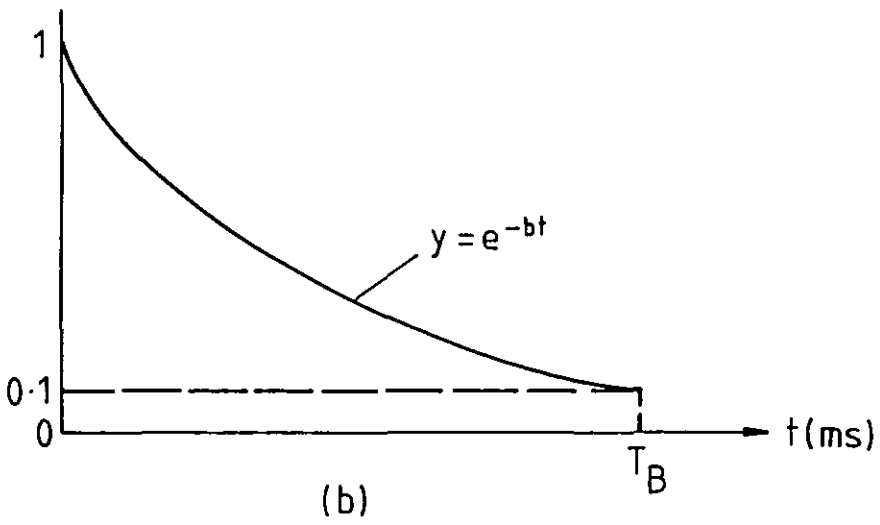


FIG. 5·24 EXPERIMENTAL WEIGHTING FUNCTIONS

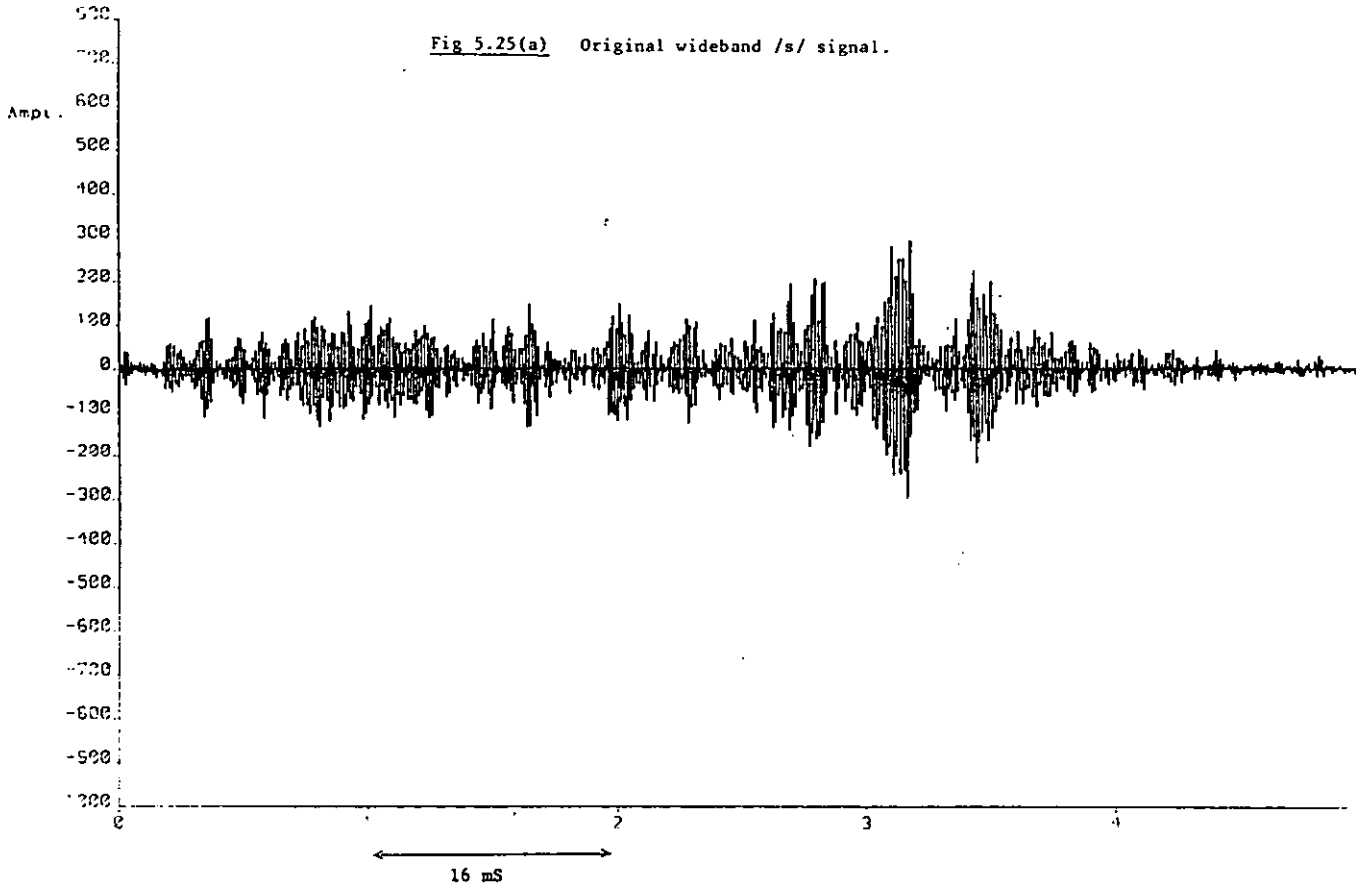


Fig 5.25(a) Original wideband /s/ signal.

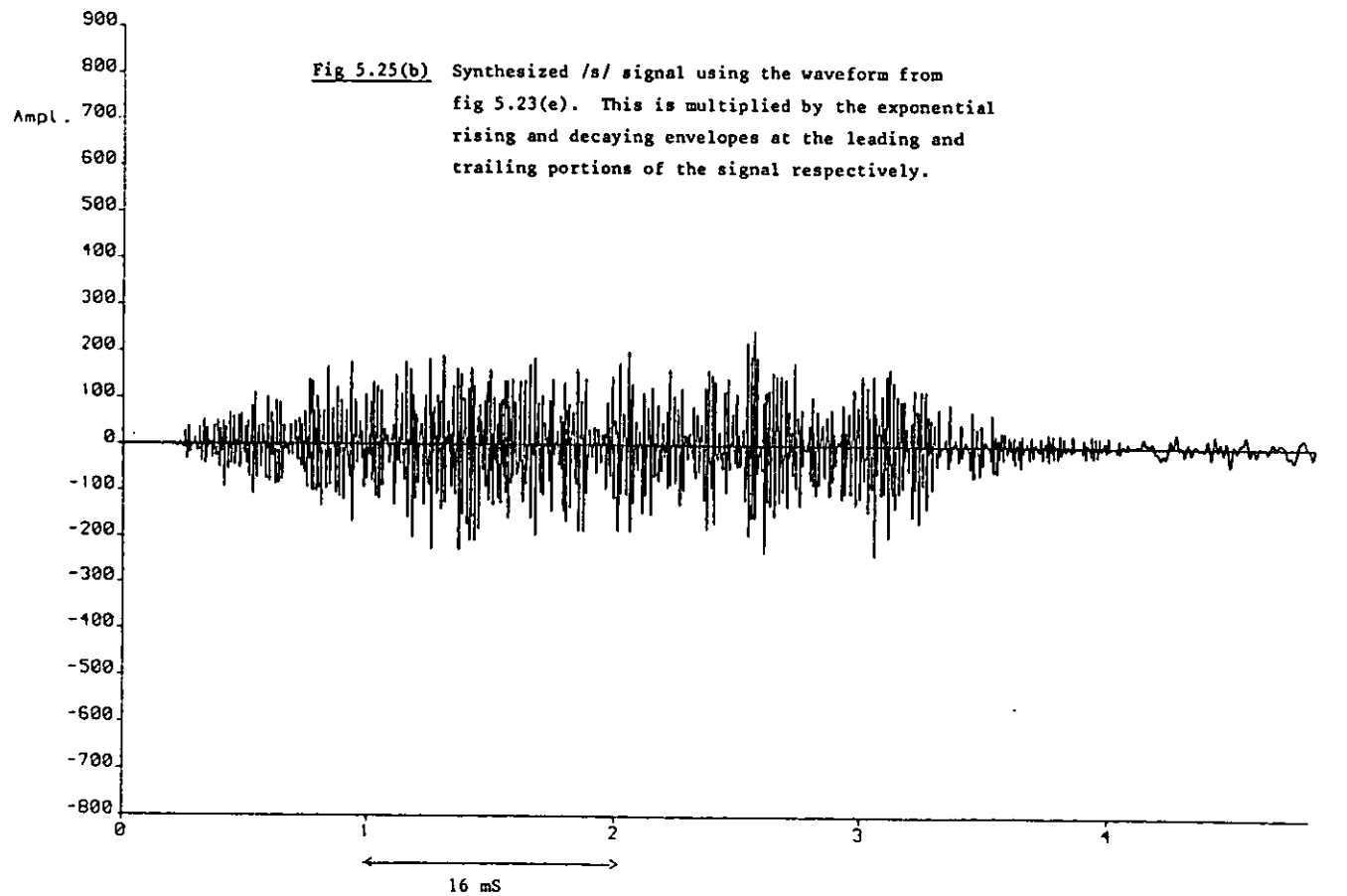
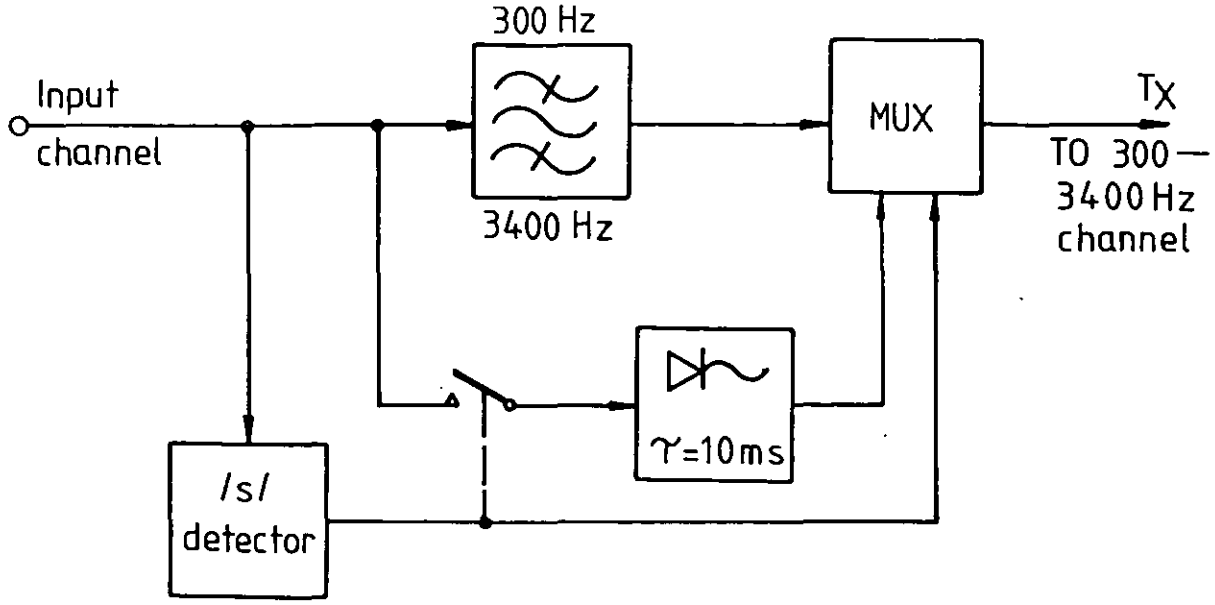
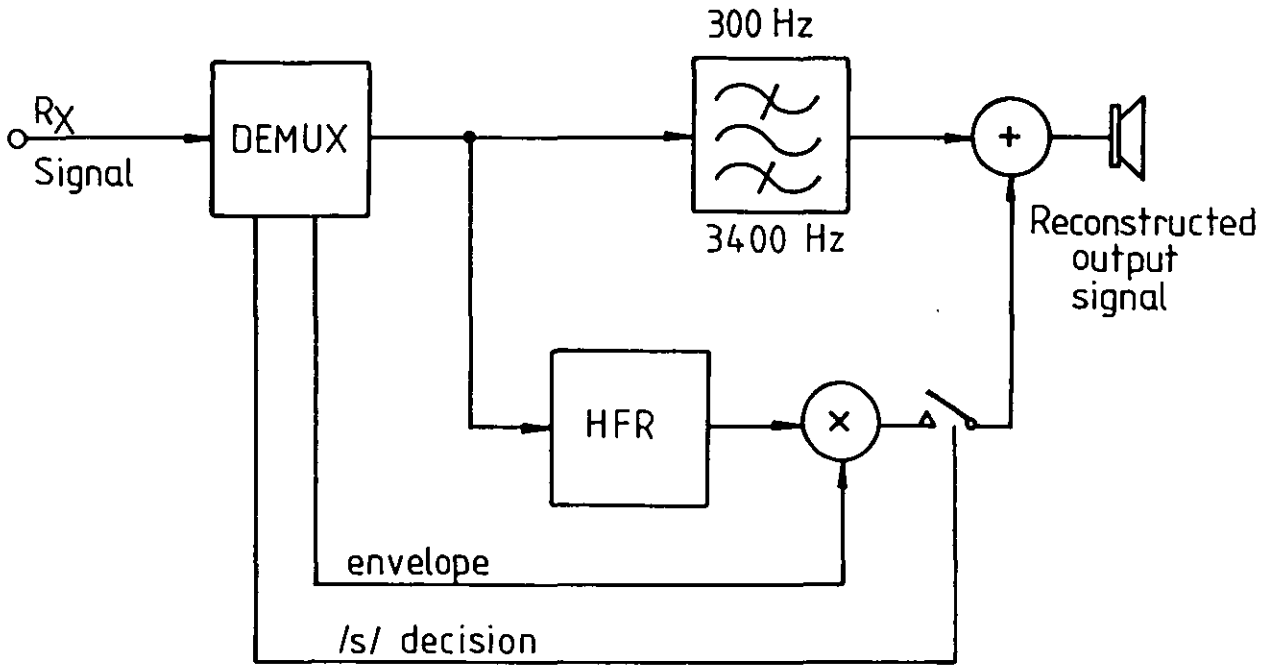


Fig 5.25(b) Synthesized /s/ signal using the waveform from fig 5.23(e). This is multiplied by the exponential rising and decaying envelopes at the leading and trailing portions of the signal respectively.



(a)



(b)

FIG. 5·26 HFR FOR /s/ SOUNDS INCORPORATING THE AMPLITUDE ENVELOPE WAVEFORM

Fig 5.27(a) Input signal to the network shown in fig 5.26.

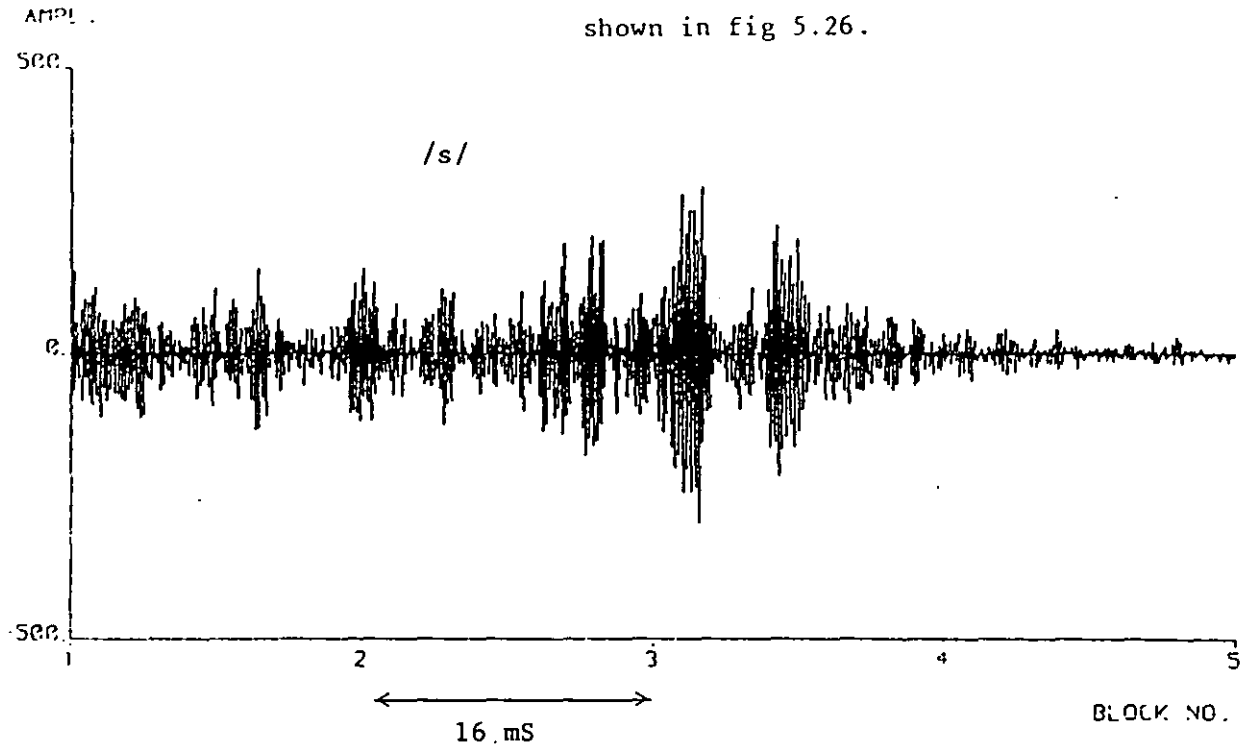
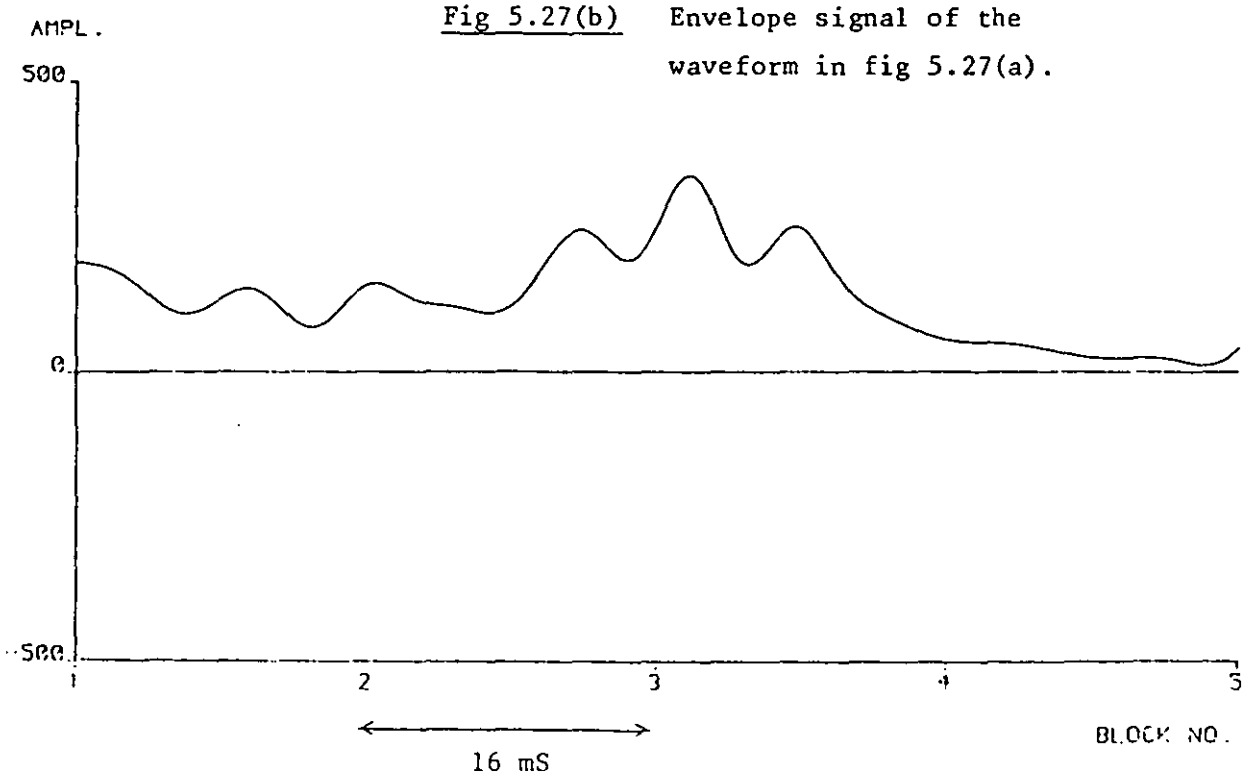


Fig 5.27(b) Envelope signal of the waveform in fig 5.27(a).



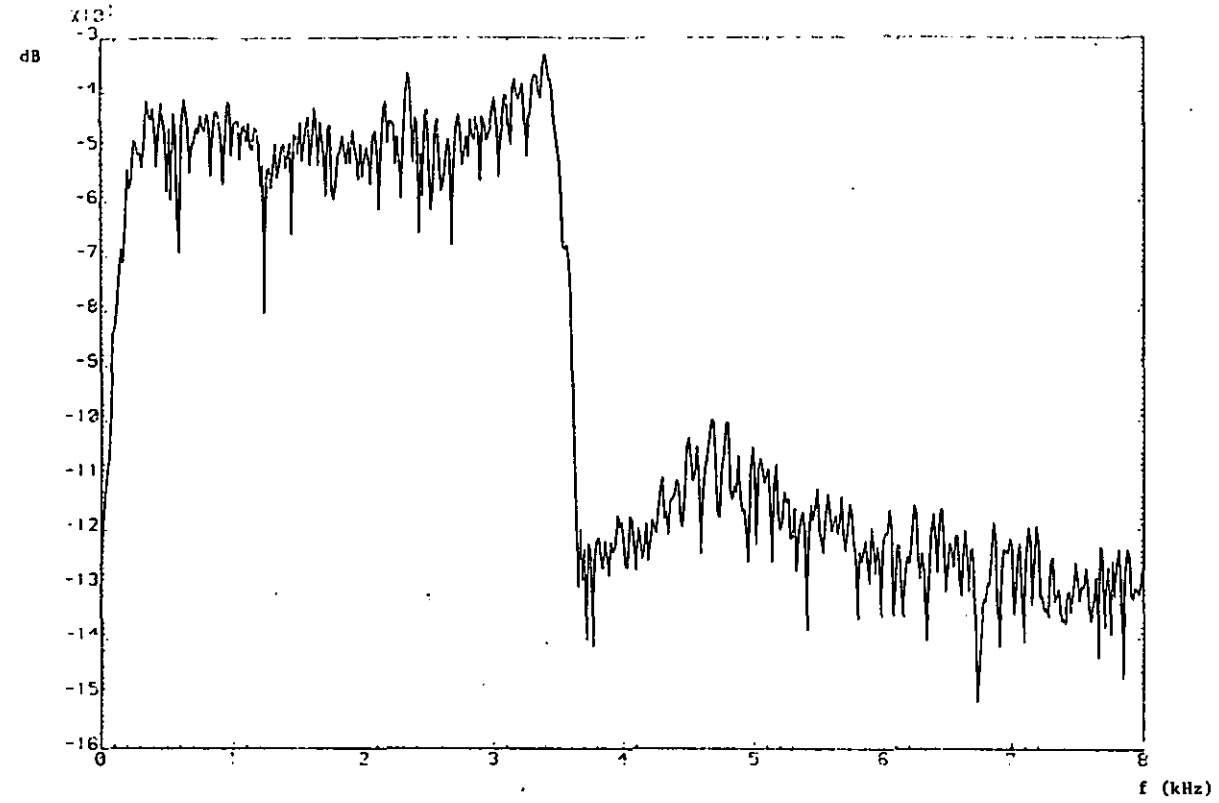


Fig 5.28(a) BPF Spectrum for an /s/ signal.

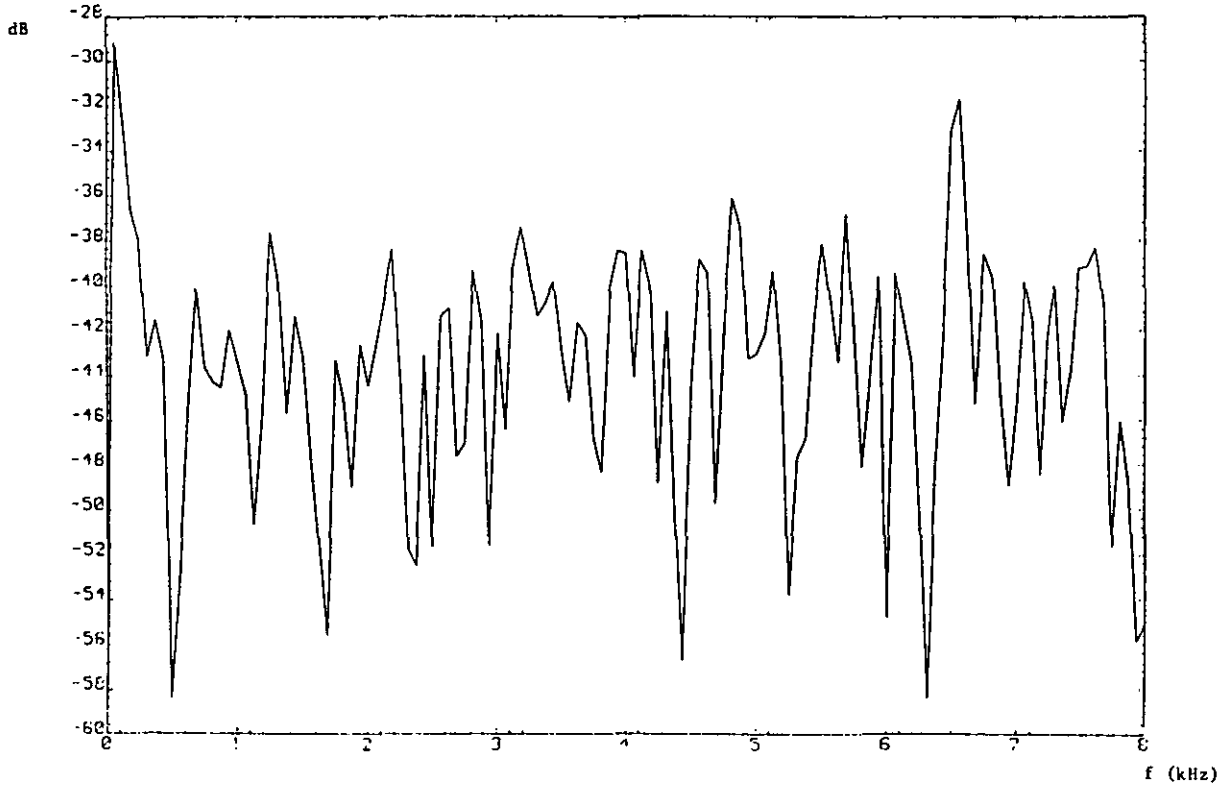


Fig 5.28(b) Expanded spectrum of an /s/ signal.

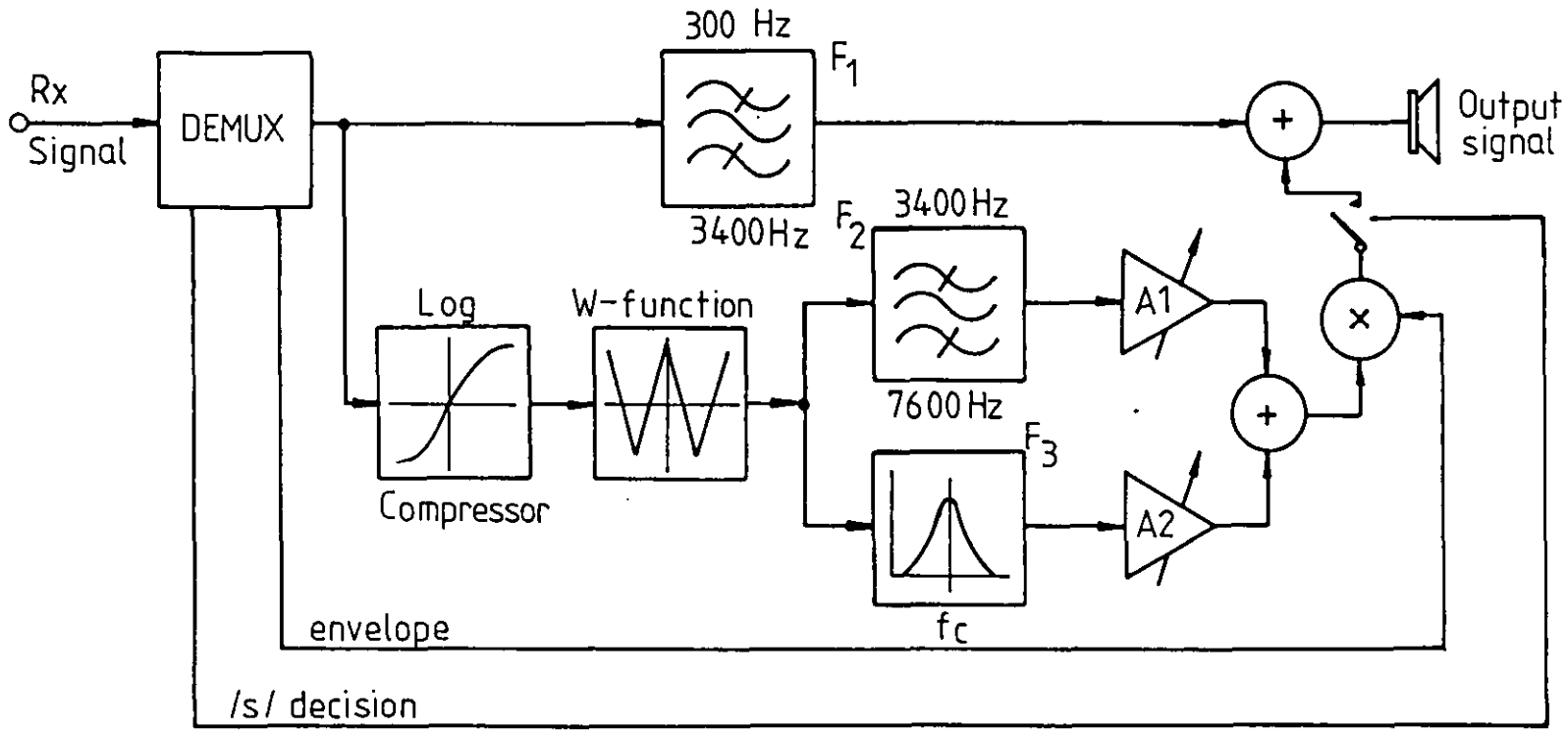

Rx
Signal
DEMUX
300 Hz
F1
3400Hz
3400Hz
F2
Log
W-function
Compressor
7600Hz
F3
fc
A1
A2
+
+
×
Output
signal
envelope
/s/ decision

FIG. 5·29

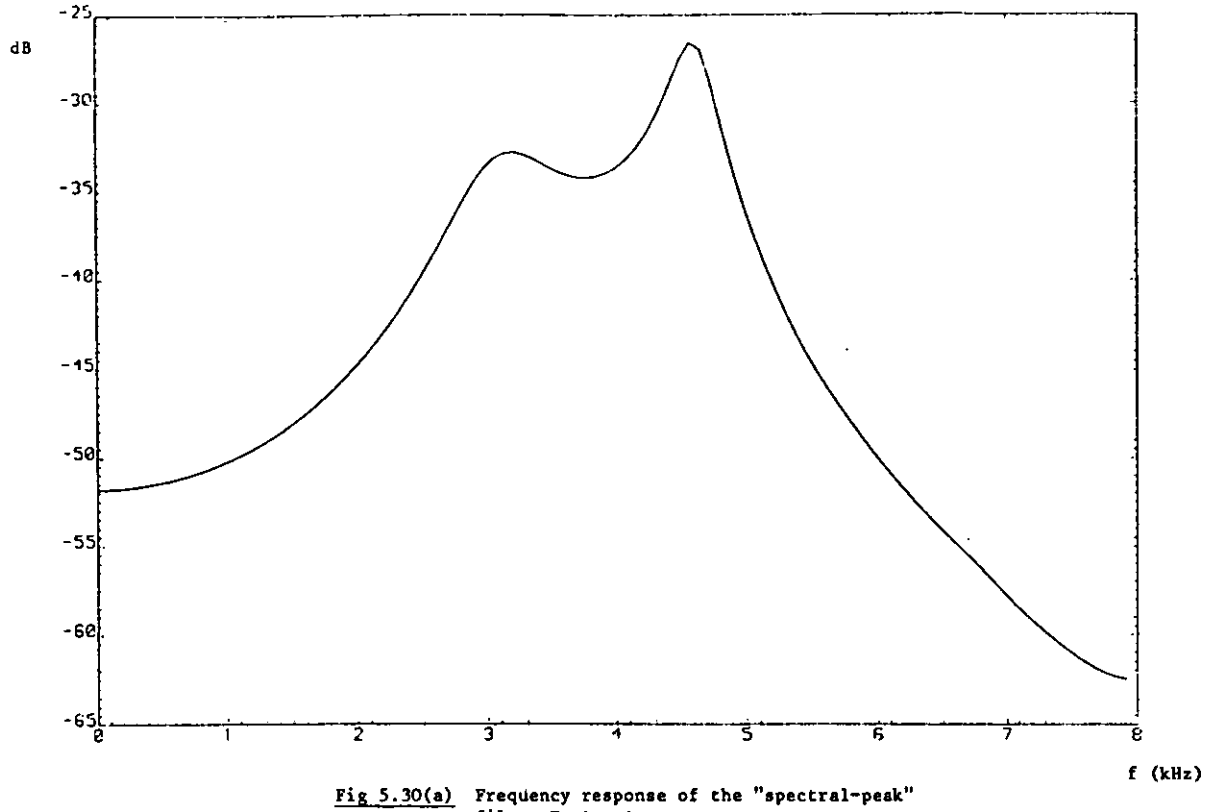


Fig 5.30(a) Frequency response of the "spectral-peak" filter F_3 in fig 5.29.

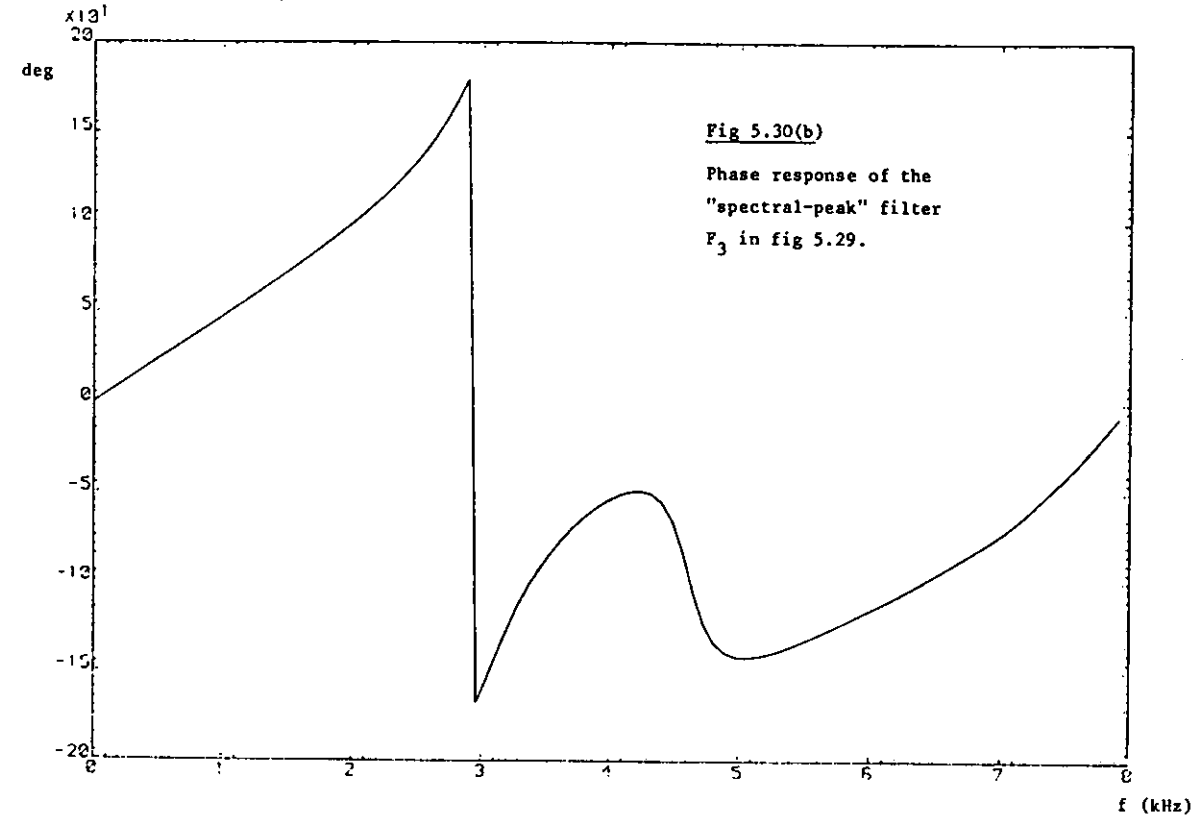


Fig 5.30(b)

Phase response of the "spectral-peak" filter F_3 in fig 5.29.

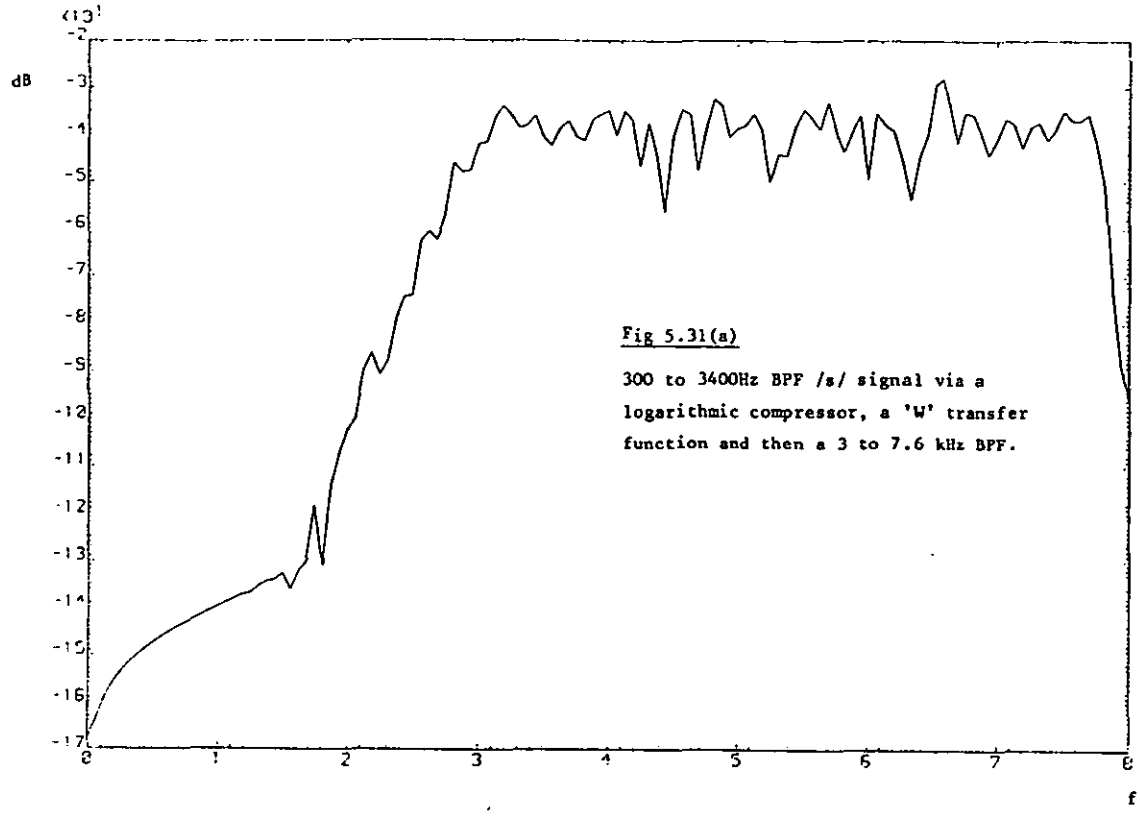


Fig 5.31(a)

300 to 3400Hz BPF /s/ signal via a logarithmic compressor, a 'W' transfer function and then a 3 to 7.6 kHz BPF.

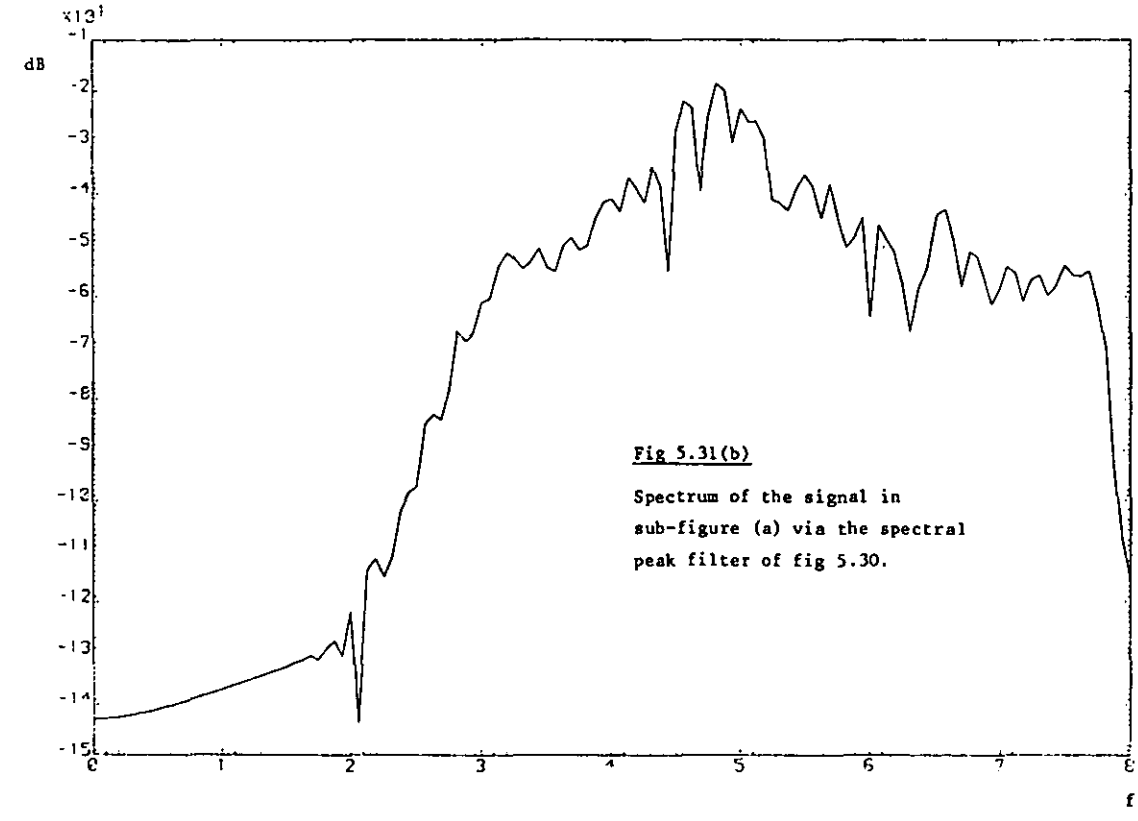


Fig 5.31(b)

Spectrum of the signal in sub-figure (a) via the spectral peak filter of fig 5.30.

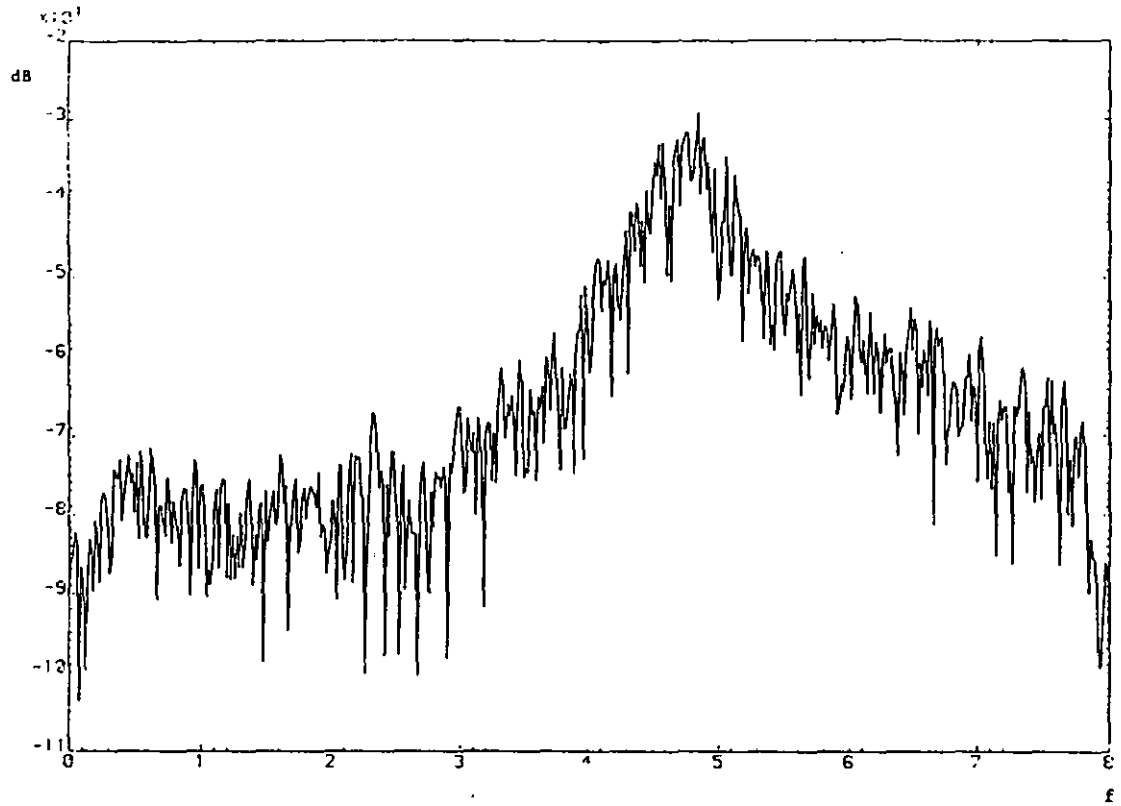


Fig 5.31(c) HFR Spectrum for an /s/ signal.

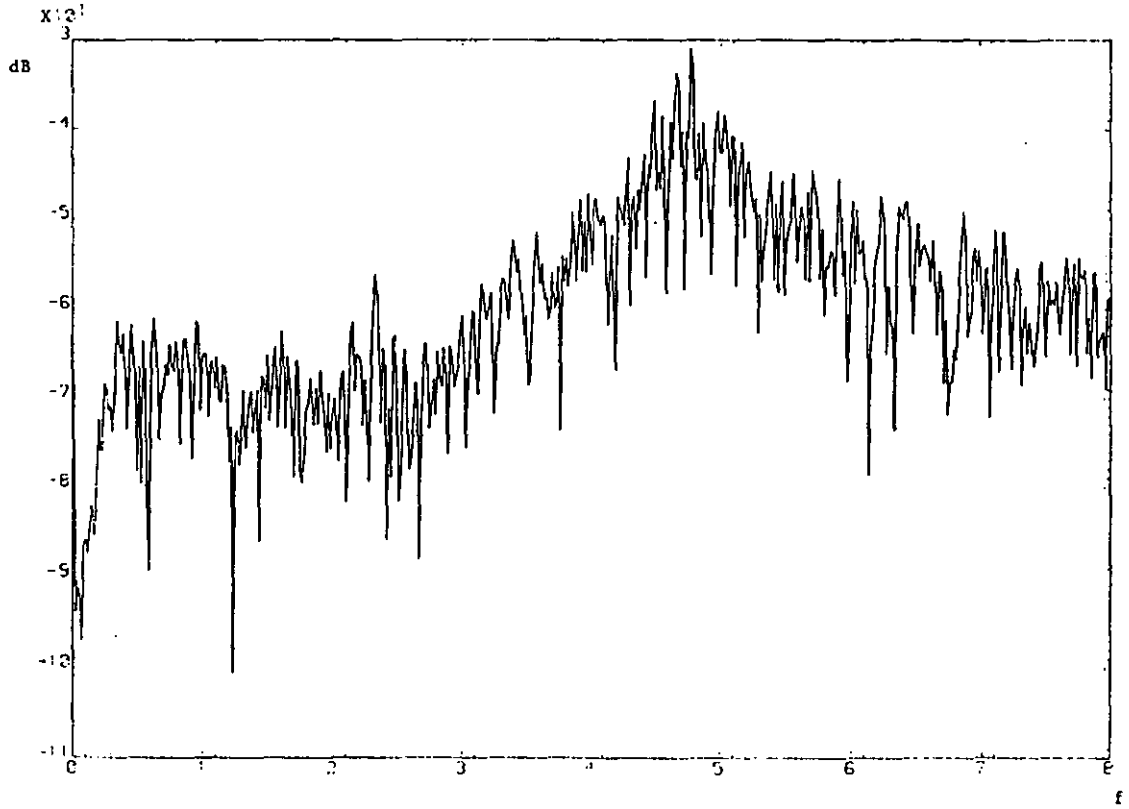


Fig 5.31(d) Original spectrum for an /s/ signal.

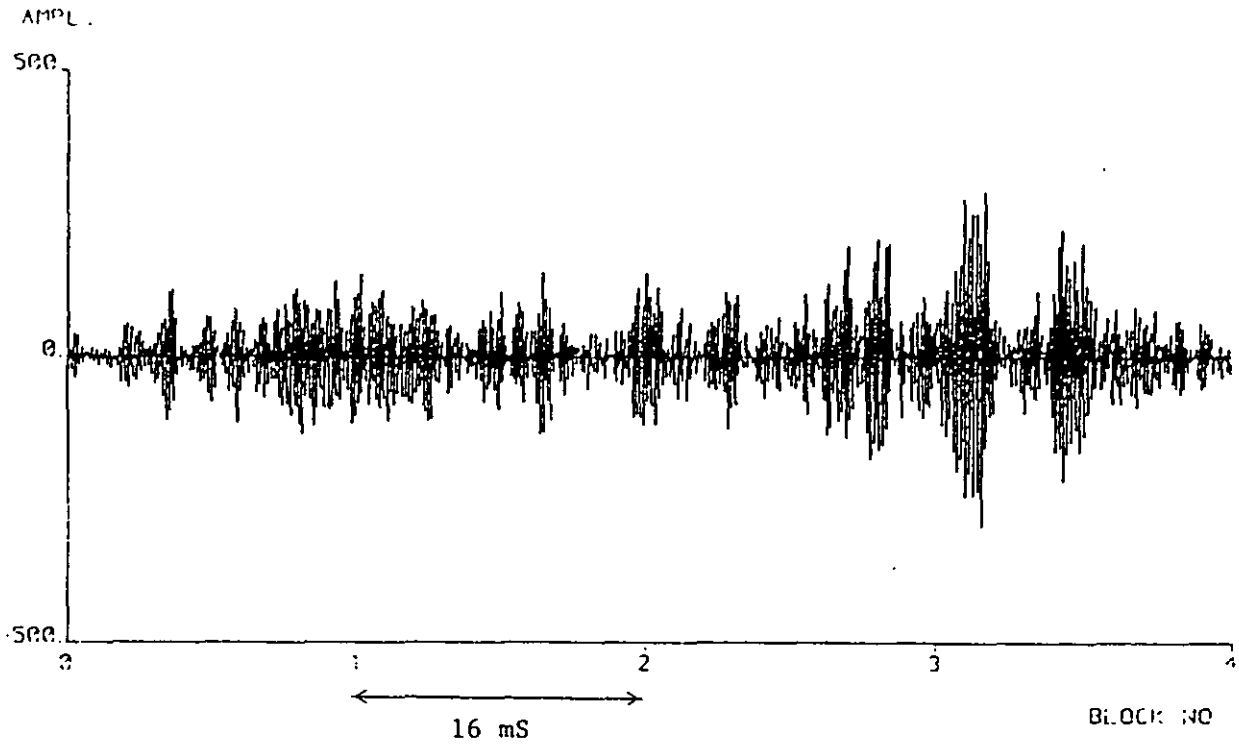


Fig 5.32(a) Original /s/ signal
(0.3 - 7.6 kHz)

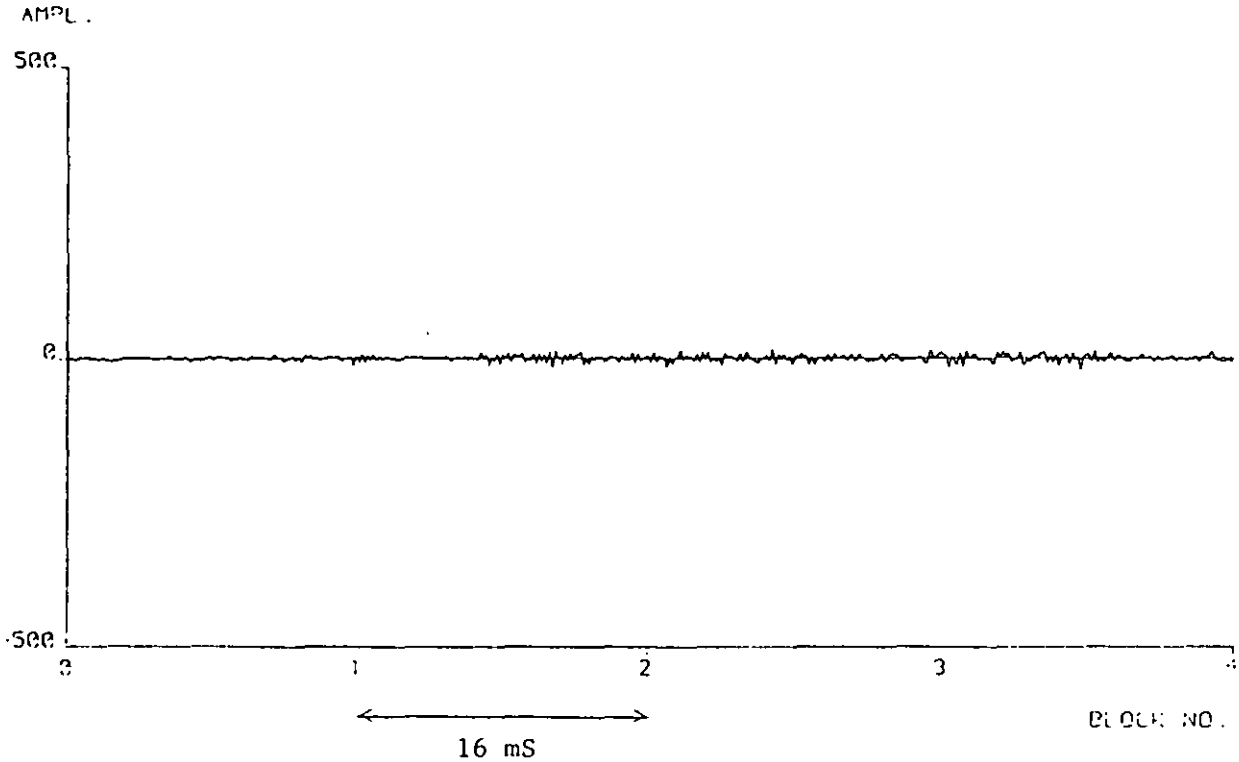


Fig 5.32(b) BPF /s/ signal
(0.3 - 3.4 kHz)

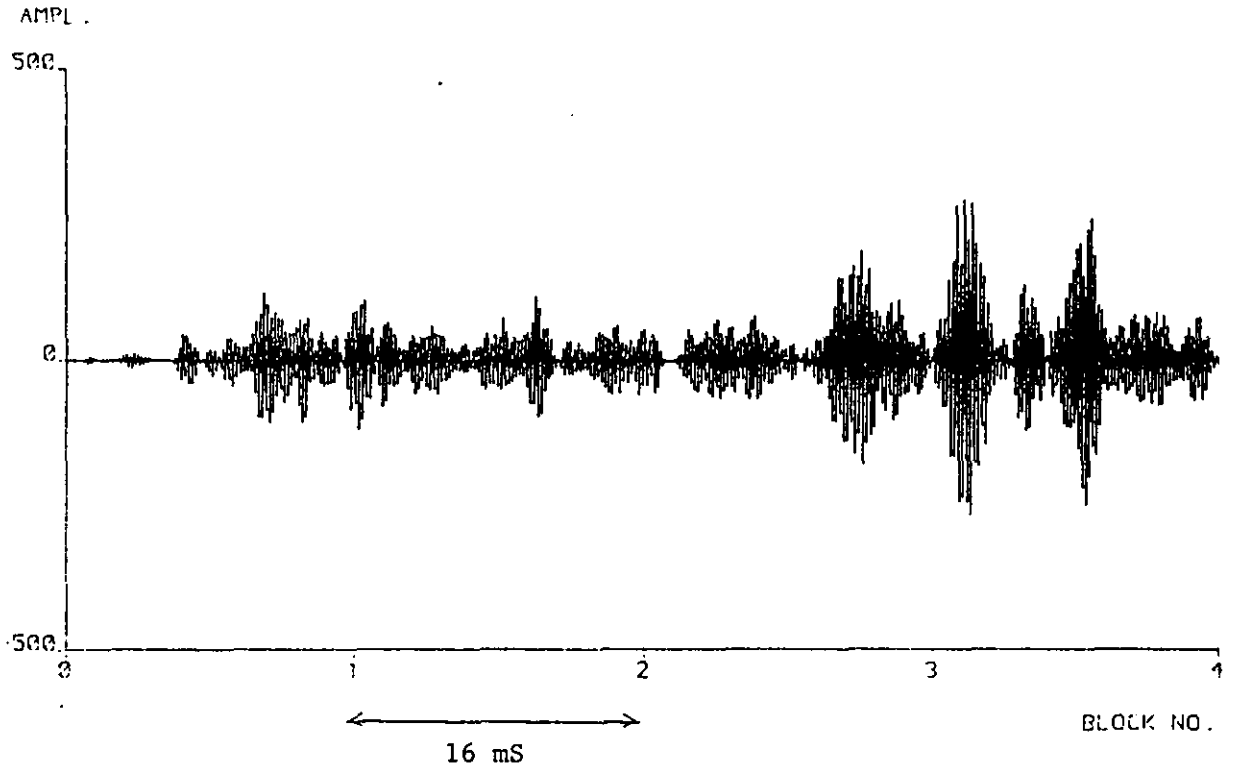


Fig 5.32(c) /s/ signal processed by h.f. band regeneration.

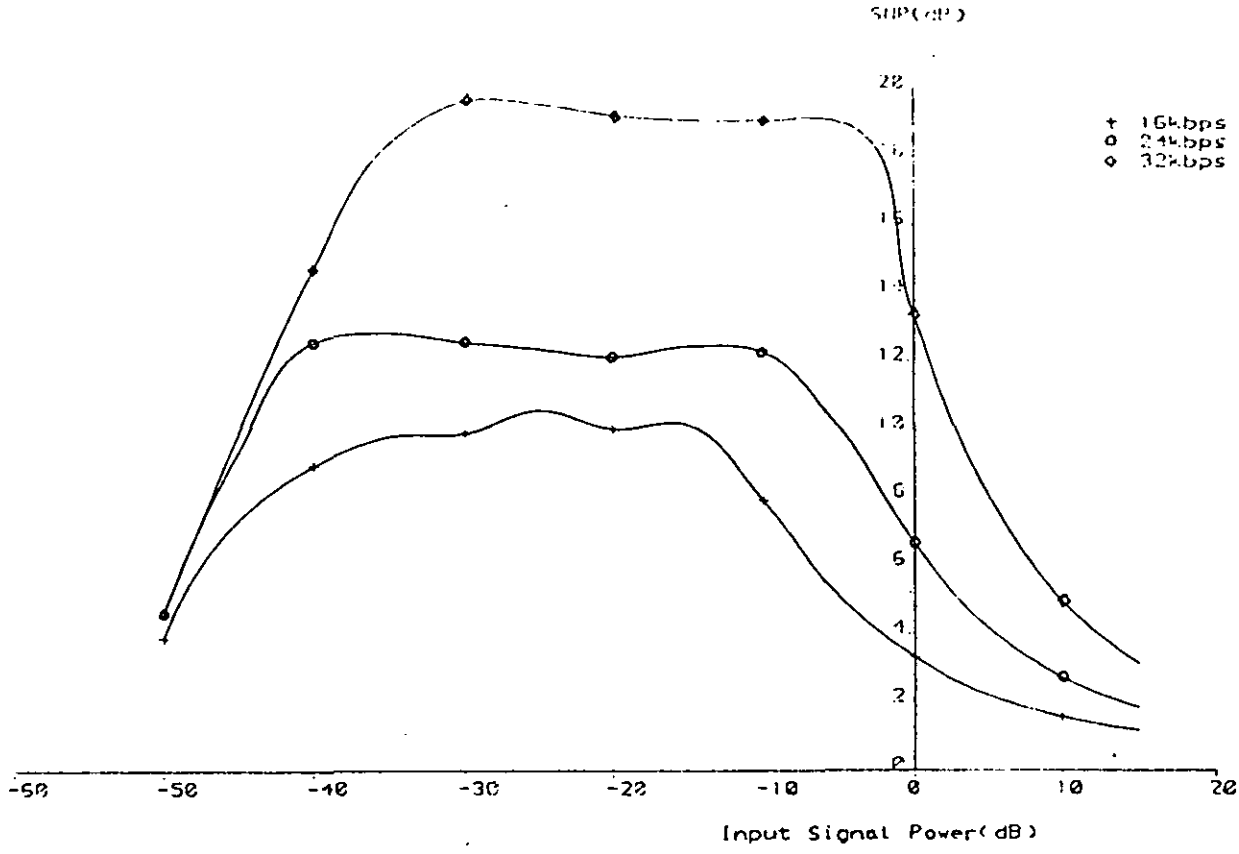


Fig 5.33 SNR performance of APCM for the narrow band speech signal at the transmission rates of 16,24 and 32 kb/s against input sig. power. (ref.78)

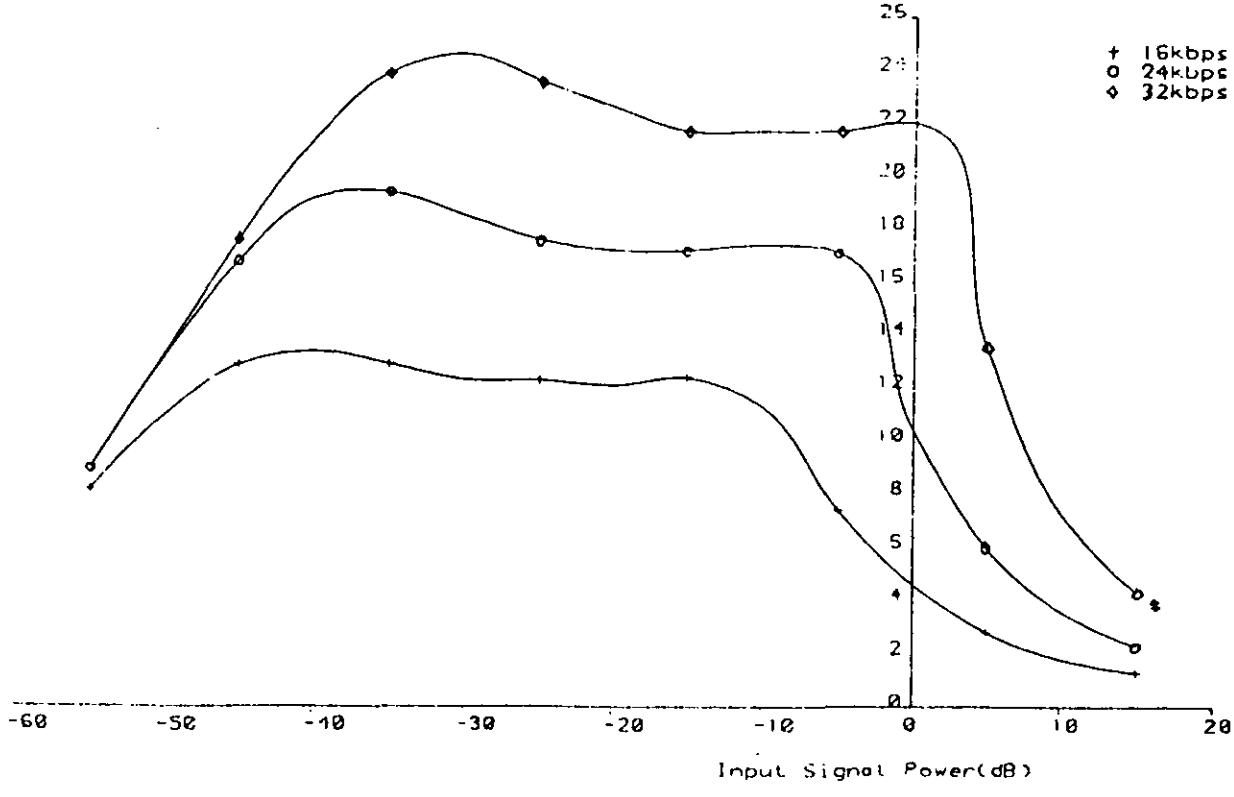


Fig 5.34 SNR performance of ADPCM for the narrow band speech signal at the transmission rates of 16,24 and 32 kb/s against input sig. power. (ref.78)

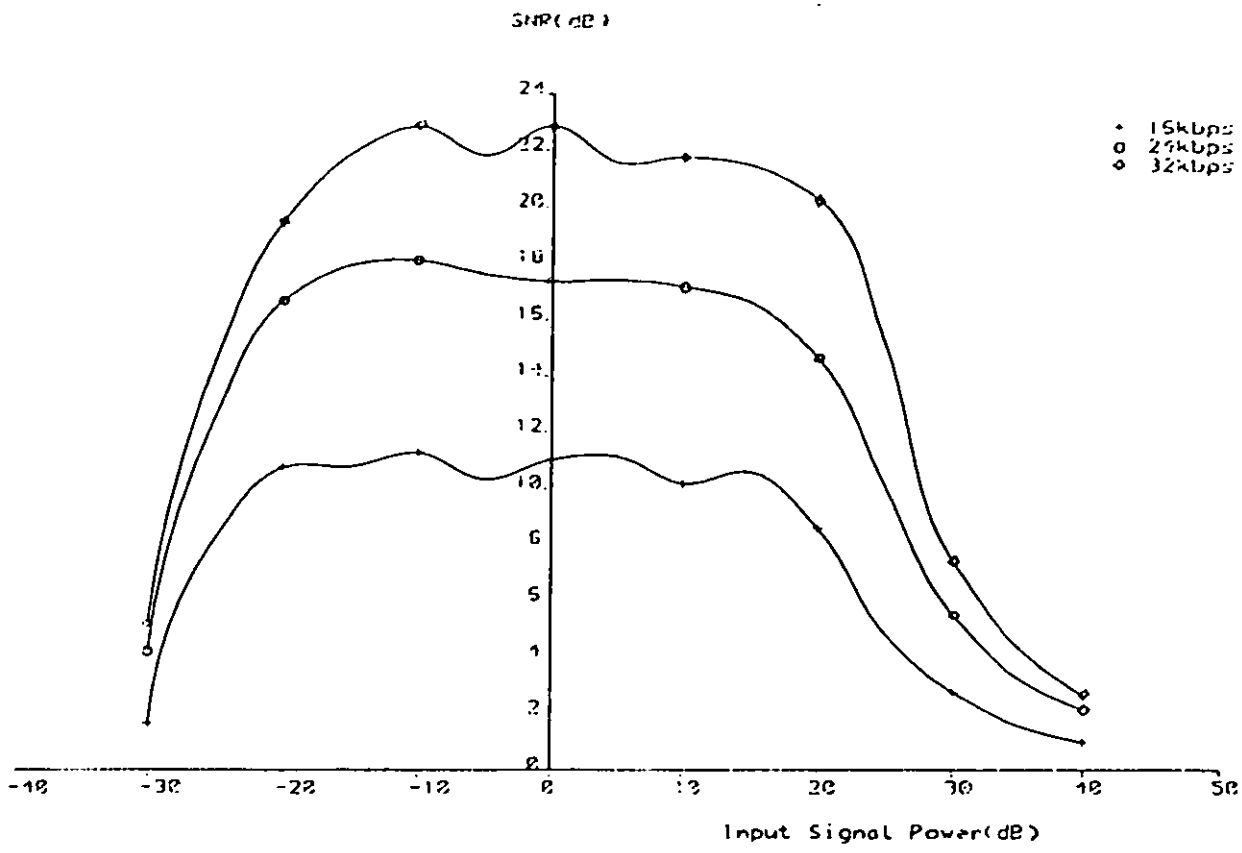


Fig 5.35 SNR performance of CFDM for the narrow band speech signal at the transmission rates of 16,24 and 32 kb/s against input sig. power.

(ref.78)

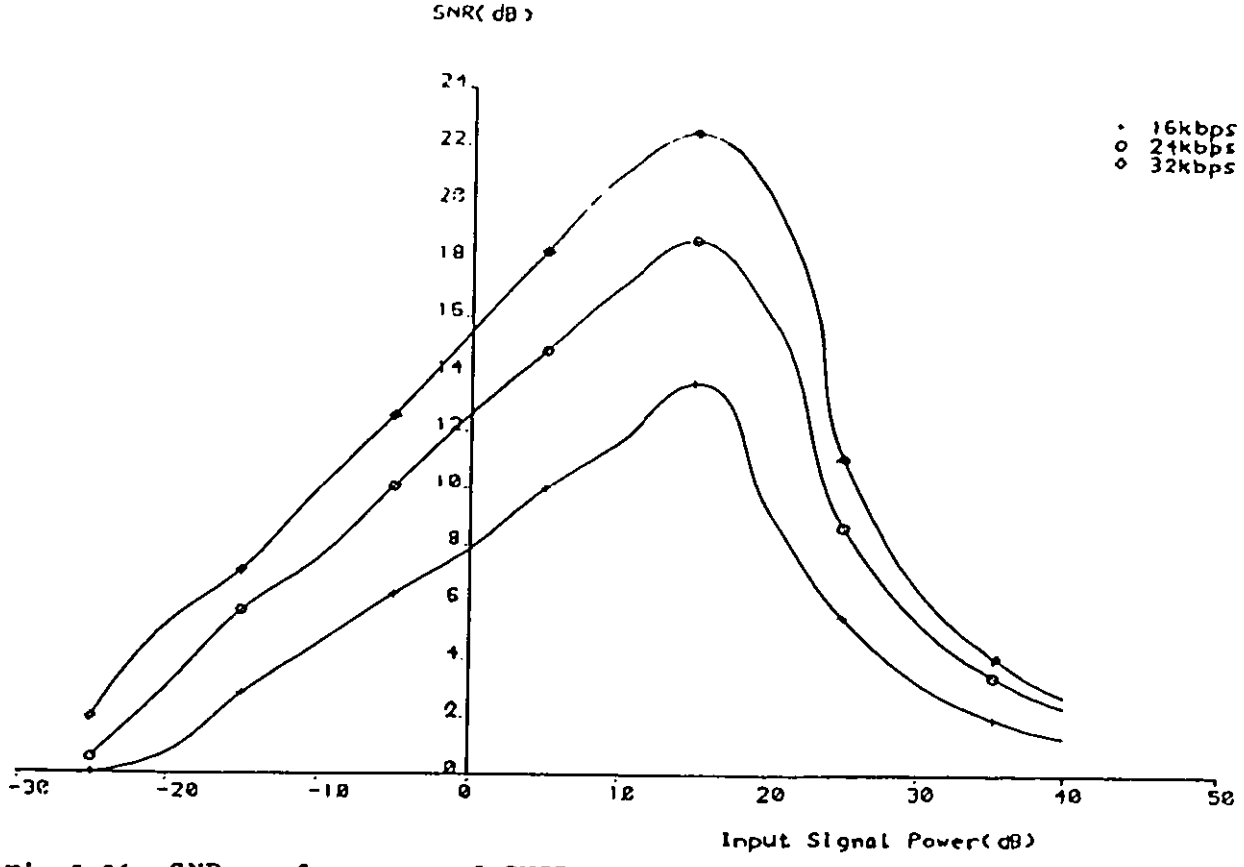


Fig 5.36 SNR performance of CVSD for the narrow band speech signal at the transmission rates of 16,24 and 32 kb/s against input sig. power.

(ref.78)

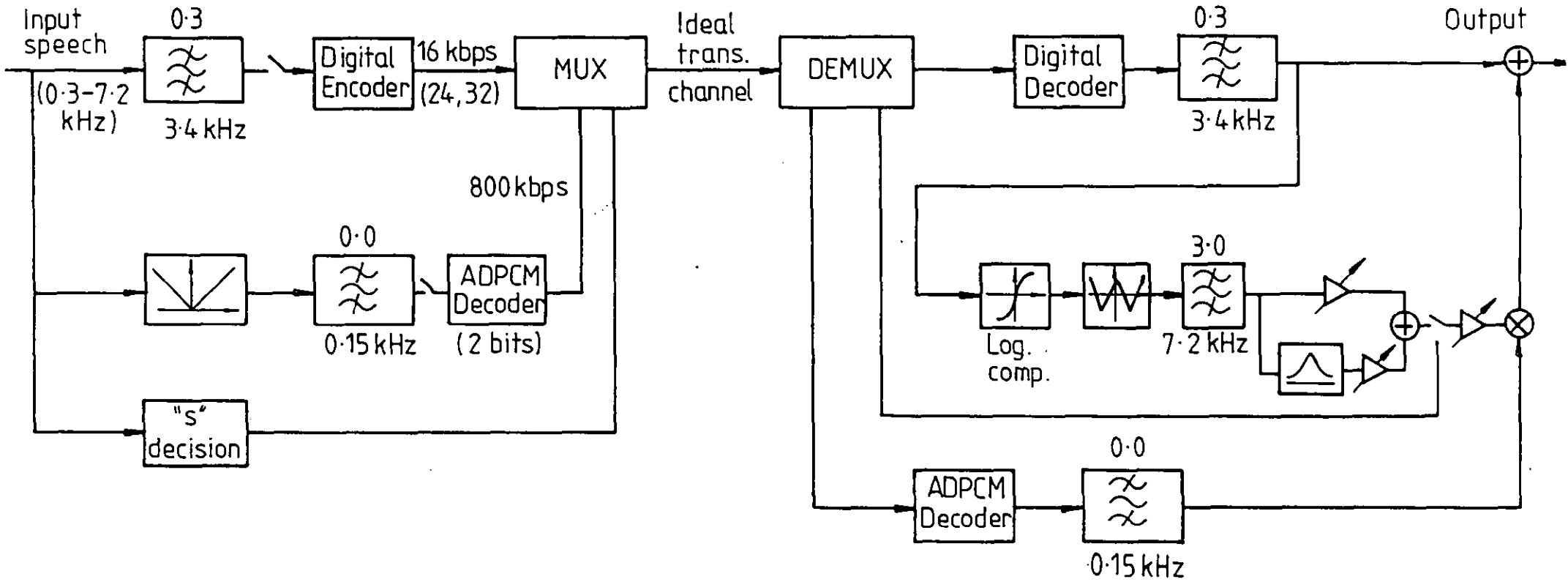


FIG. 5·37 SYSTEM BLOCK DIAGRAM OF SPEECH ENHANCEMENT USING DIGITAL CODECS (REF 78)

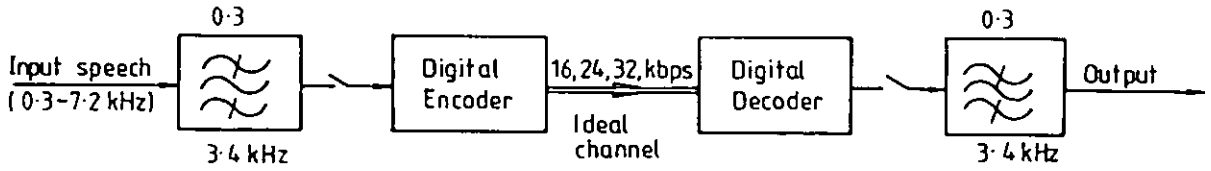


FIG. 5·38 DIGITAL SPEECH TRANSMISSION SYSTEM WITHOUT QUALITY ENHANCEMENT (Ref. 78)

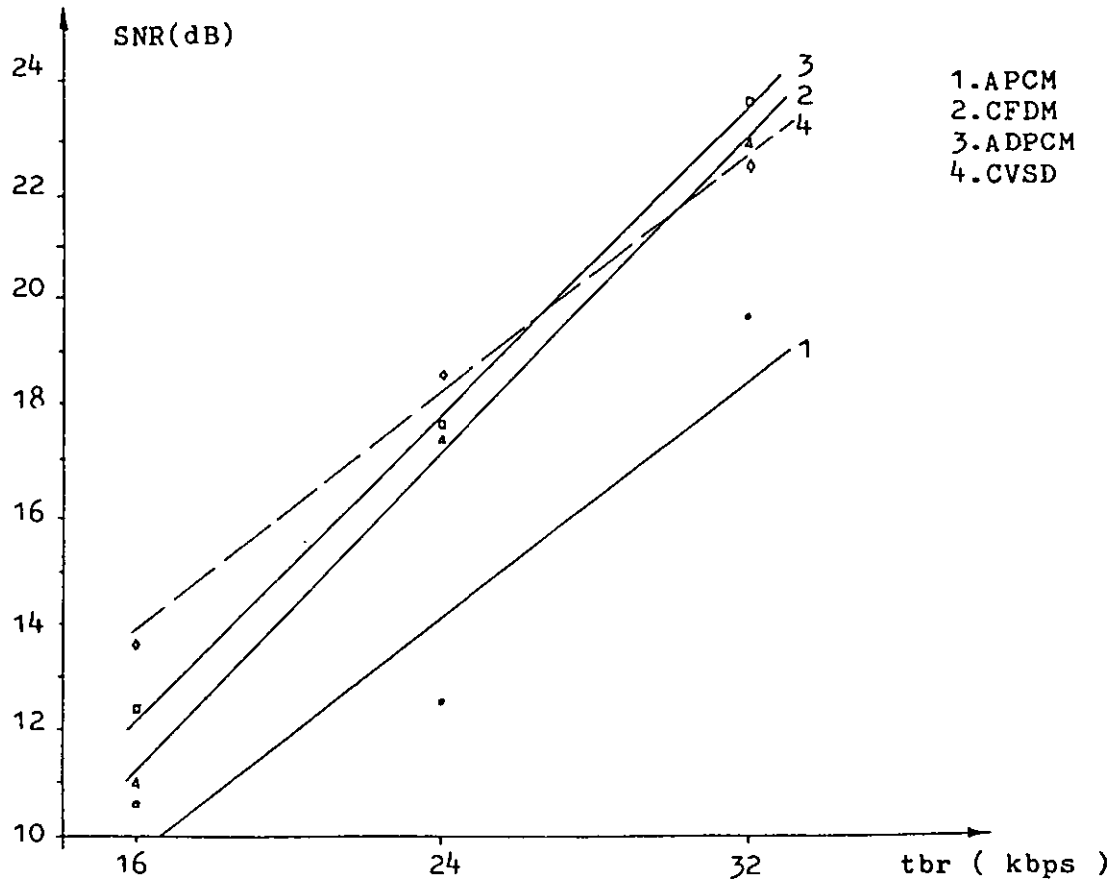


Fig 5.39 SNR performance against transmission bit rate. (tbr)
(ref.78)

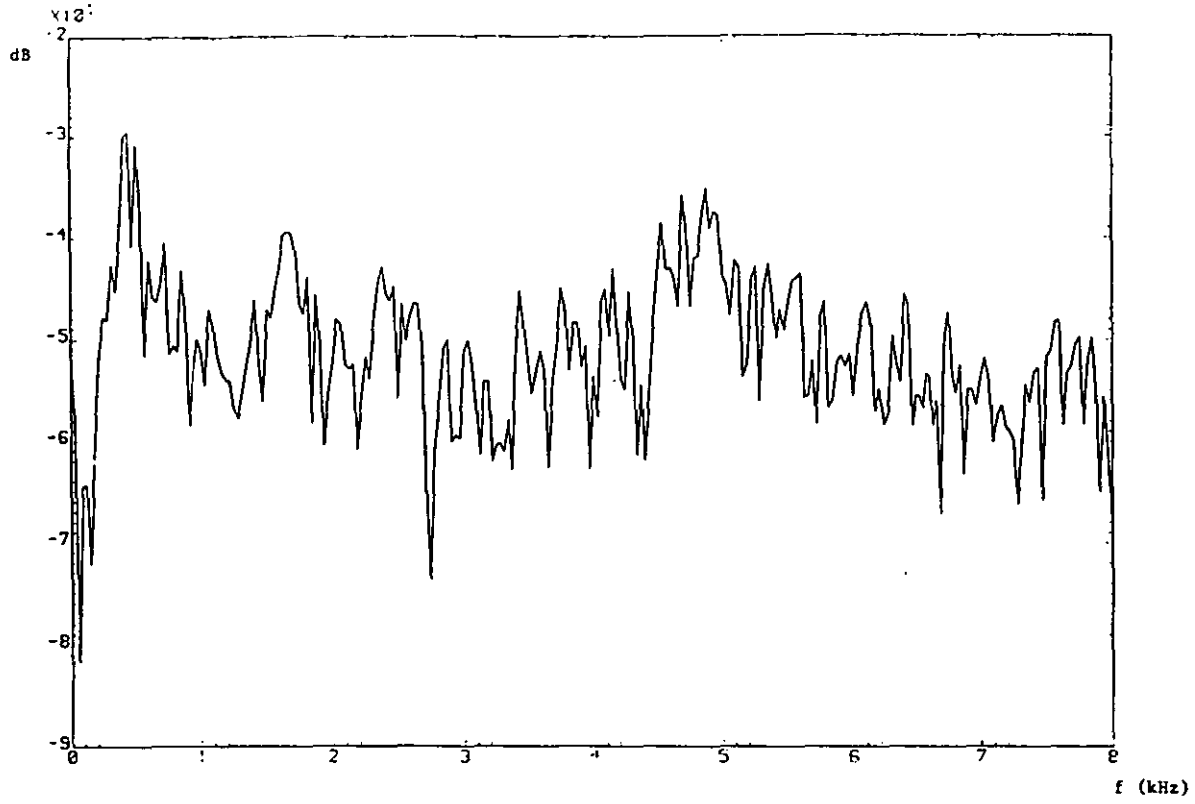


Fig 5.40(a) Spectrum of the leading portion of the /s/ signal from "fist ".

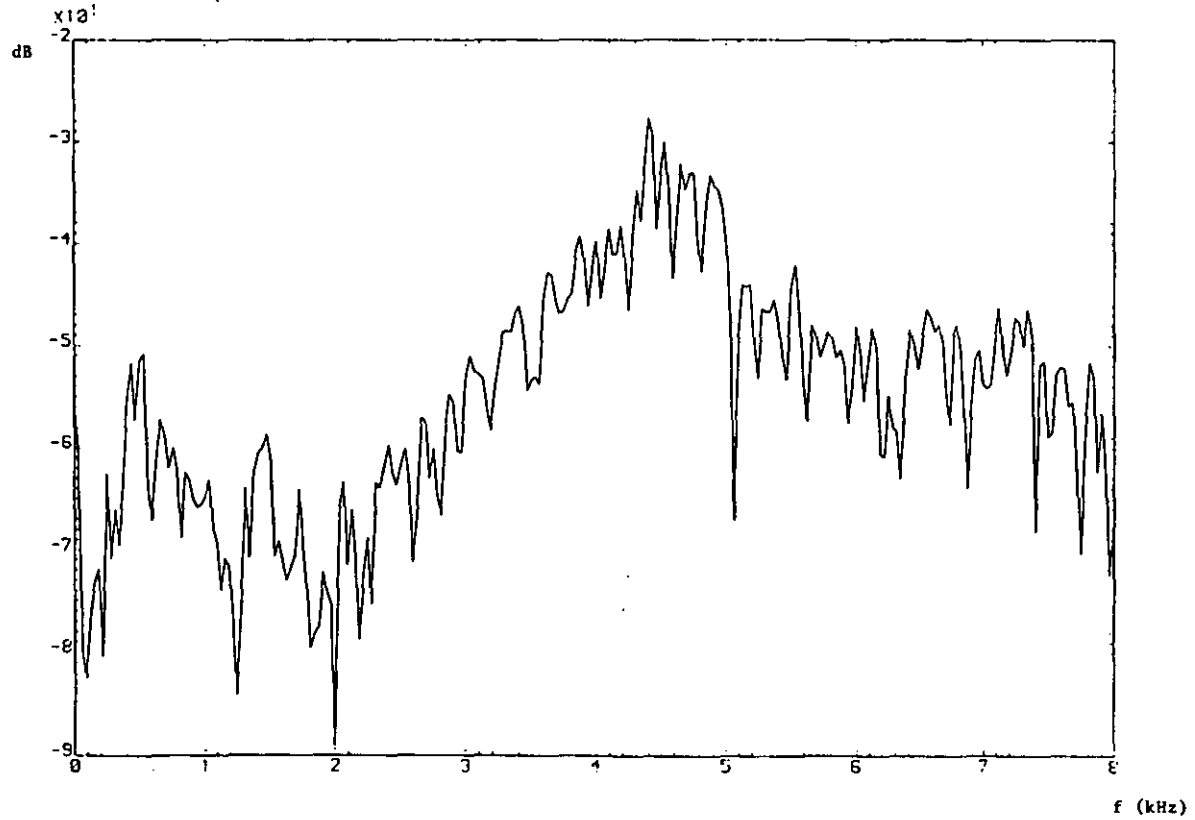


Fig 5.40(b) Spectrum of the trailing portion of the /s/ signal in "fist ".

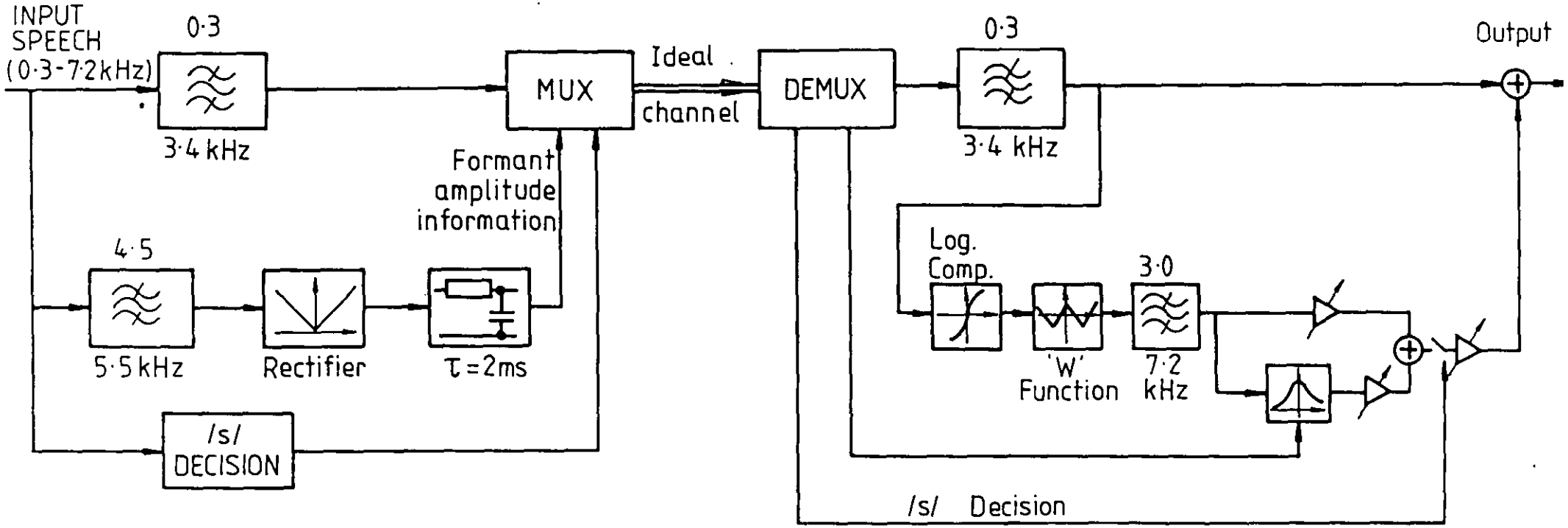


FIG. 5·41 SUGGESTED SPEECH TRANSMISSION SYSTEM WITH HIGH FREQUENCY ENHANCEMENT FOR FUTURE INVESTIGATION (REF 78)

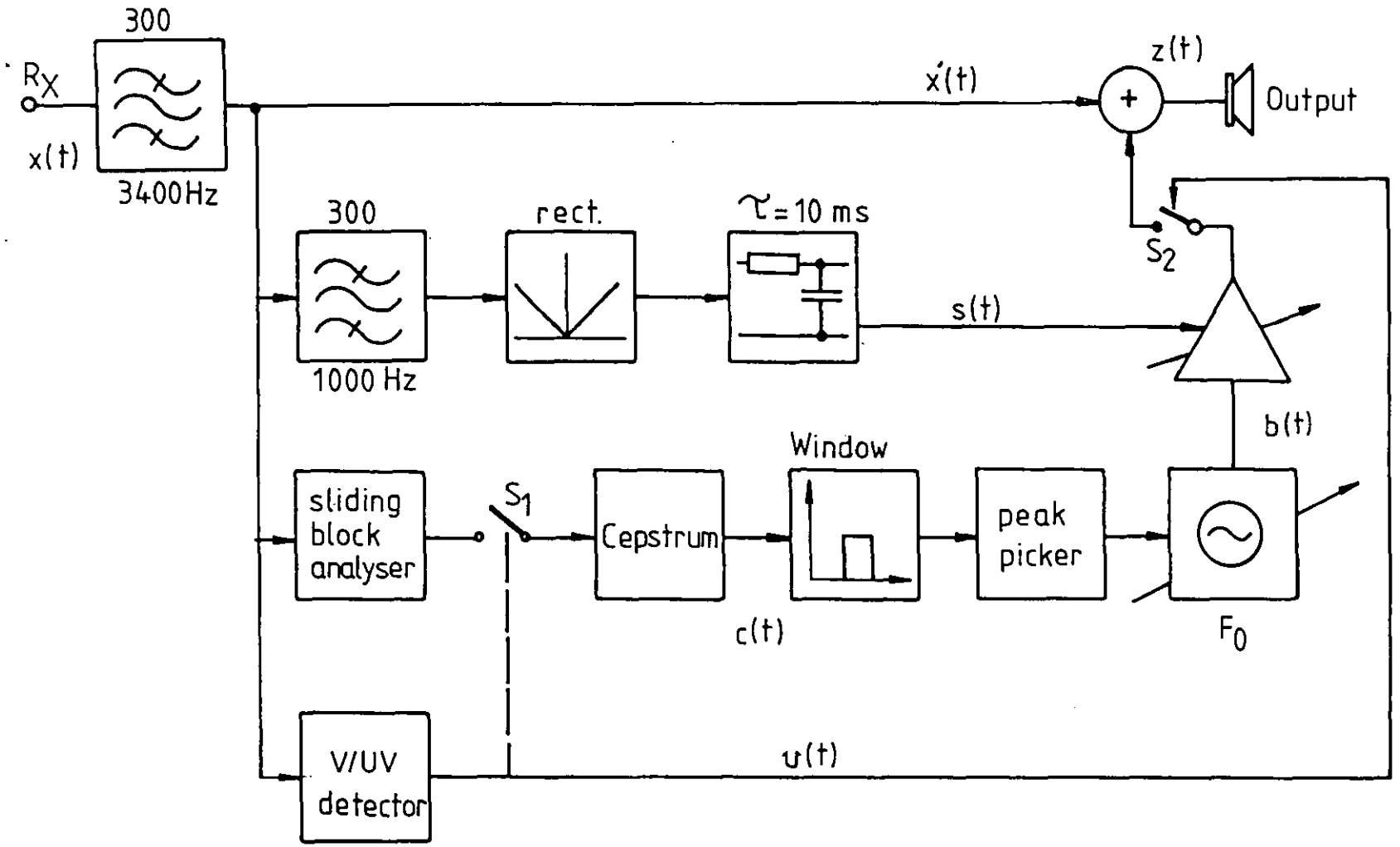


FIG. 5·42 BASEBAND SYNTHESIS

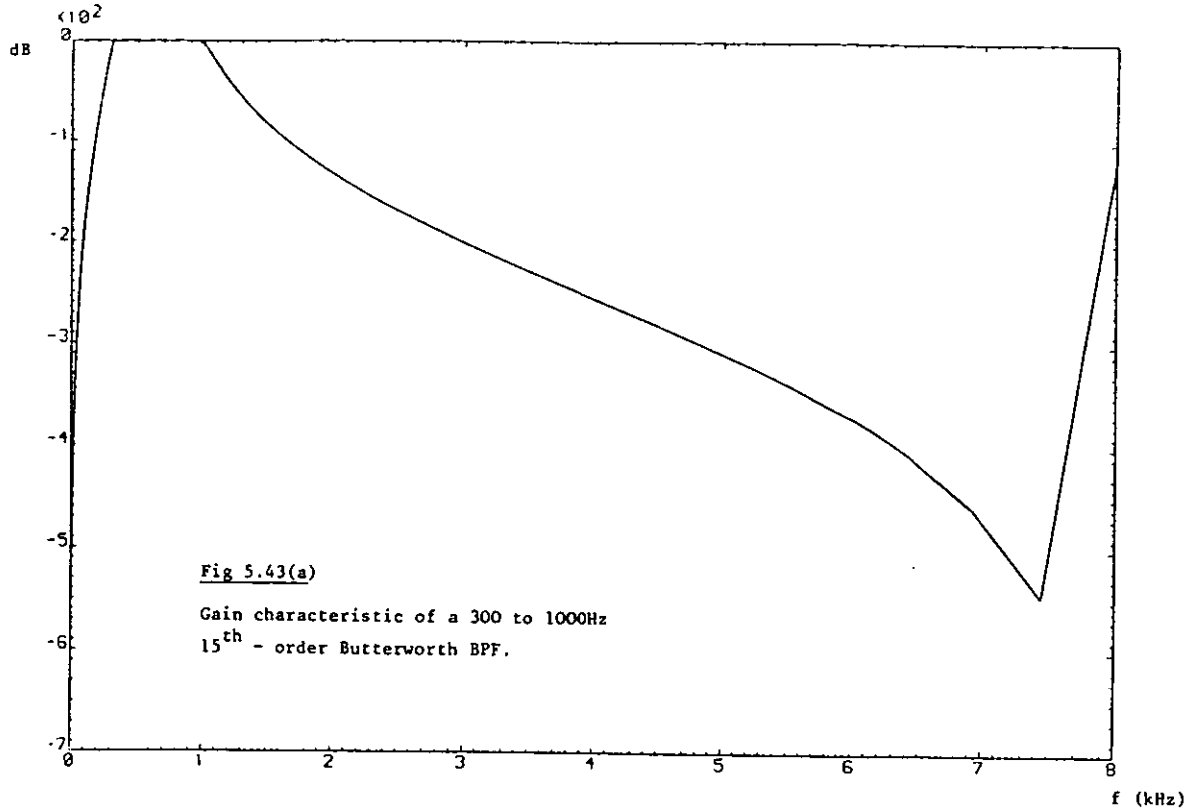


Fig 5.43(a)

Gain characteristic of a 300 to 1000Hz 15^{th} - order Butterworth BPF.

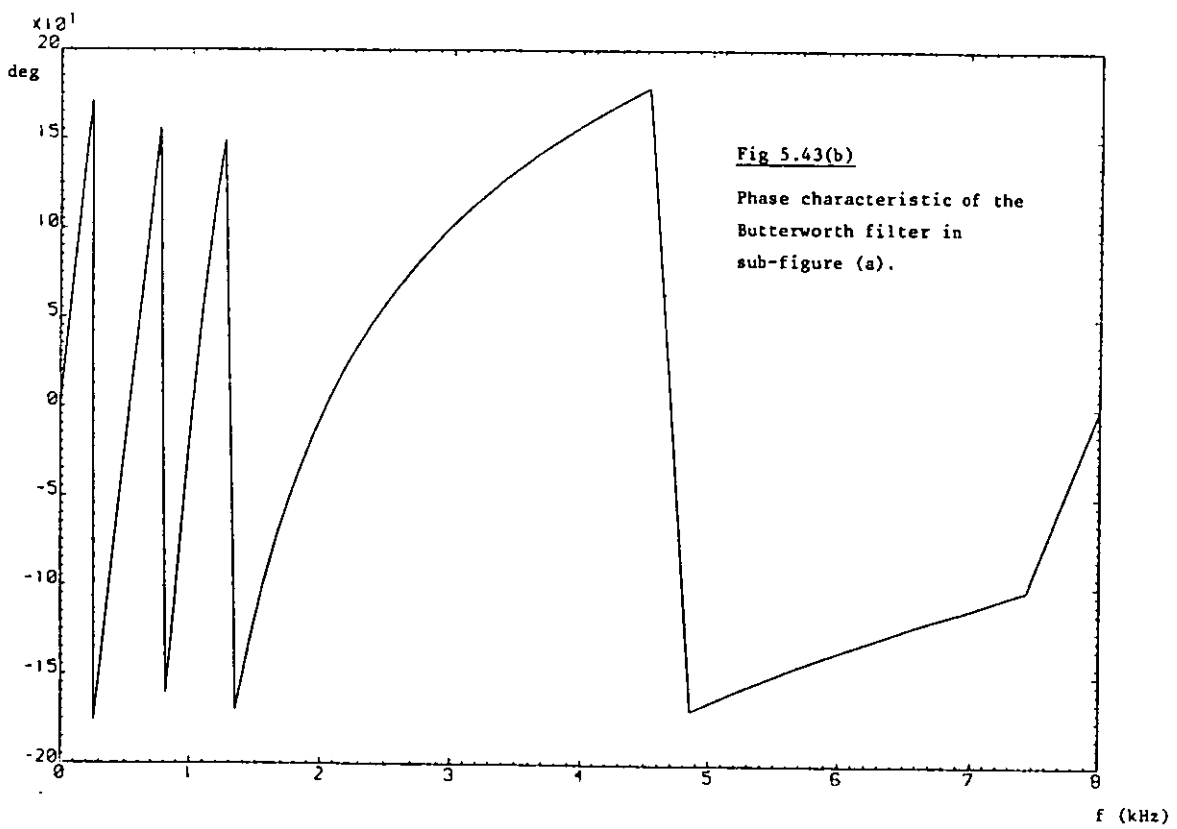


Fig 5.43(b)

Phase characteristic of the Butterworth filter in sub-figure (a).

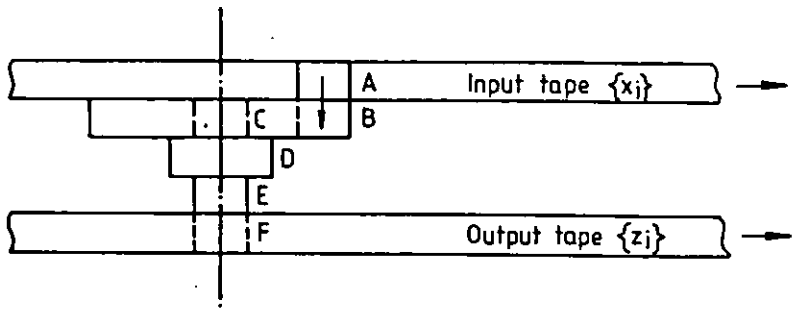


FIG. 5·44 SLIDING BLOCK REGIME

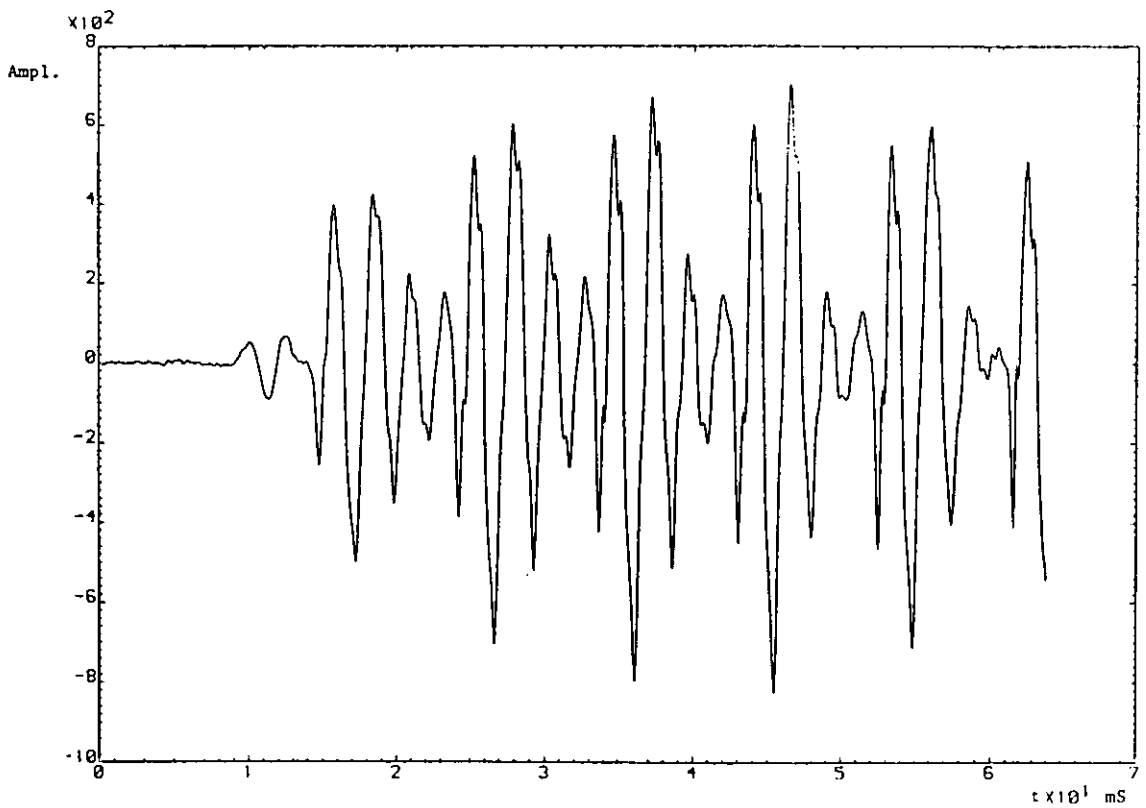


Fig 5.45(a) Input waveform to cepstrum analyser corresponding to block 'B' in fig 5.44.

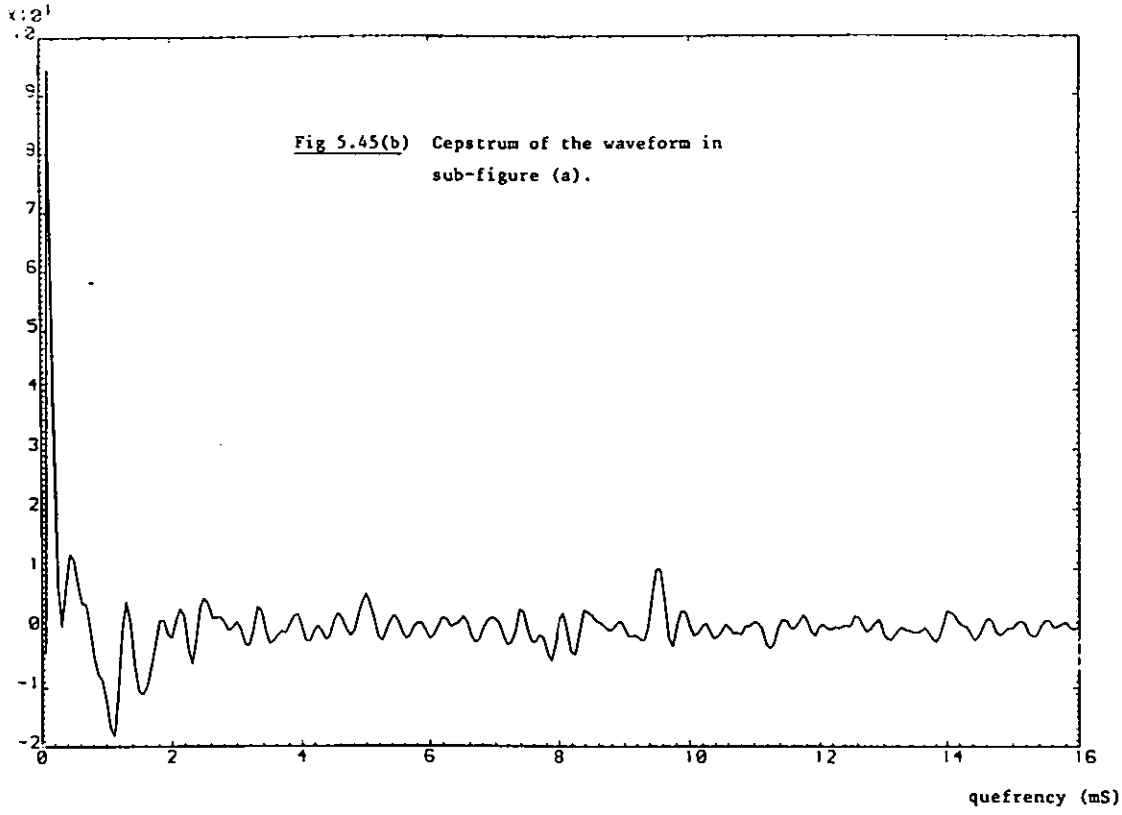


Fig 5.45(b) Cepstrum of the waveform in sub-figure (a).

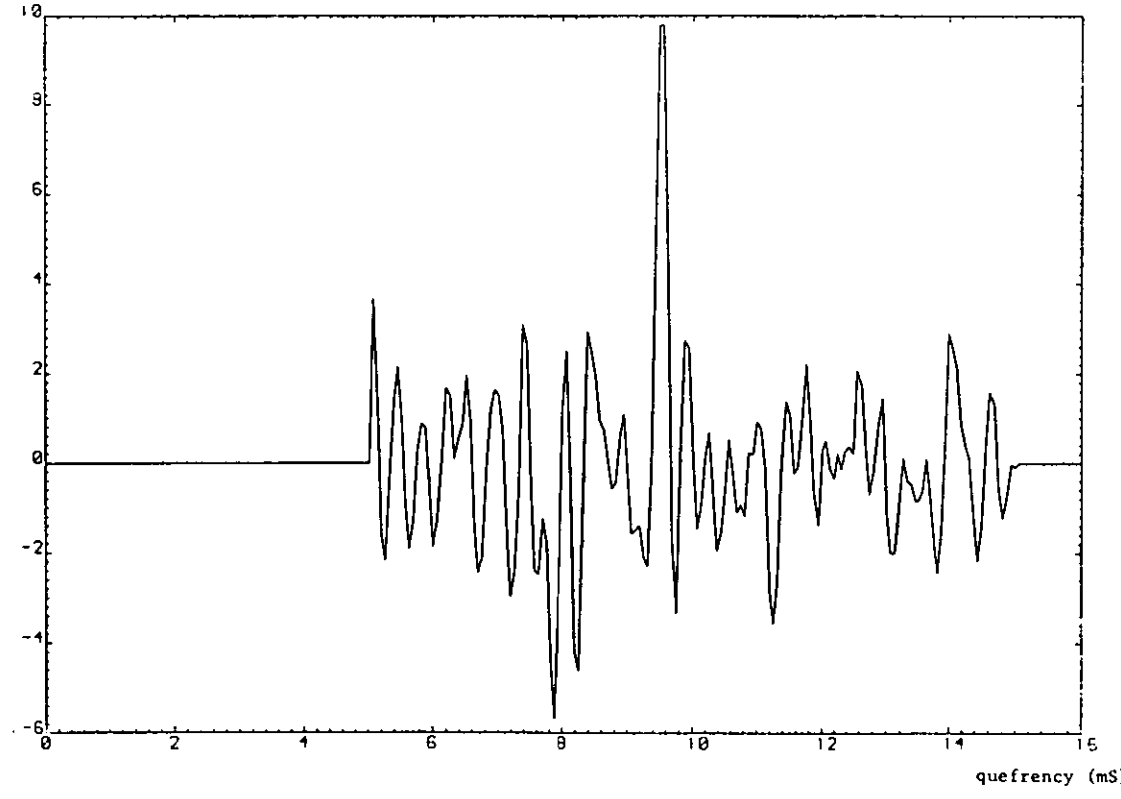


Fig 5.45(c) Truncated cepstrum of (b) from 5-15mS.

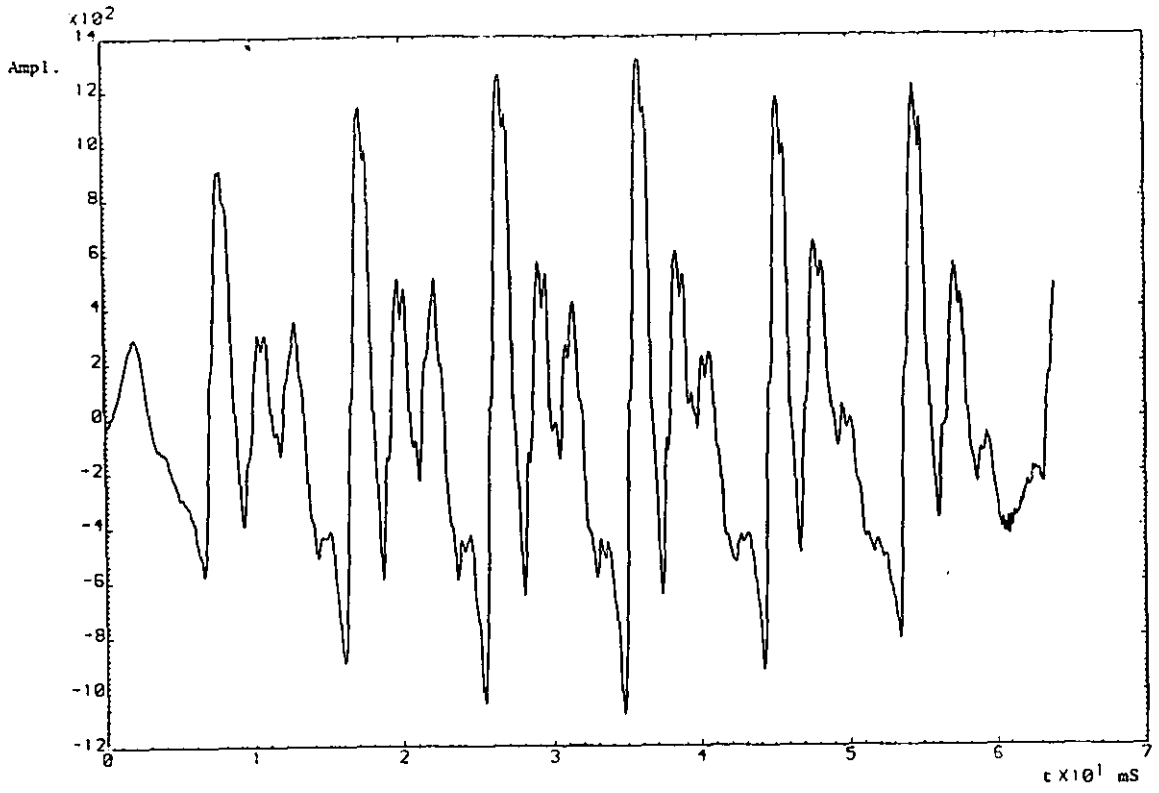


Fig 5.46(a) (i) Original 0 - 3.4 kHz waveform for an /I/ sound in "sister".

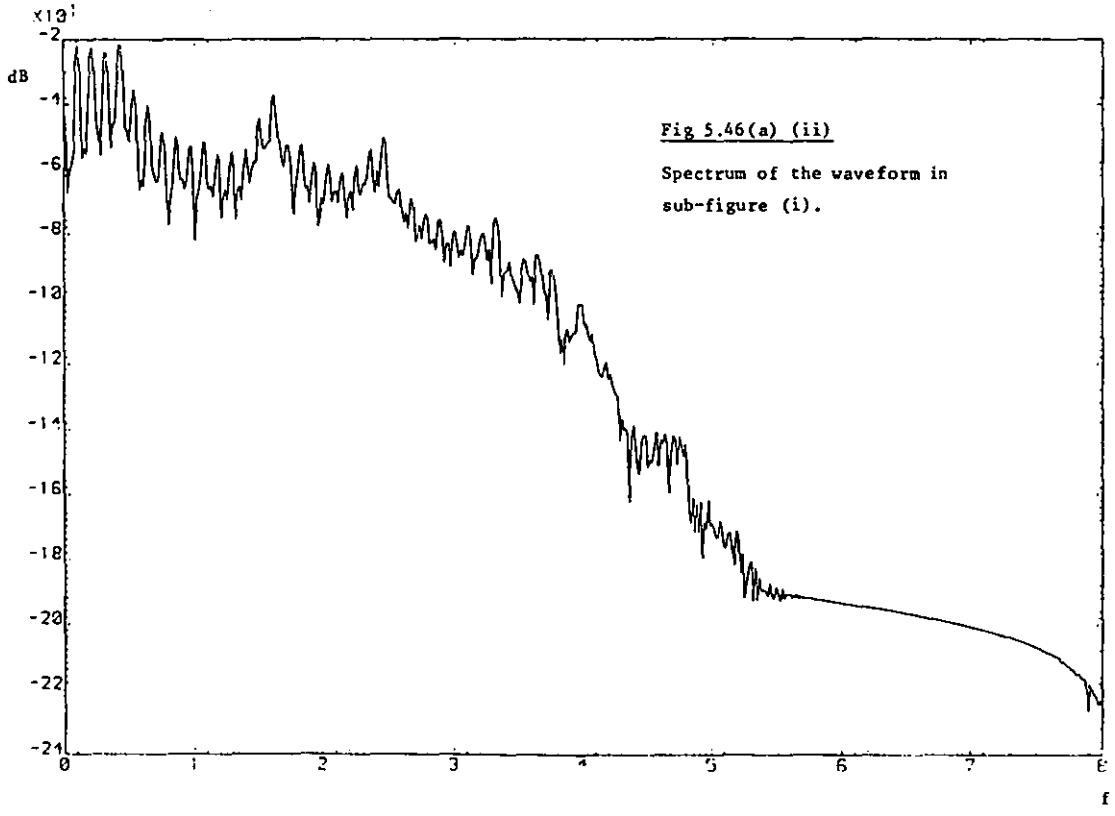


Fig 5.46(a) (ii)

Spectrum of the waveform in sub-figure (i).

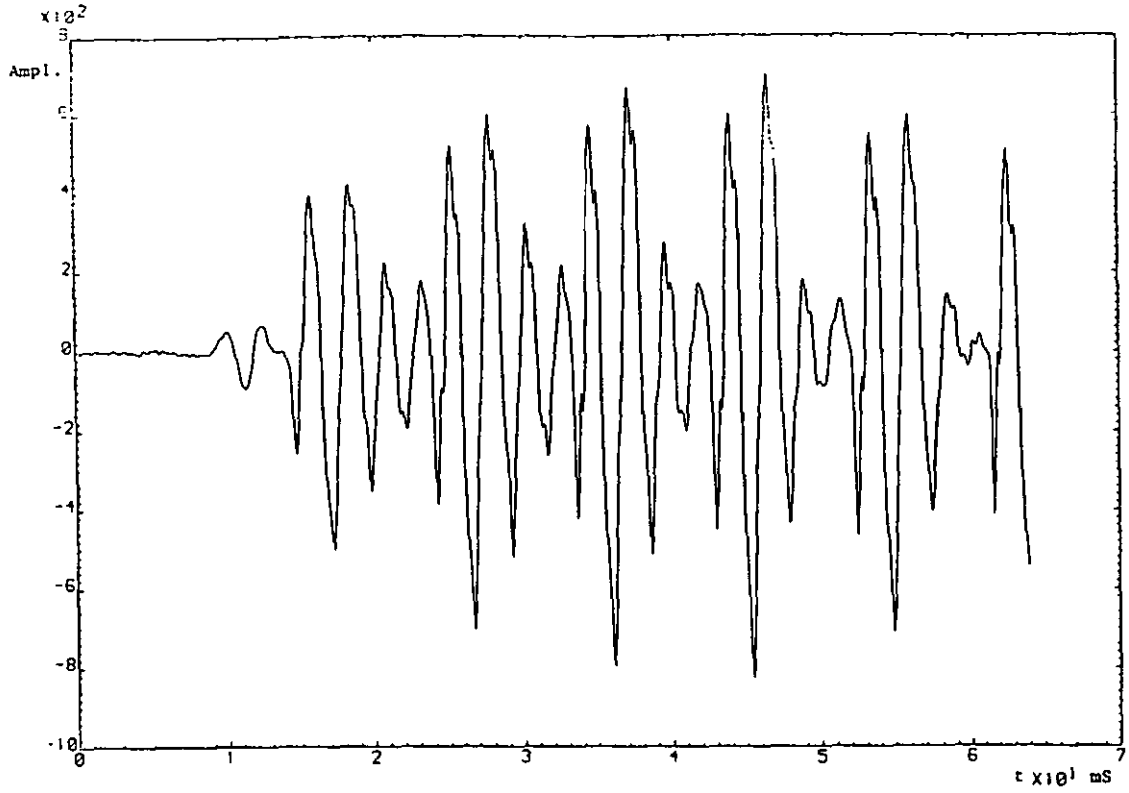


Fig 5.46(b) (i) Waveform of a 0.3 - 3.4 kHz bandlimited /I/ sound from "sister".

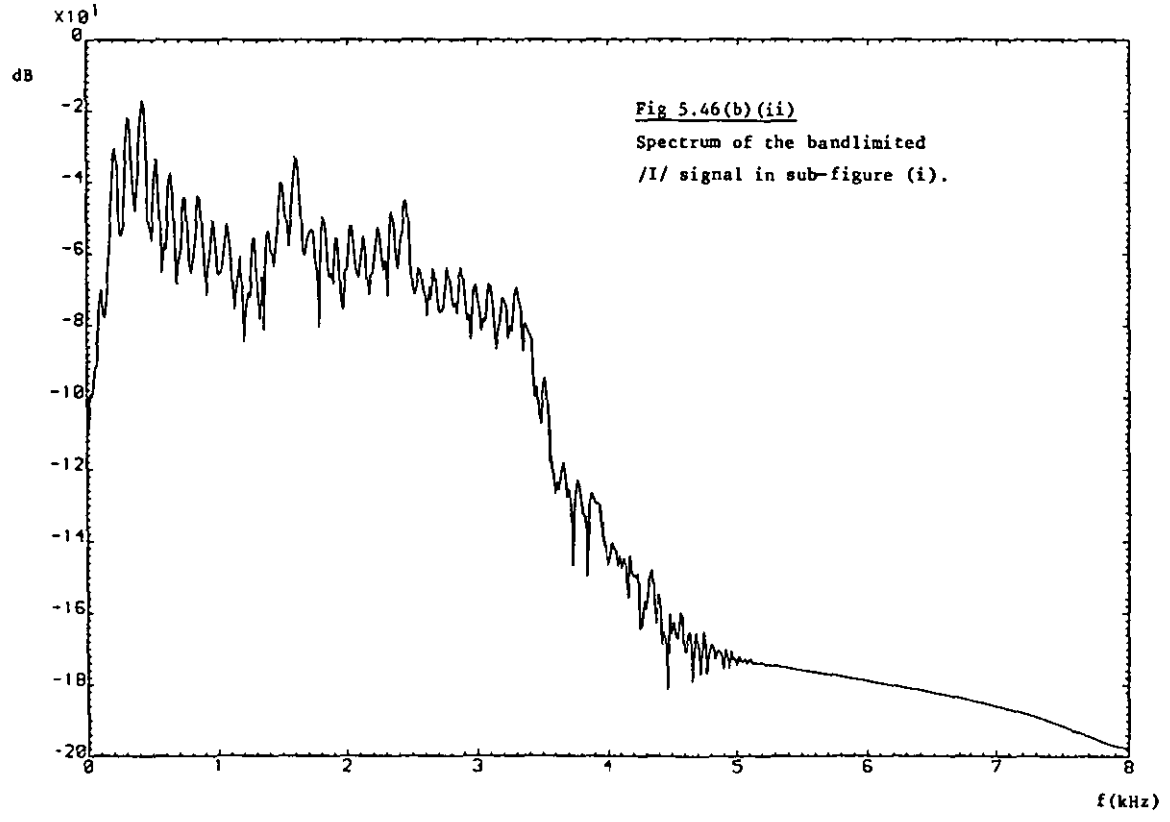


Fig 5.46(b)(ii)
Spectrum of the bandlimited /I/ signal in sub-figure (i).

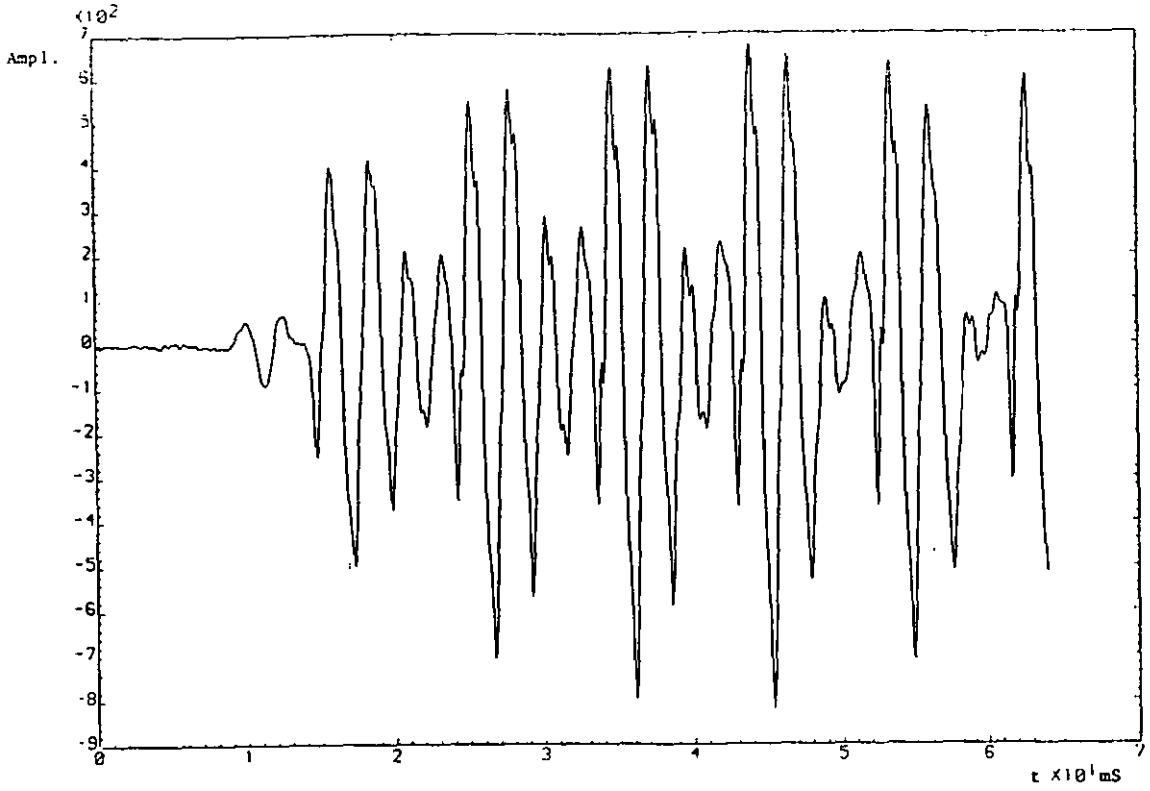


Fig 5.46(c)(i) Processed signal from the base-band synthesiser.

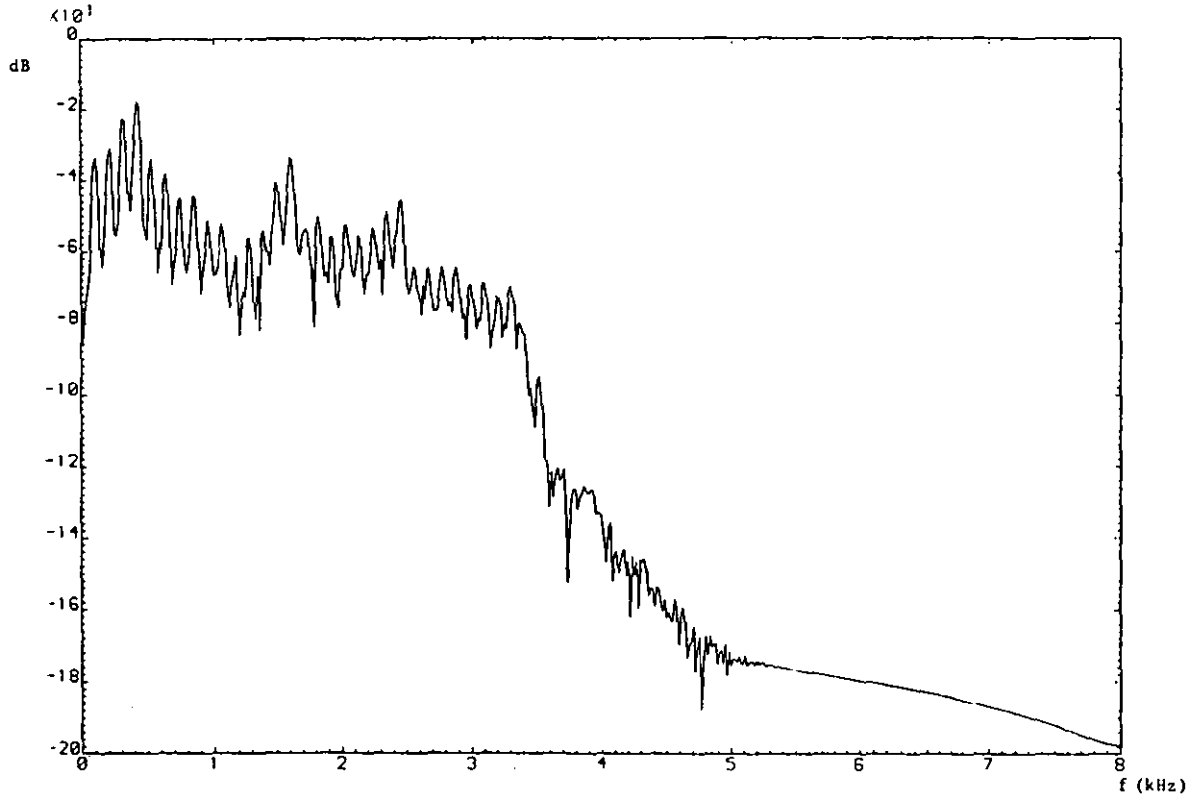


Fig 5.46(c)(ii) Spectrum of the processed signal.

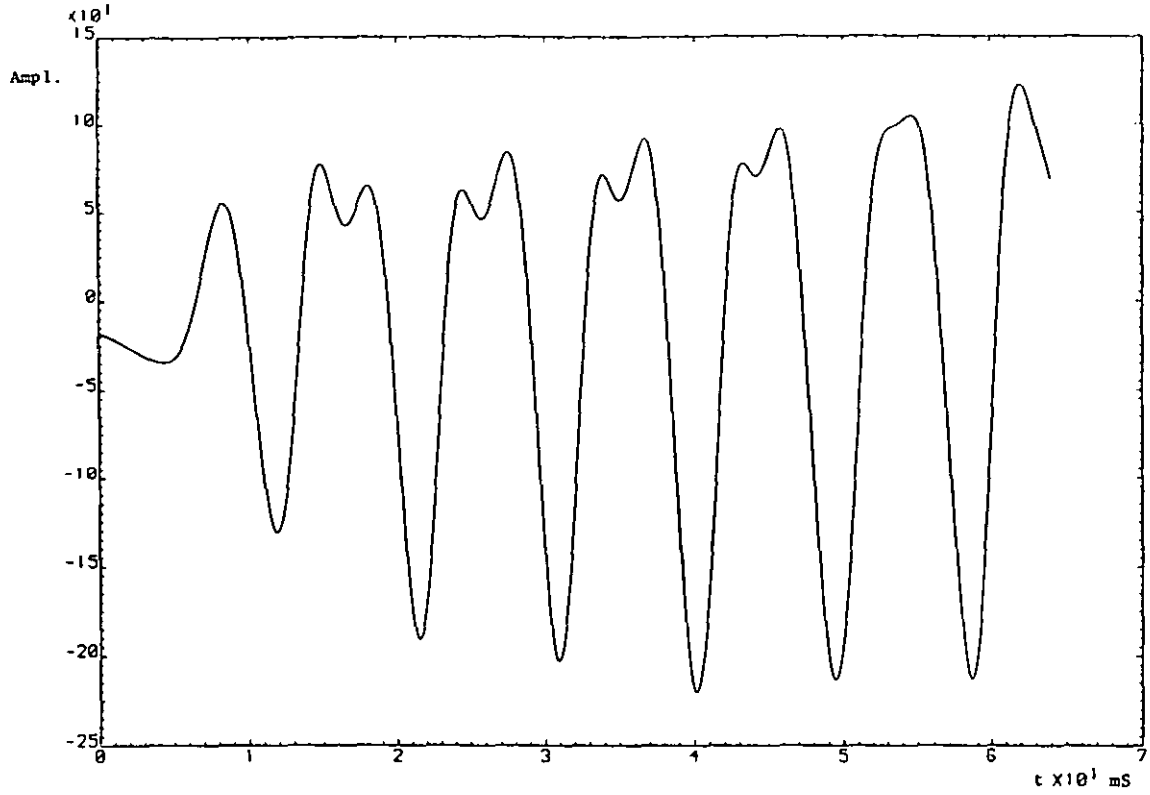


Fig 5.46(d) Original signal for an /I/ sound bandlimited from 0 to 300 Hz.

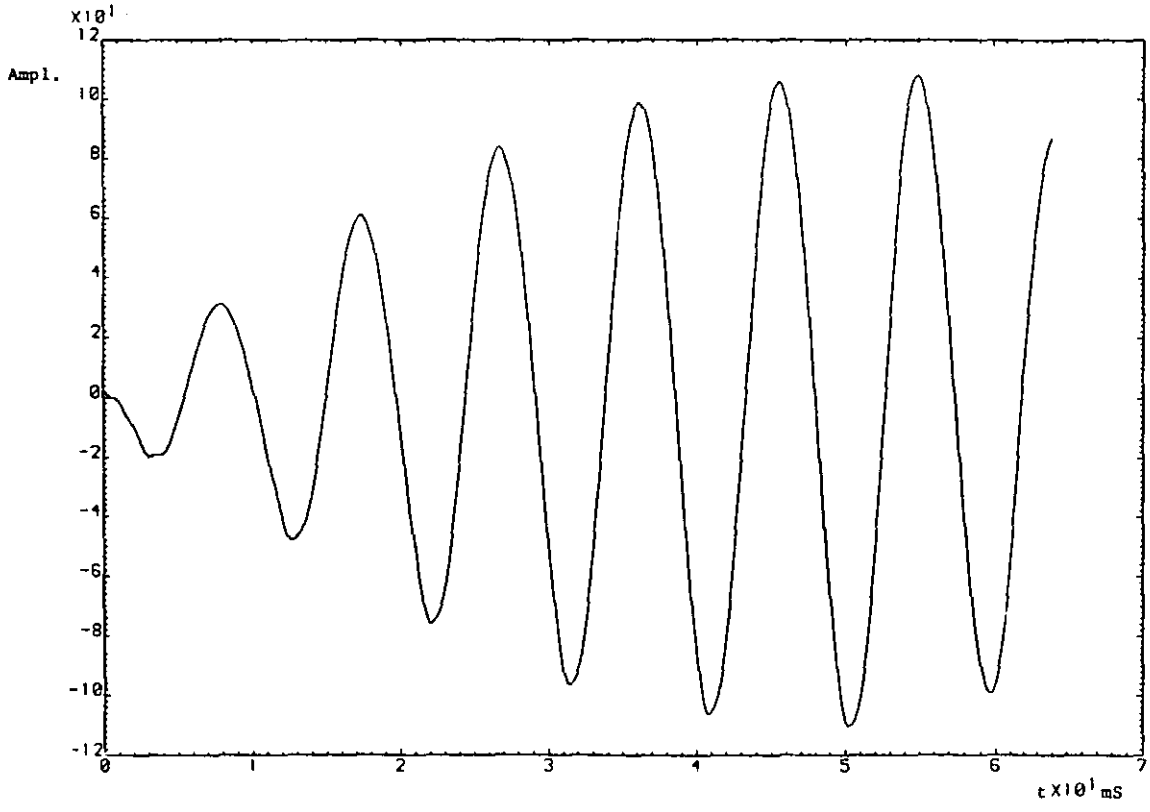


Fig 5.46(e)(i) 0 to 300 Hz synthesised signal.

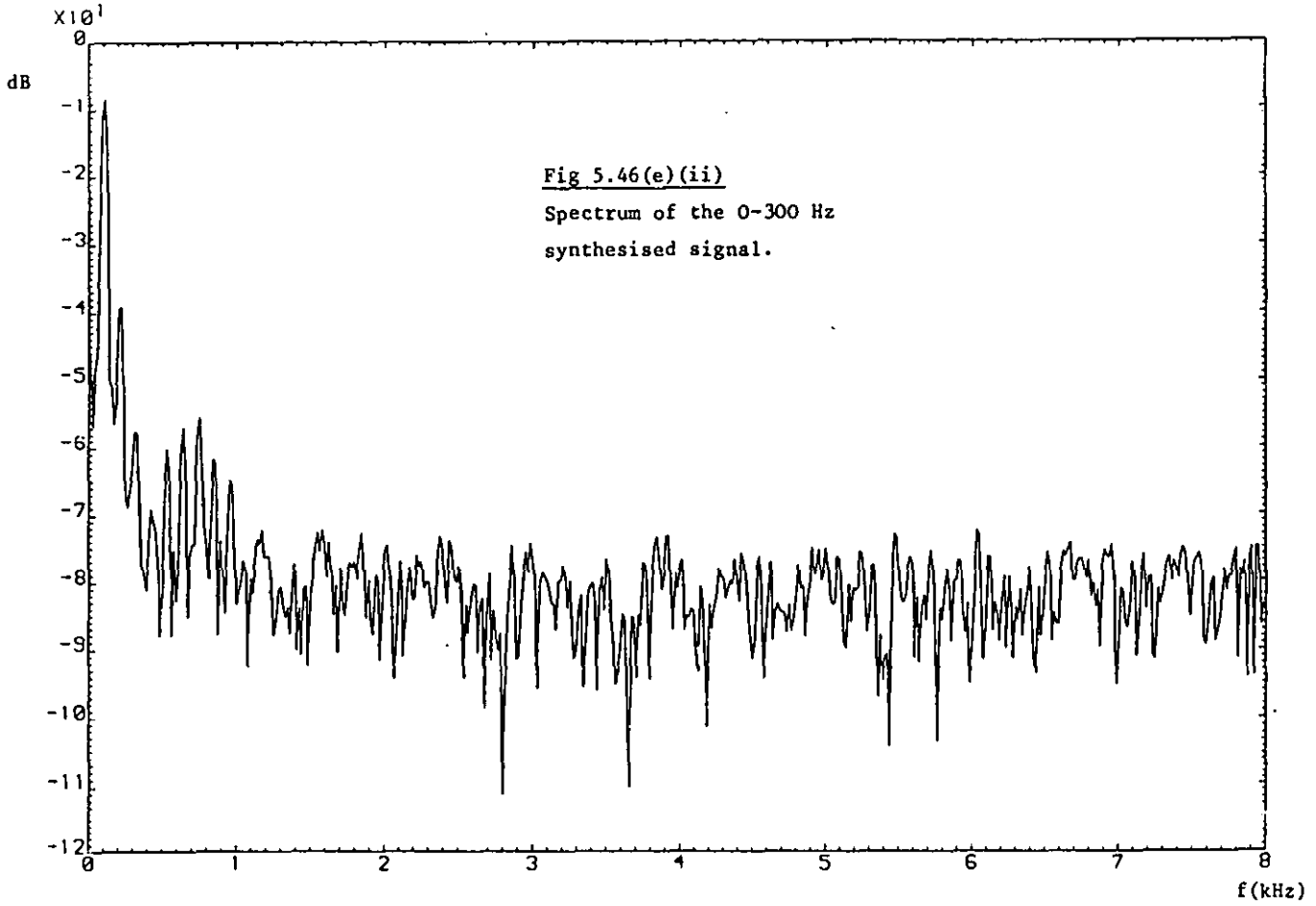


Fig 5.46(e)(ii)

Spectrum of the 0-300 Hz synthesised signal.

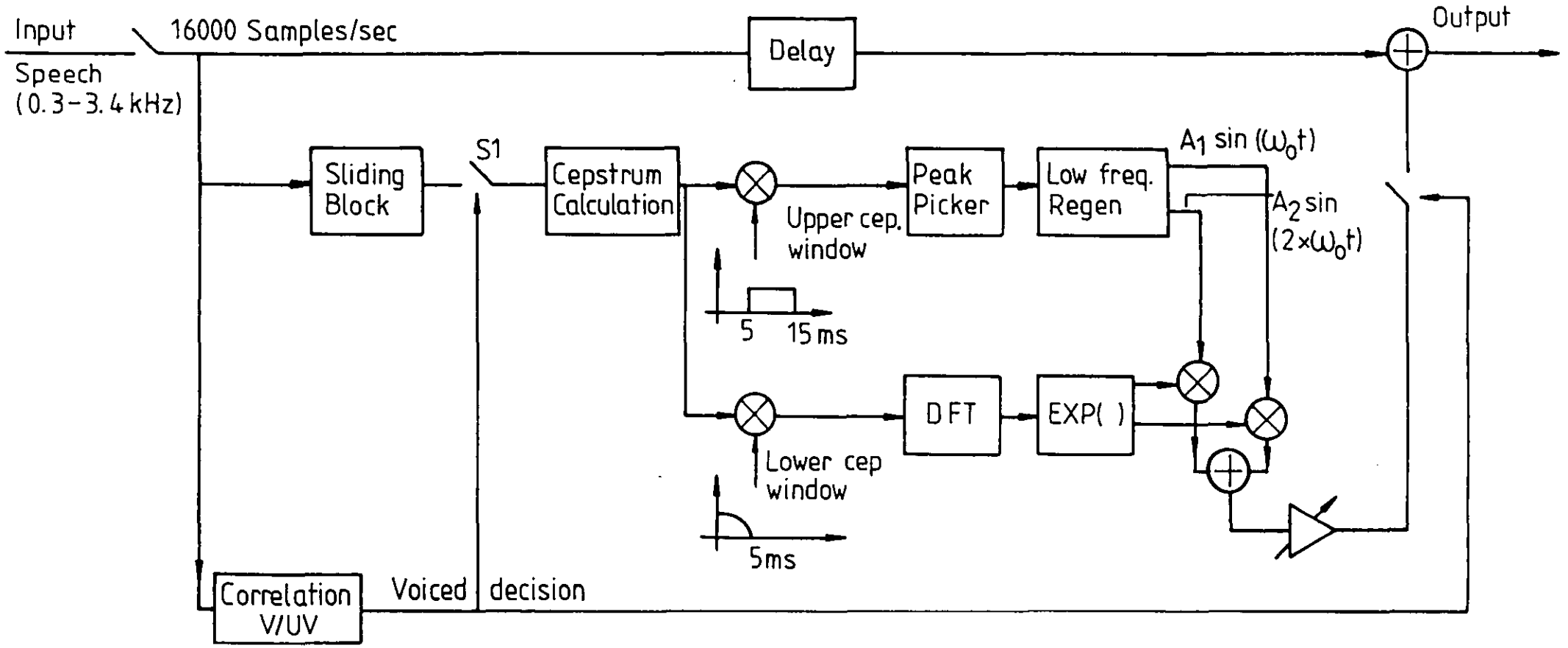


FIGURE 5·47 ENHANCEMENT OF TELEPHONE SPEECH QUALITY BY LOW FREQUENCY REGENERATION (REF 78)

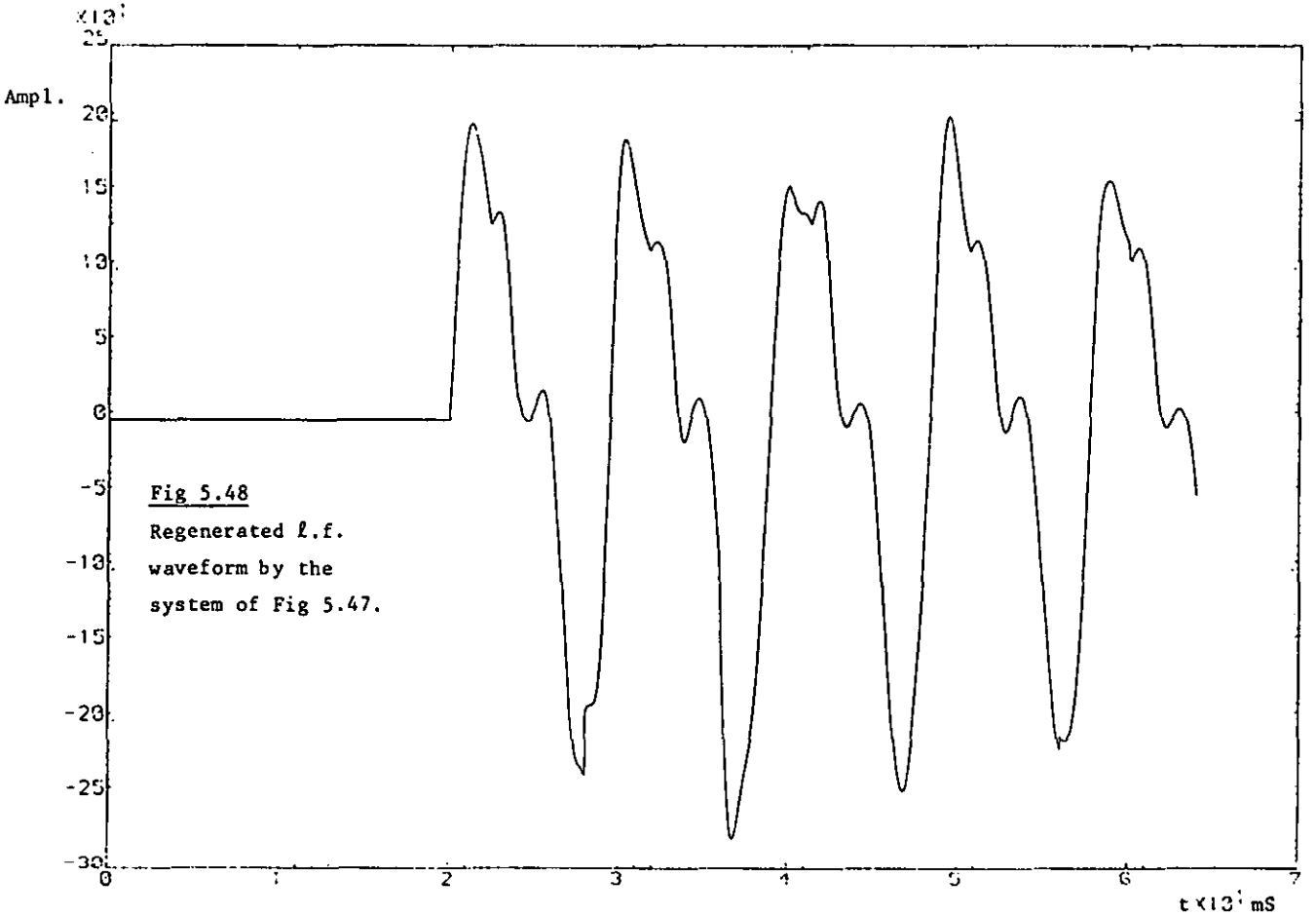


Fig 5.48
Regenerated ℓ.f. waveform by the system of Fig 5.47.

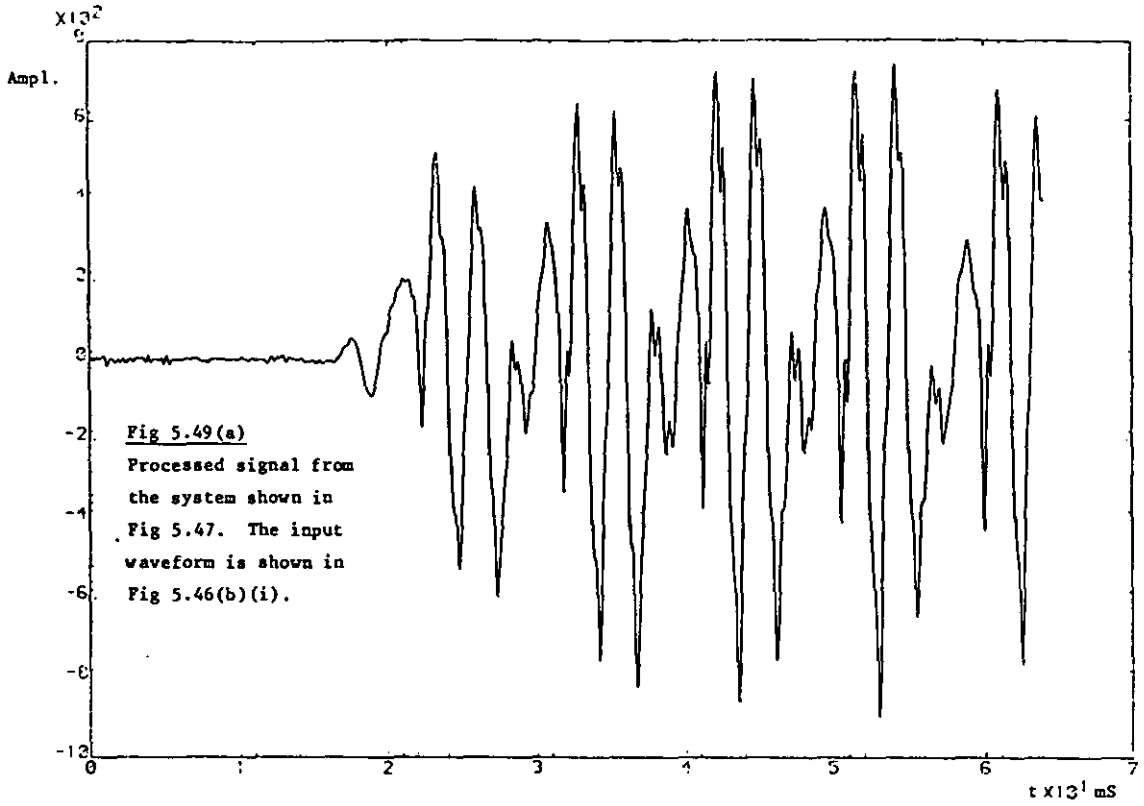


Fig 5.49(a)
Processed signal from the system shown in Fig 5.47. The input waveform is shown in Fig 5.46(b)(i).

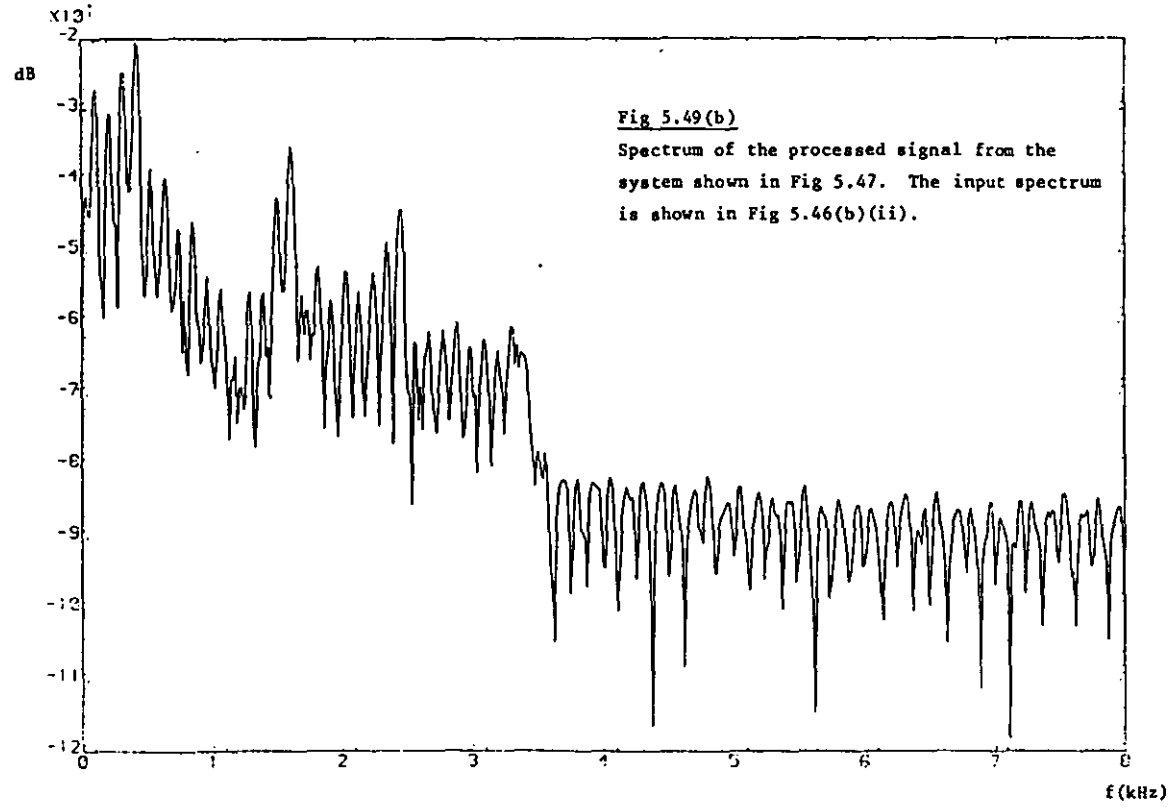


Fig 5.49(b)
Spectrum of the processed signal from the system shown in Fig 5.47. The input spectrum is shown in Fig 5.46(b)(ii).

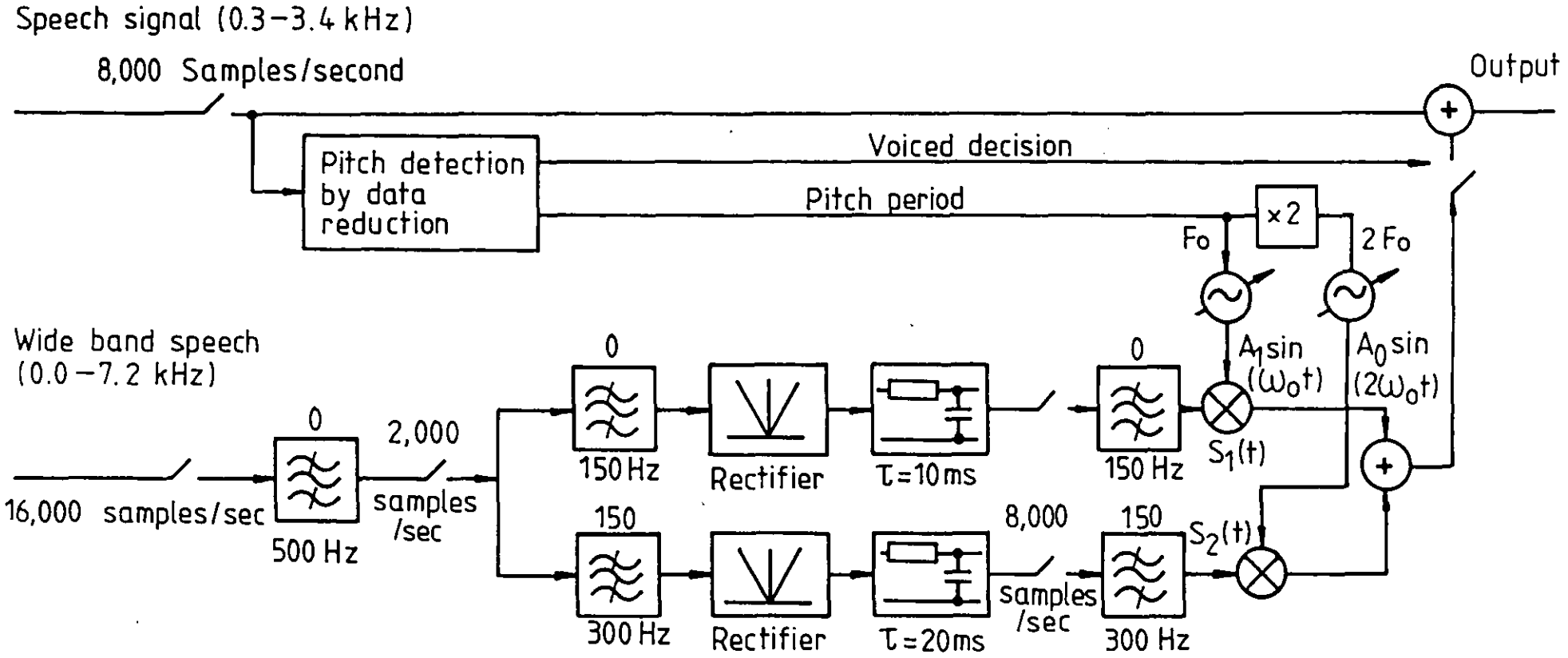


FIG. 5·50 ENHANCEMENT OF TELEPHONE SPEECH QUALITY BY LOW FREQUENCY REGENERATION (REF 78)

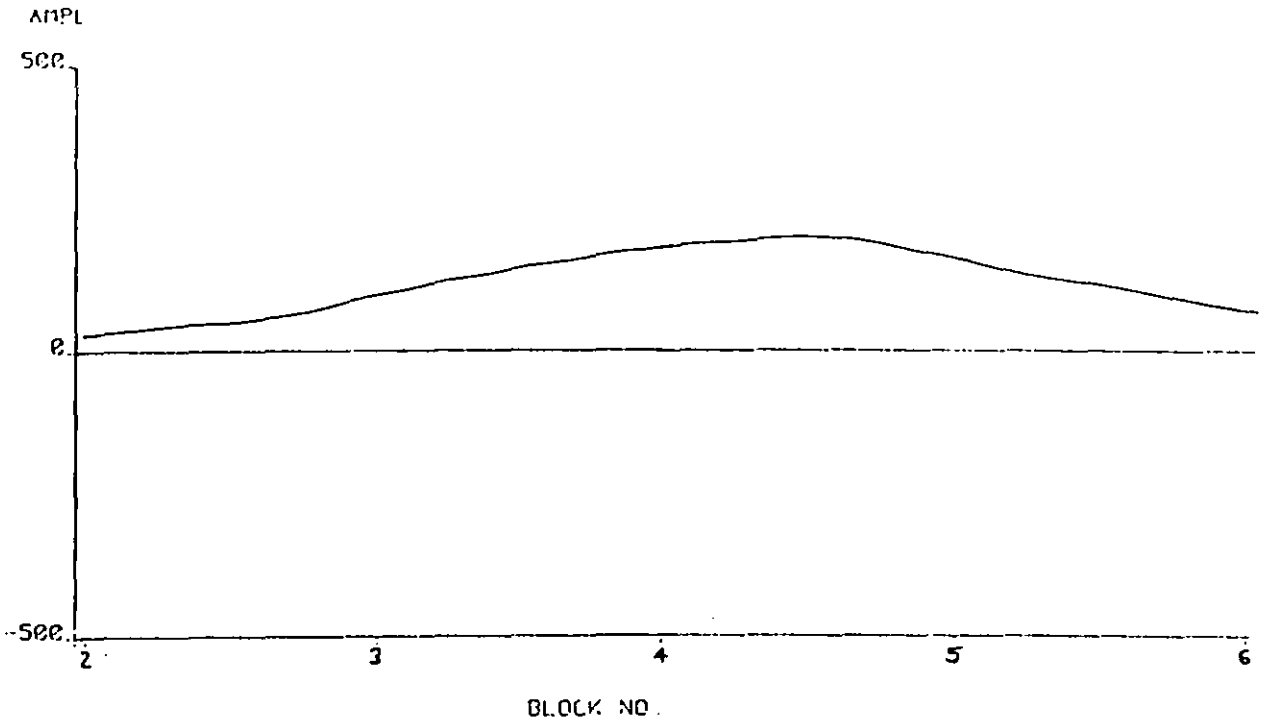


Fig 5.51(a). Envelope waveform of /I/ in "sister" from the 150 Hz LPF. This is used to shape the regenerated fundamental frequency signal (ref.78).

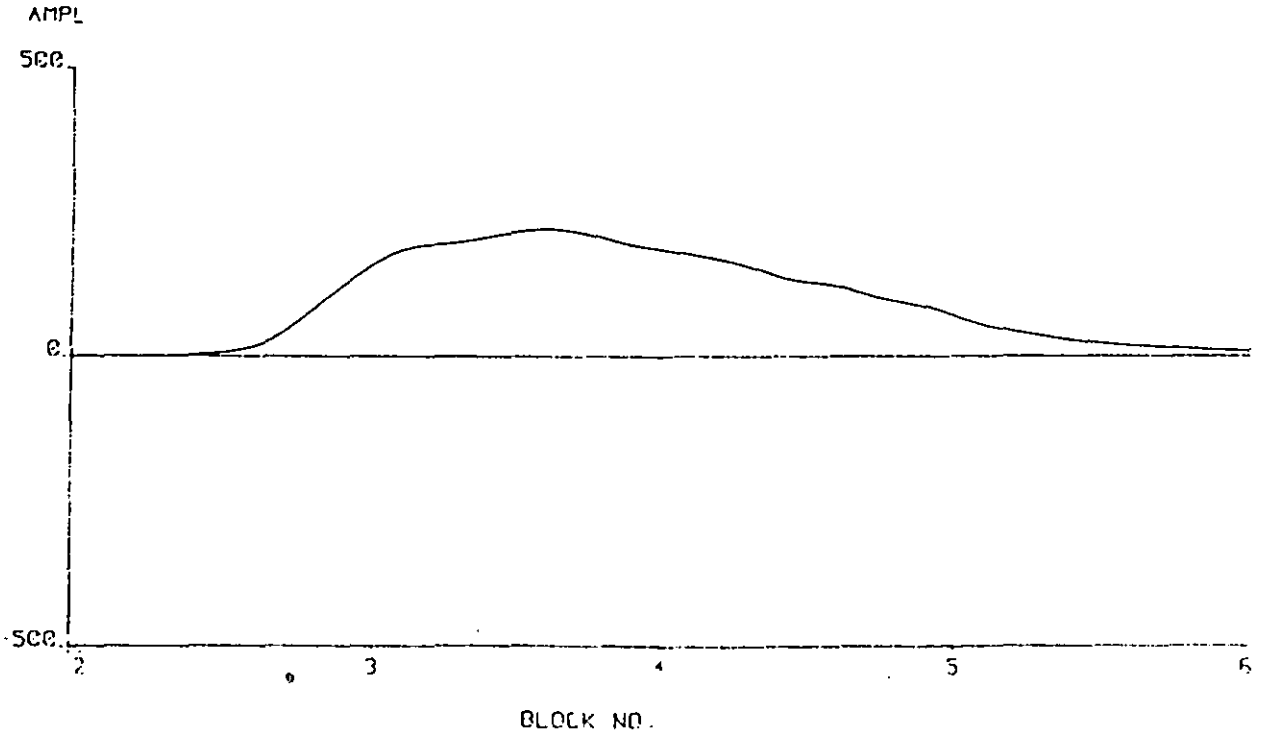


Fig 5.51(b). Envelope waveform of /I/ in "sister" from the output of the 150-300 Hz BPF. This is used to shape the regenerated second harmonic signal (ref.78).

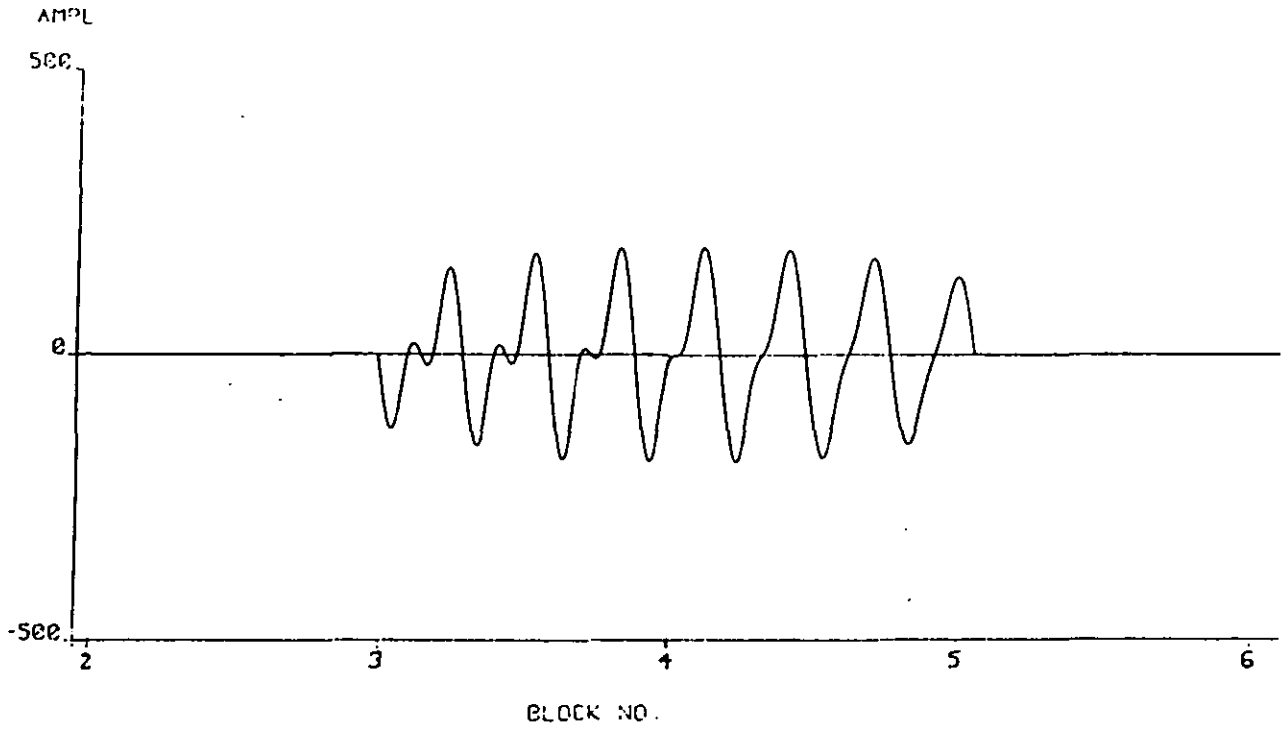


Fig. 5.51(c). The regenerated low frequency waveform by the system in Fig 5.50 (ref.78)

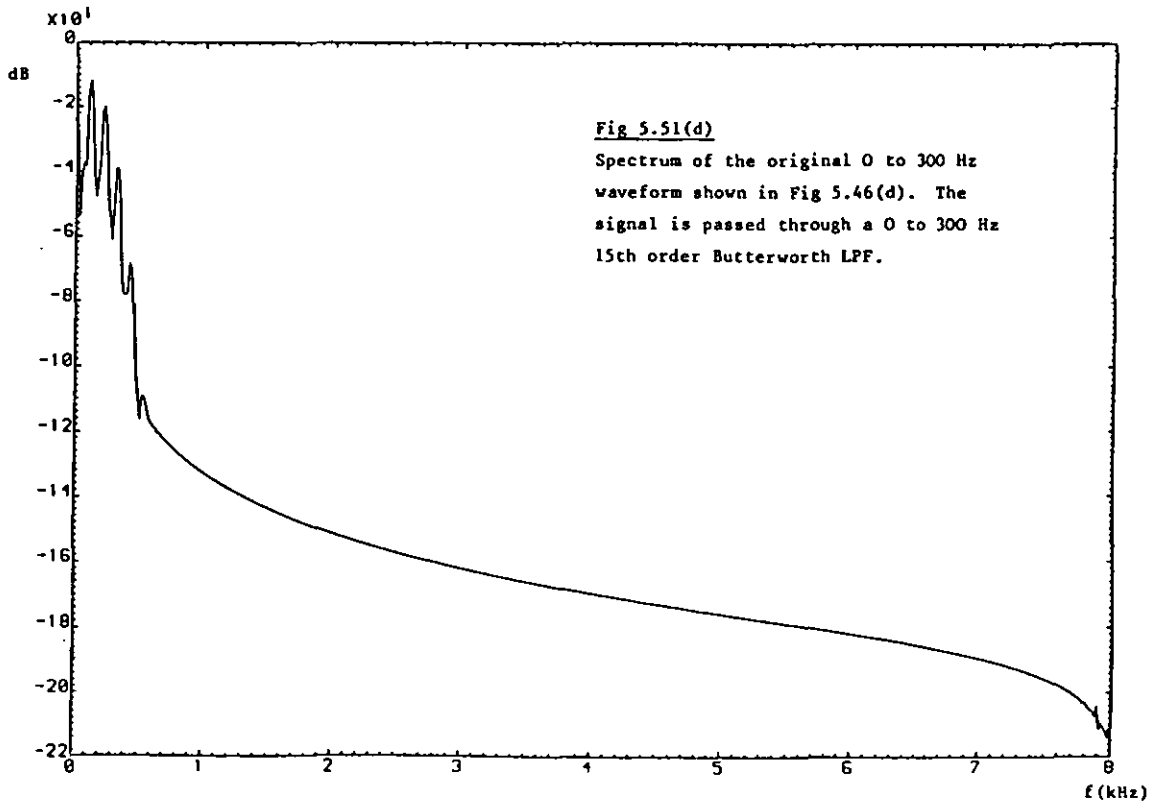


Fig 5.51(d)
Spectrum of the original 0 to 300 Hz waveform shown in Fig 5.46(d). The signal is passed through a 0 to 300 Hz 15th order Butterworth LPF.

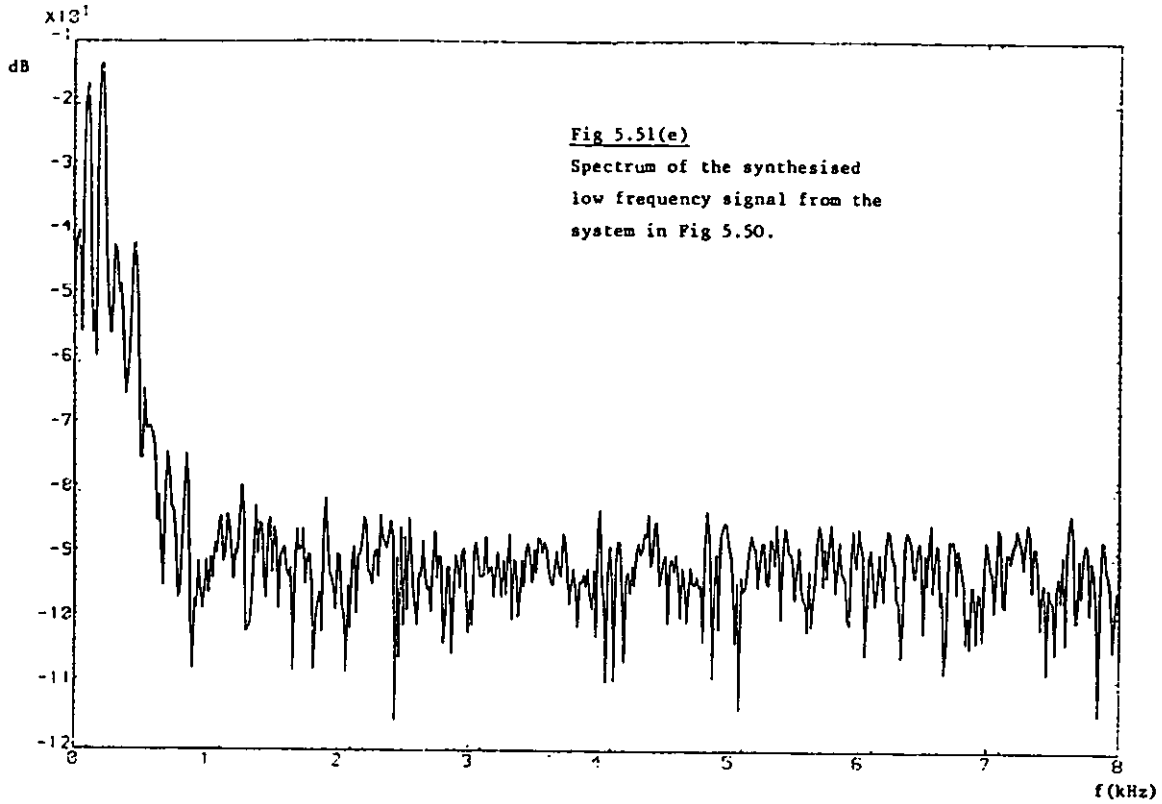


Fig 5.51(e)
Spectrum of the synthesised low frequency signal from the system in Fig 5.50.

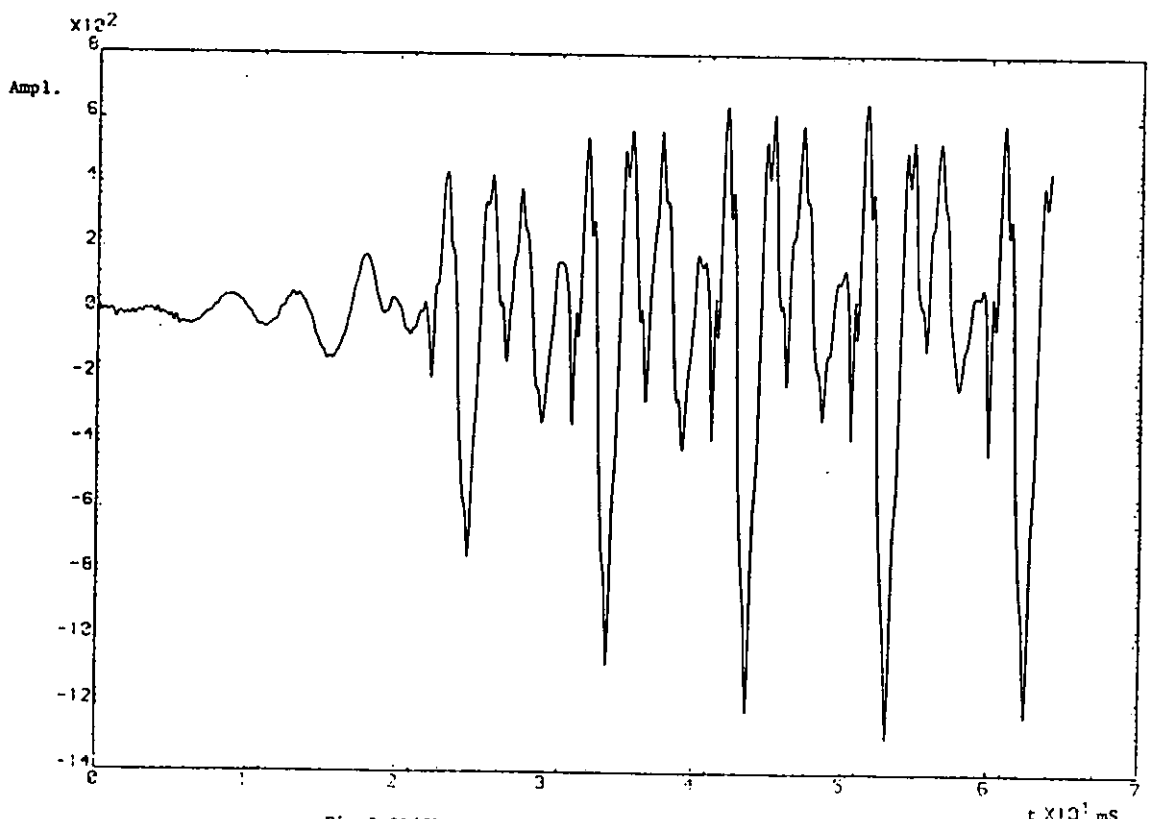


Fig 5.51(f) Processed output signal from the system shown in Fig 5.50.

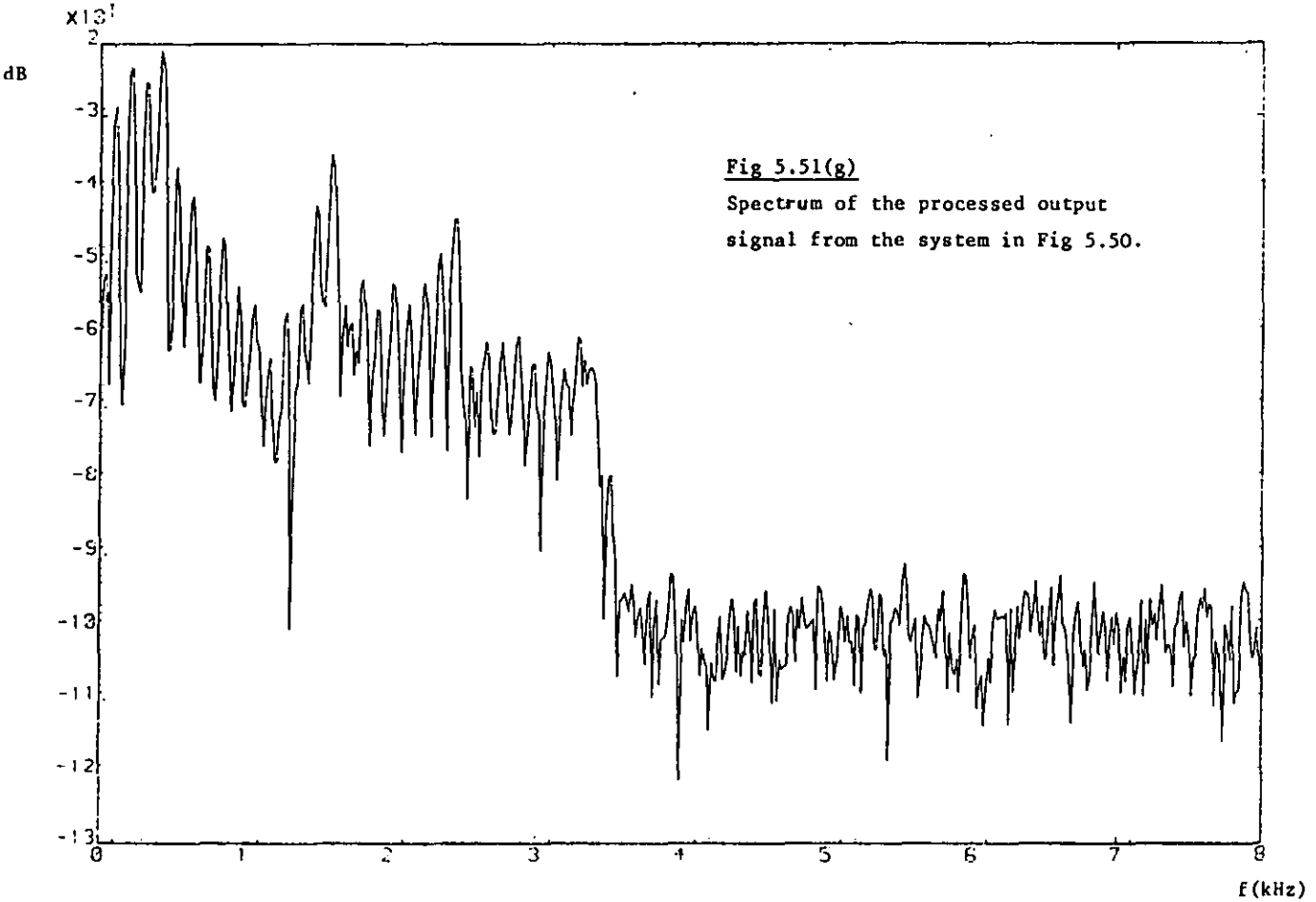


Fig 5.51(g)

Spectrum of the processed output signal from the system in Fig 5.50.

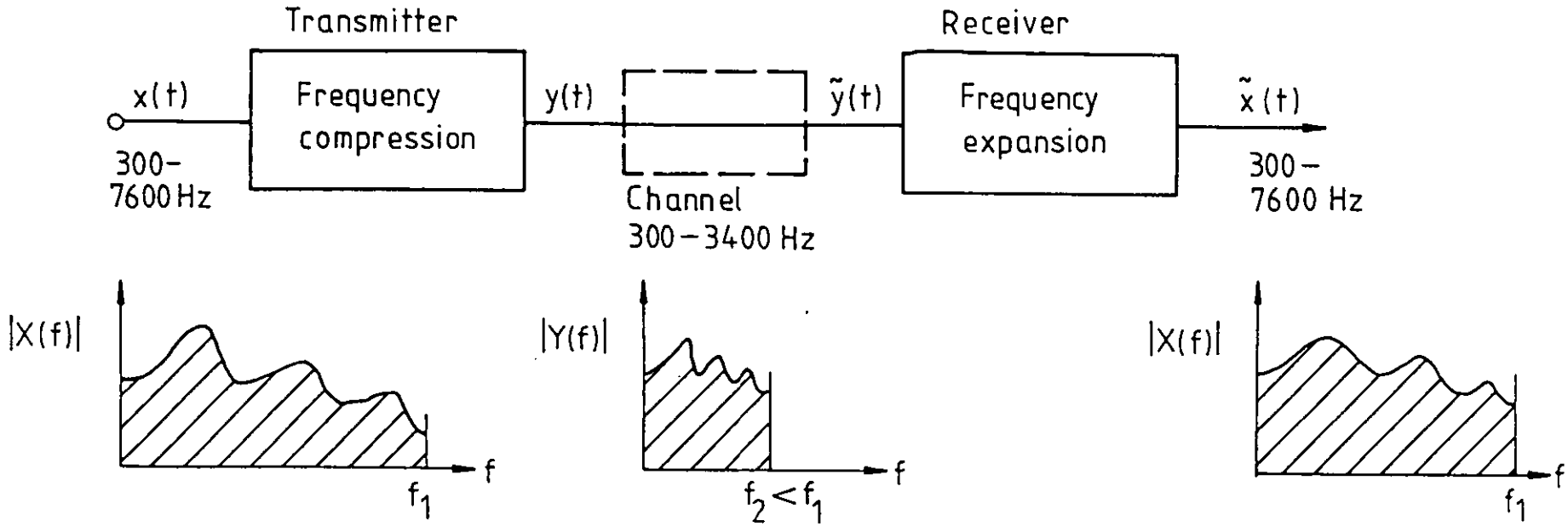
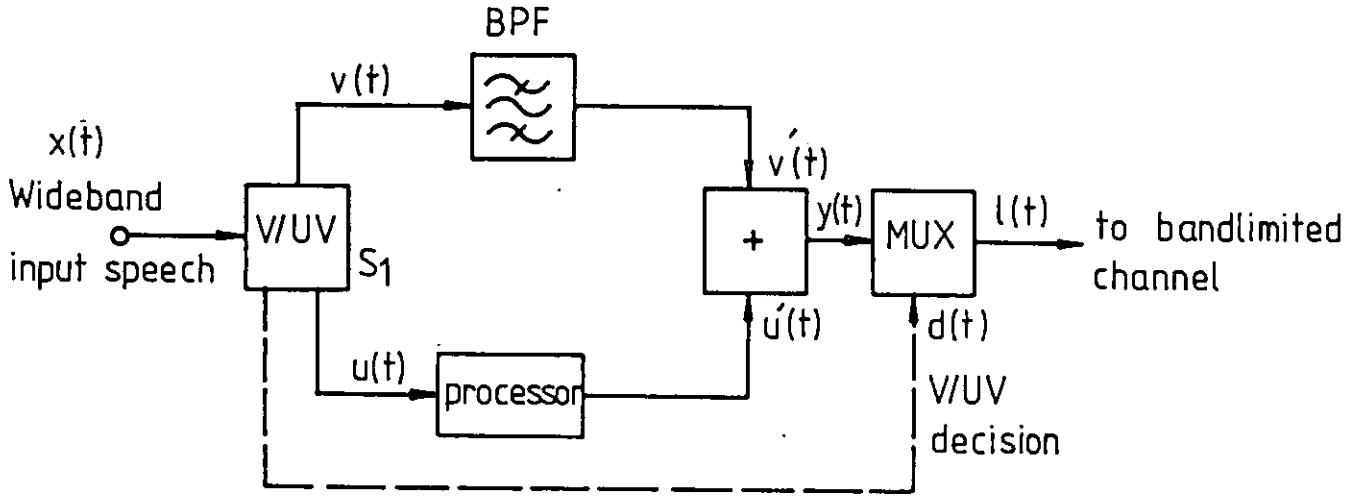
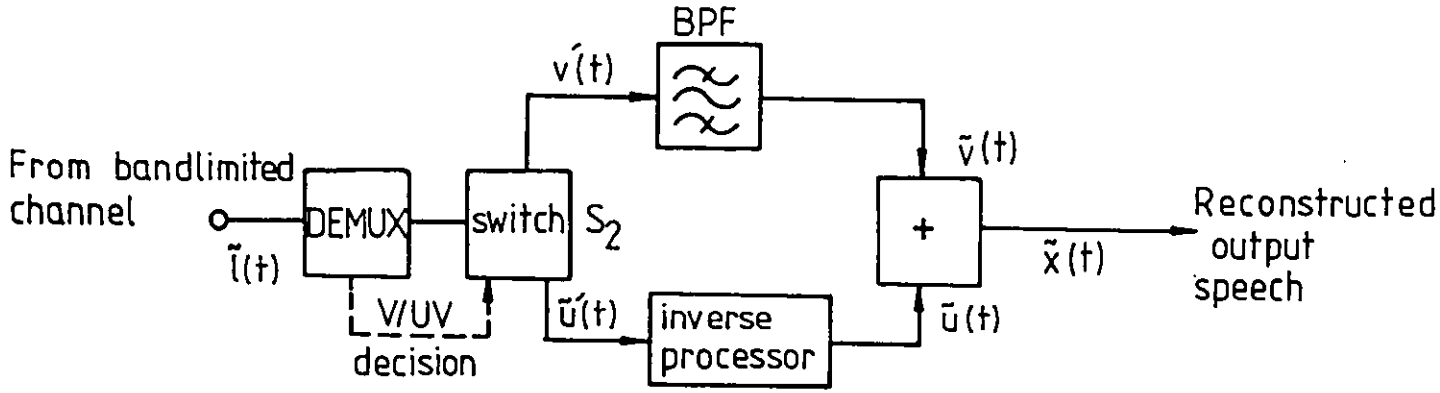


FIG. 6·1 GENERALISED FREQUENCY COMPRESSION



(a) Transmitter



(b) Receiver

FIG. 6·2 FRAMEWORK FOR BANDWIDTH COMPRESSION

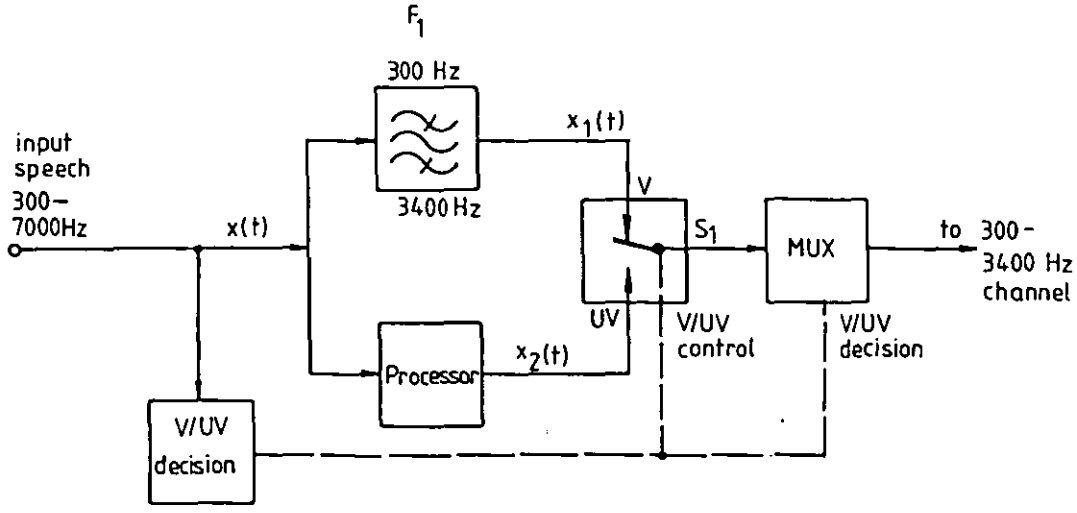


FIG. 6·3 (a) Transmitter

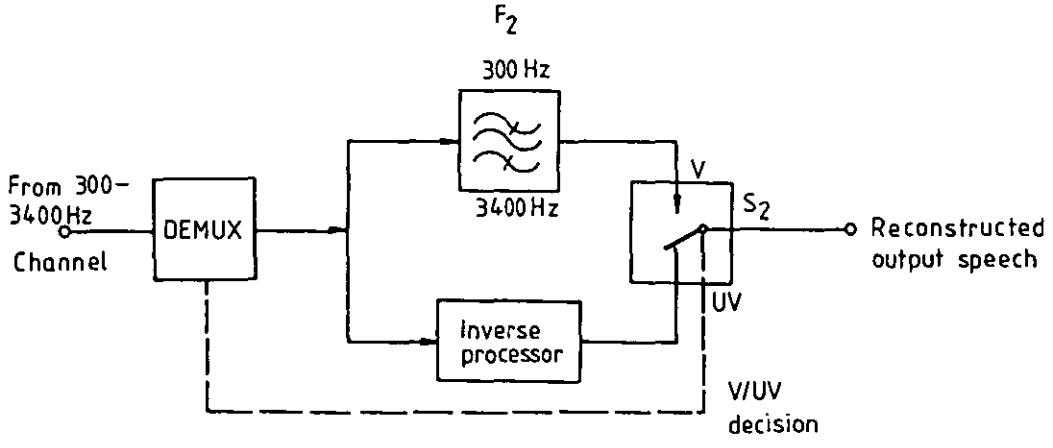


FIG. 6·3 (b) RECEIVER

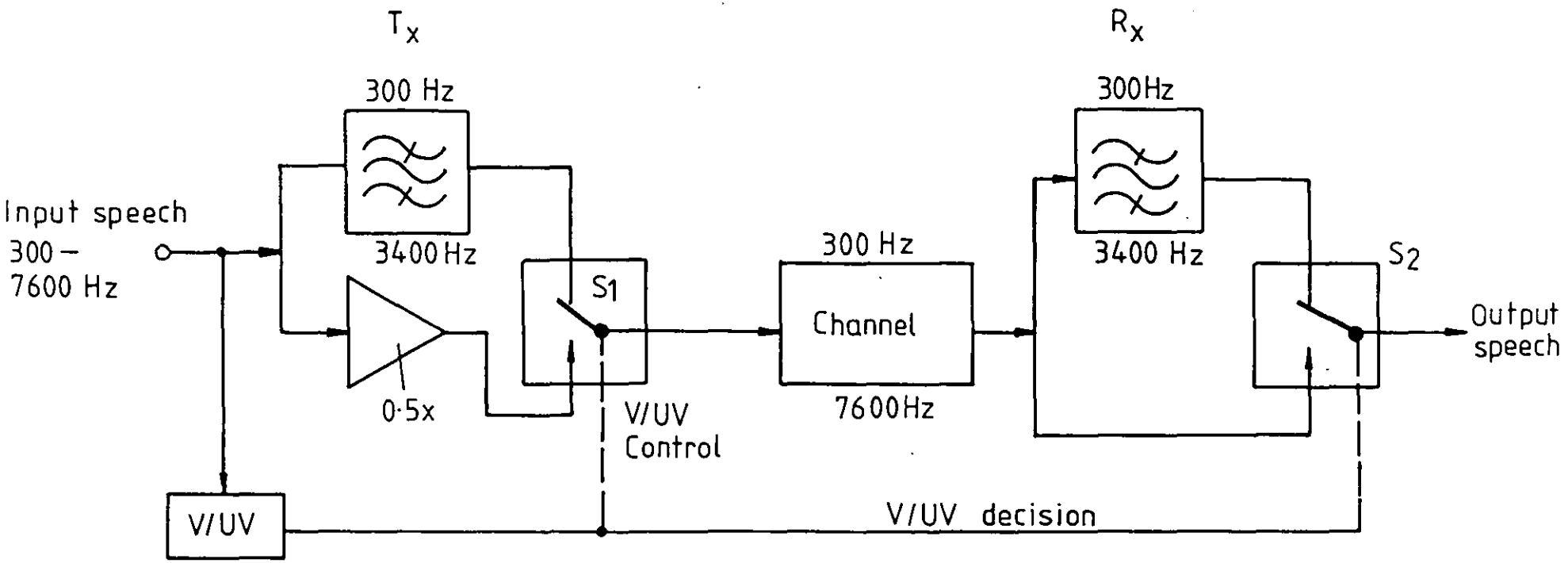

Tx
Rx
300 Hz
300Hz
Input speech
300 –
7600 Hz
3400 Hz
3400 Hz
S1
300 Hz
Channel
7600Hz
S2
Output
speech
0·5x
V/UV
Control
V/UV
V/UV decision

FIG. 6·4

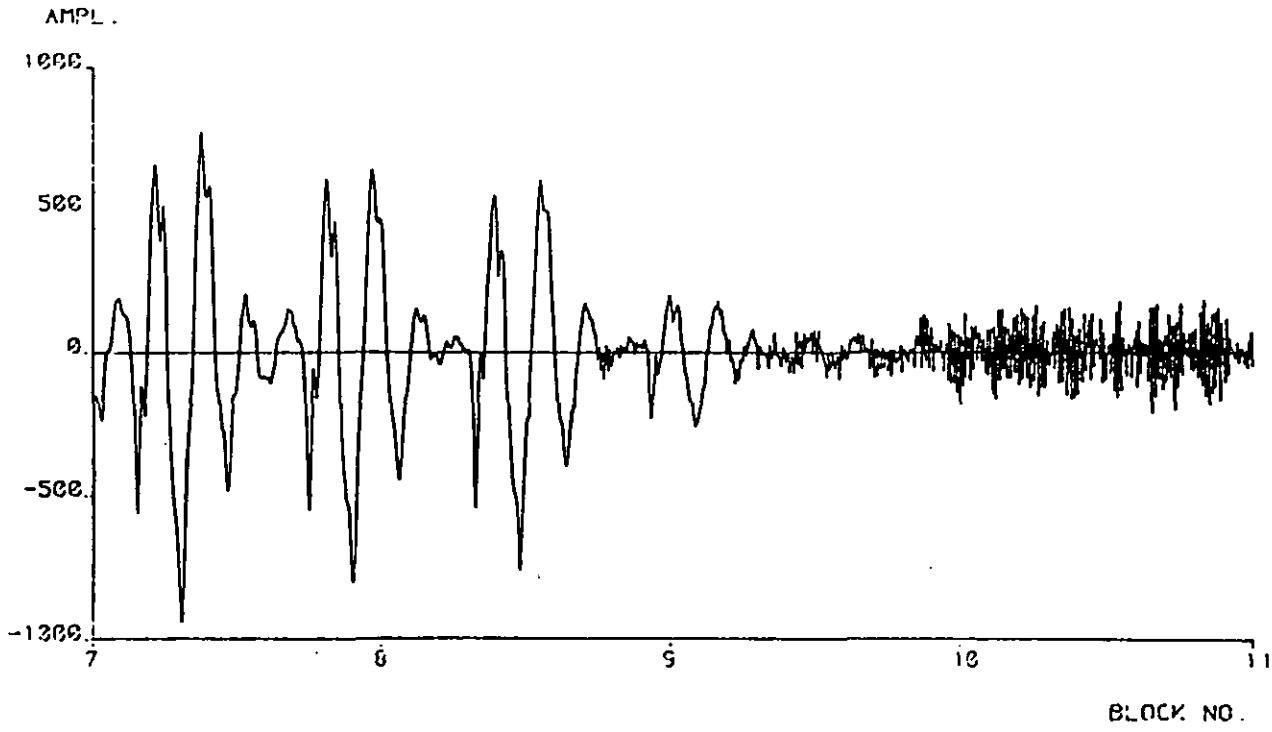


Fig 6.5(a) Time waveform for /Is/ in sister:
Original 0.3-6.0 kHz signal.

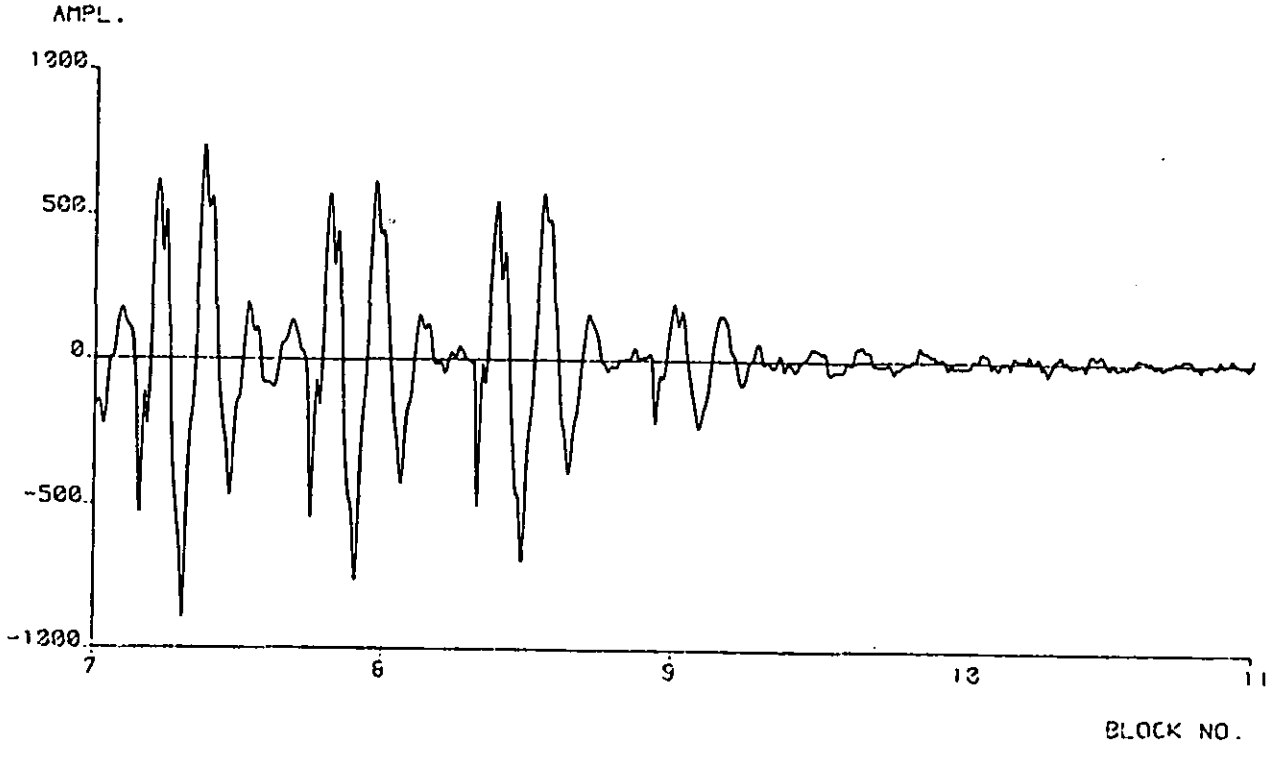


Fig 6.5(b) Bandlimited speech signal 0.3-3.4 kHz.

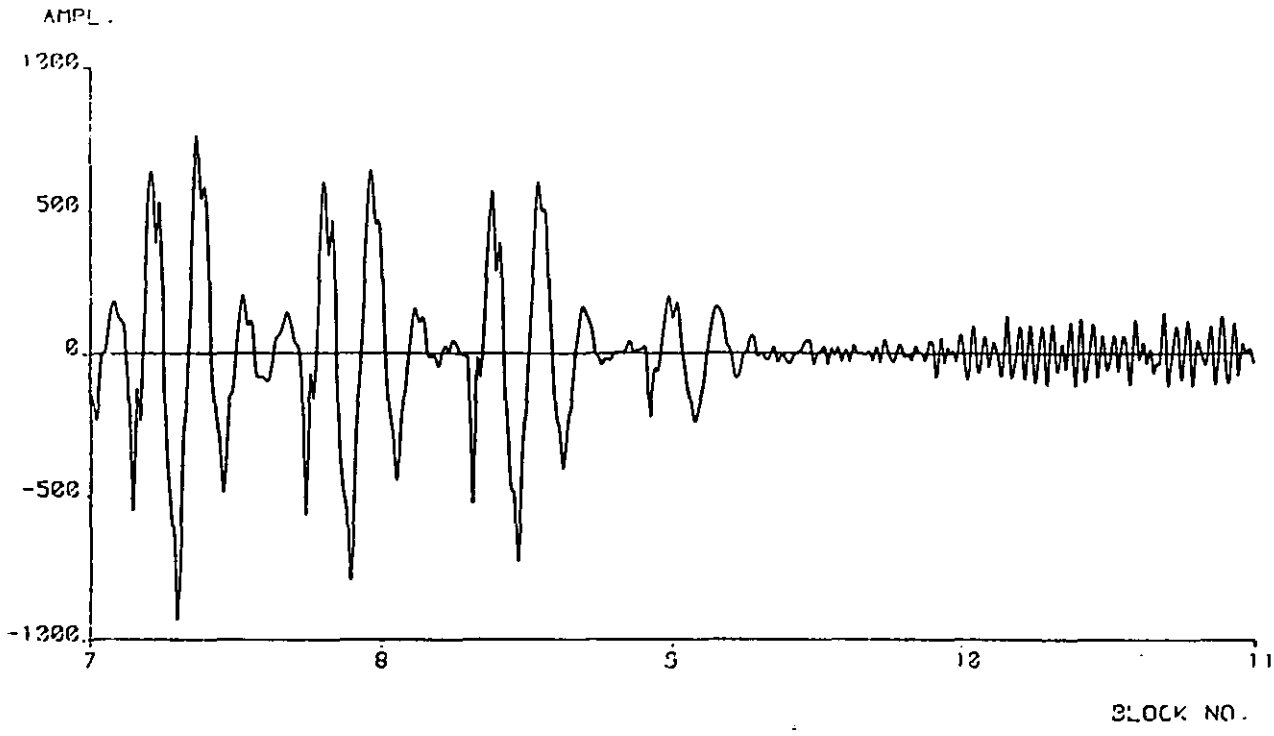


Fig 6.5(c) Transmitted VUBS signal 0.3-3.4 kHz.

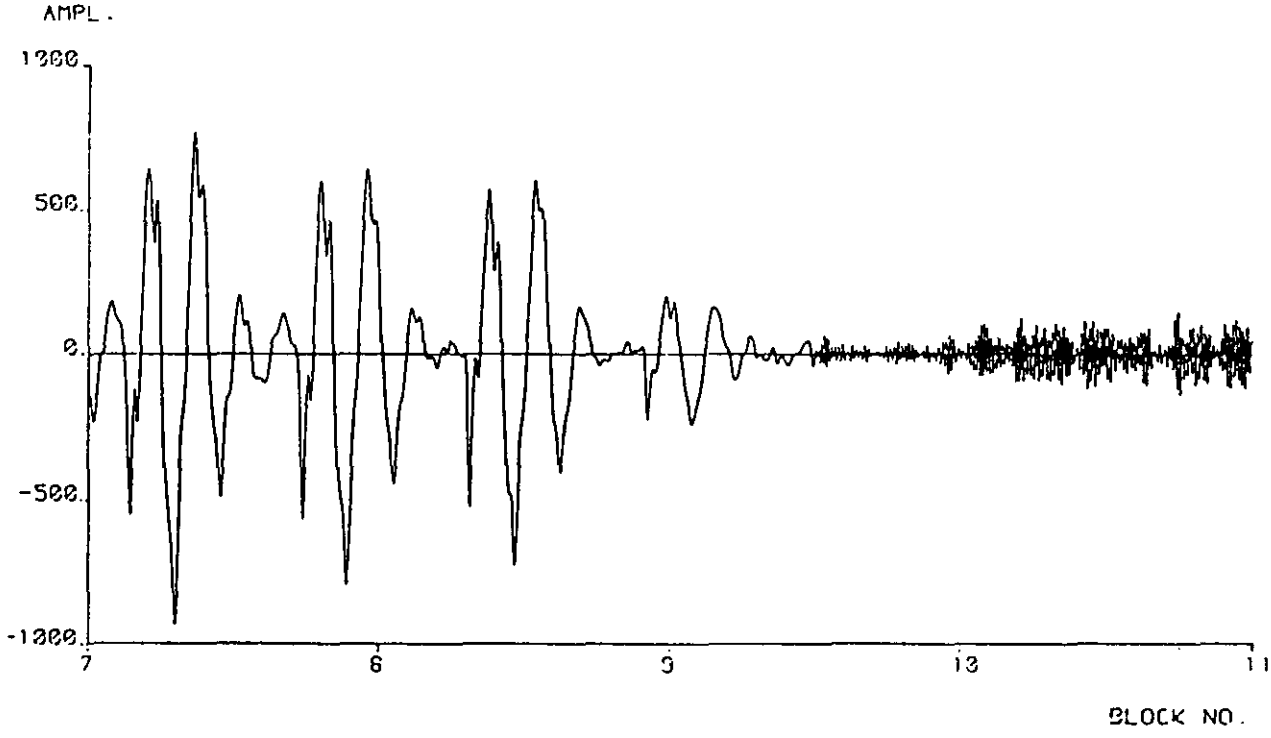
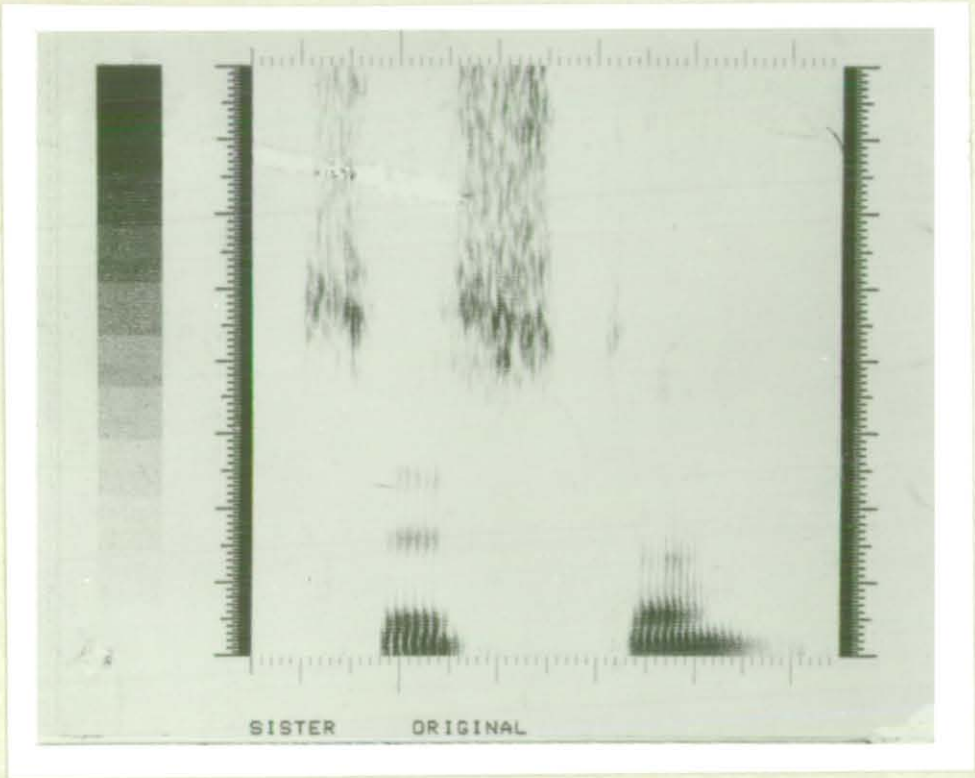
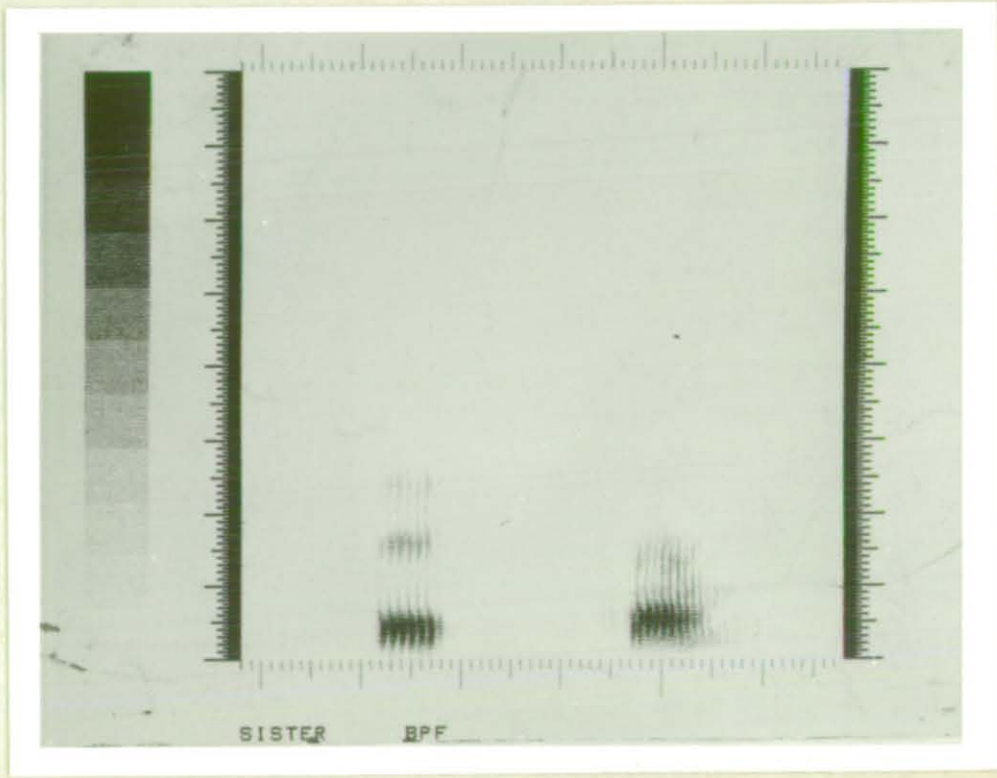


Fig 6.5(d) Reconstructed VUBS speech signal 0.3-6.0 kHz.

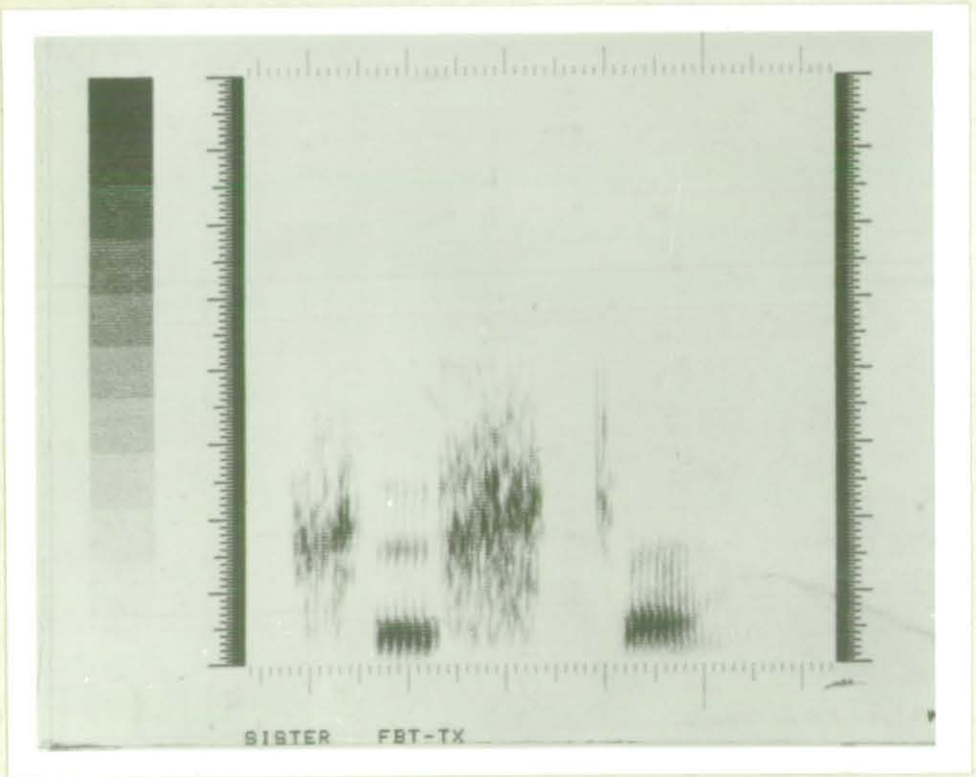


(a) Original speech signal 0.3 - 6.0 kHz

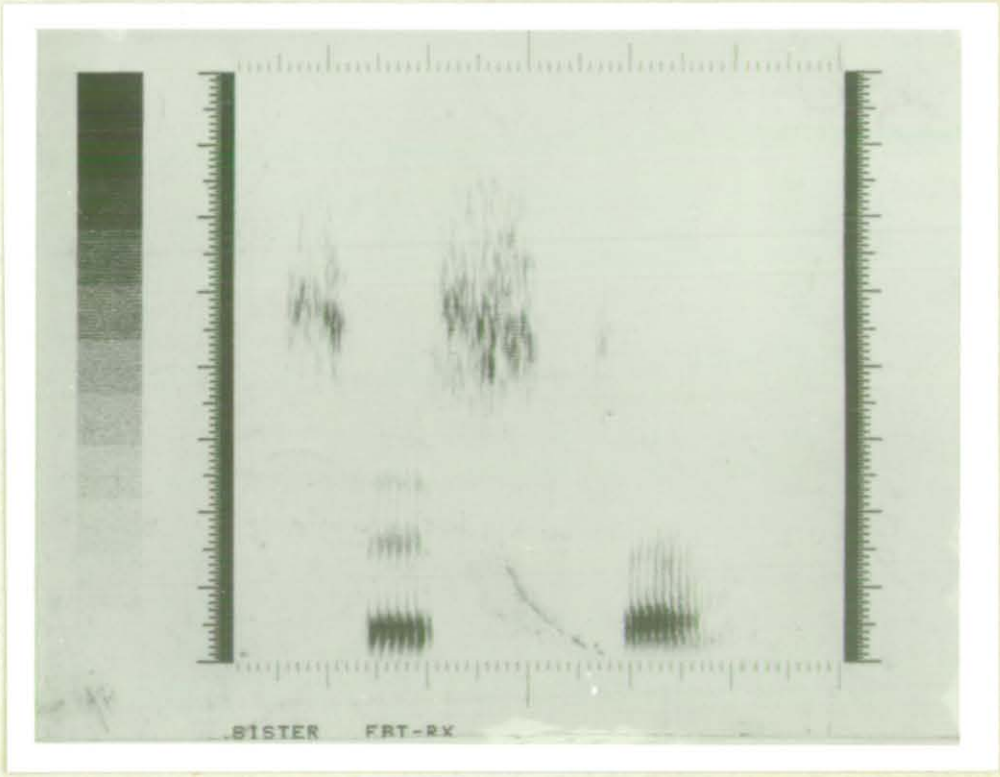


(b) Bandlimited speech signal 0.3 - 3.4 kHz

Fig. 6.6 Spectrograms for the utterance "sister". The marking depth = 40 dB



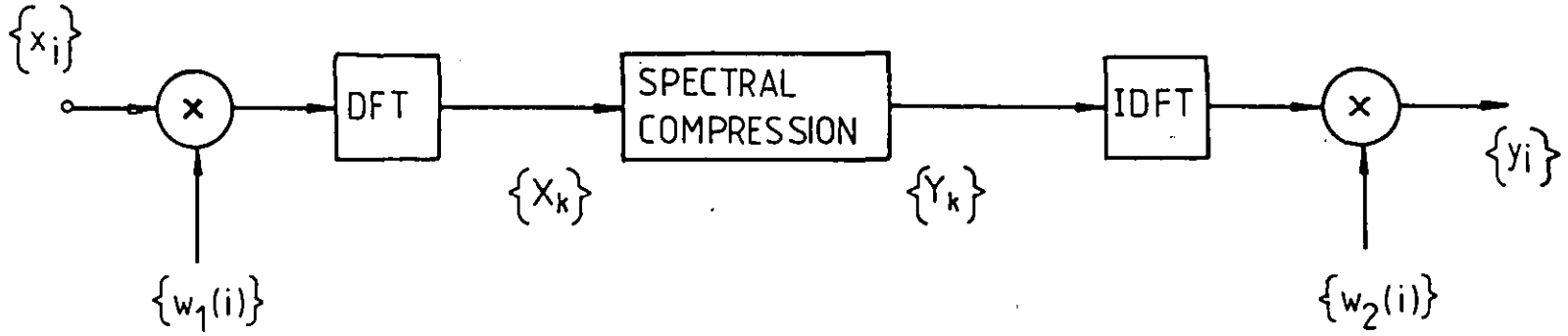
(c) Transmitted VUBS signal 0.3 - 3.4 kHz



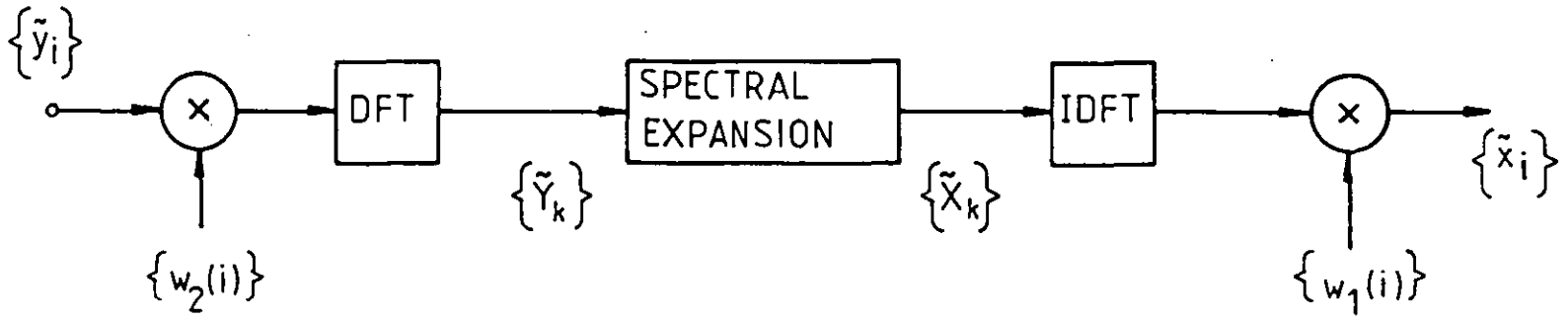
(d) Reconstructed VUBS signal 0.3 - 6.0 kHz

Fig. 6.6 (Continued)

Note:- The unvoiced transmitted signal in Fig. 6.5(c) was amplified by four times to emphasise the unvoiced signal in the spectrogram of Fig. 6.6(c). The unvoiced signal amplitude is divided by four at the receiver terminal



(a) TRANSMITTER



(b) RECEIVER

FIG. 6·7 DFT PROCESSING

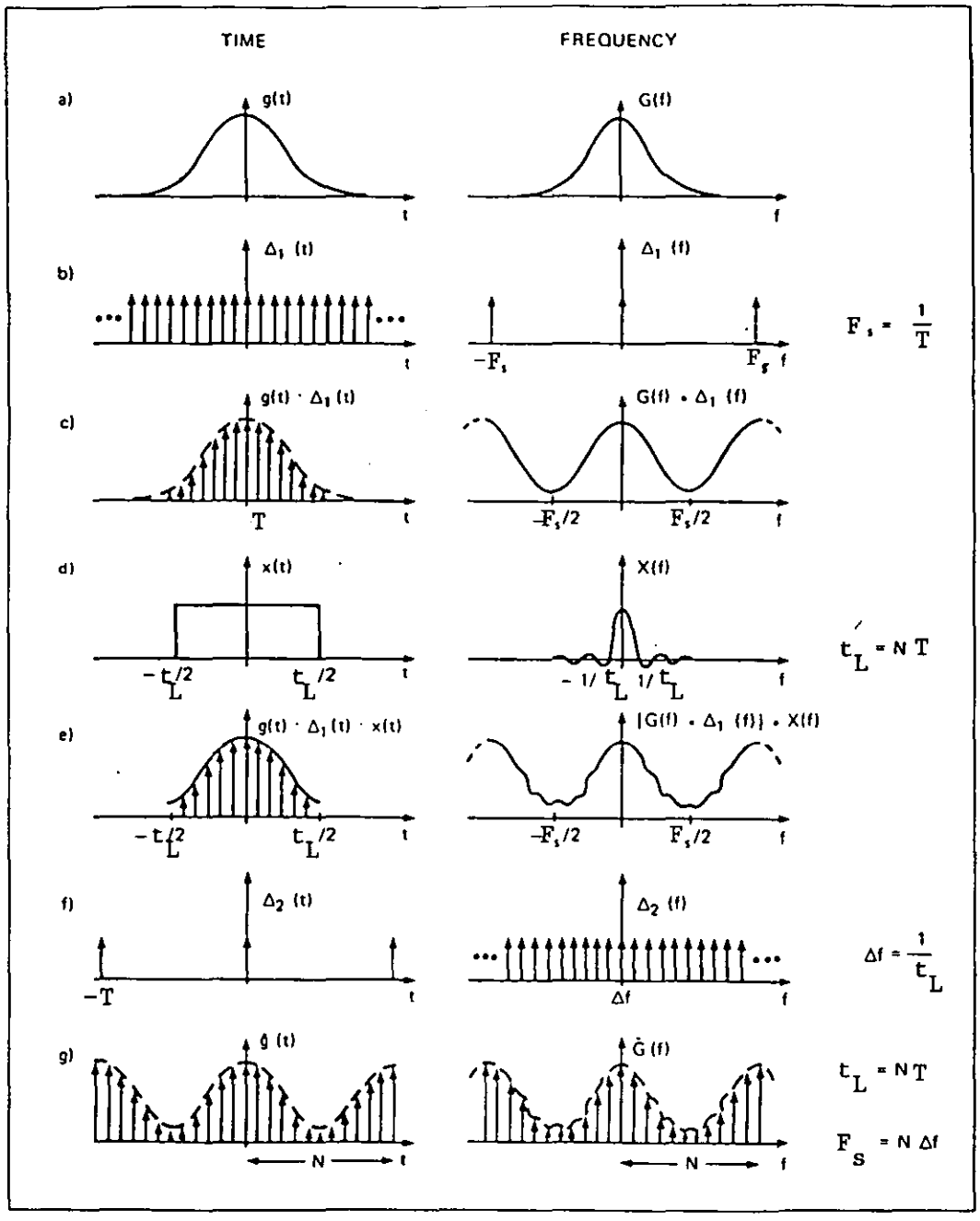
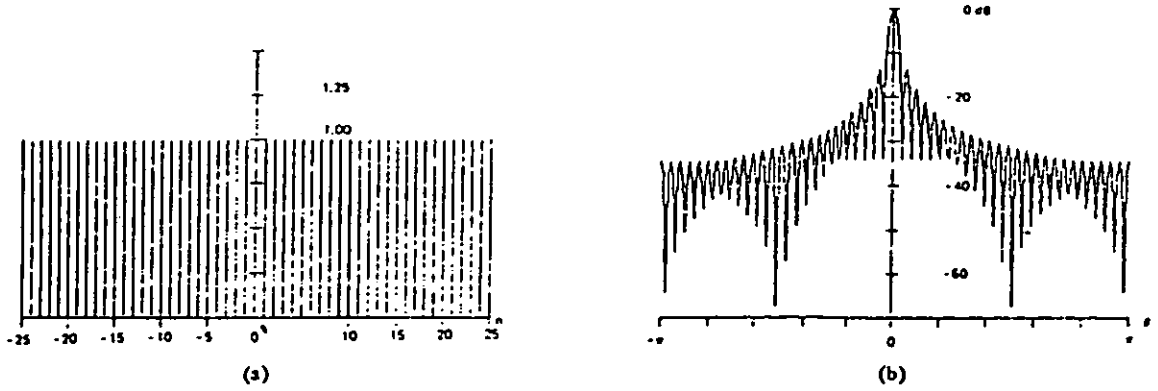


Fig 6.8 *Derivation of the Discrete Fourier Transform from the Integral Transform* (ref.83)

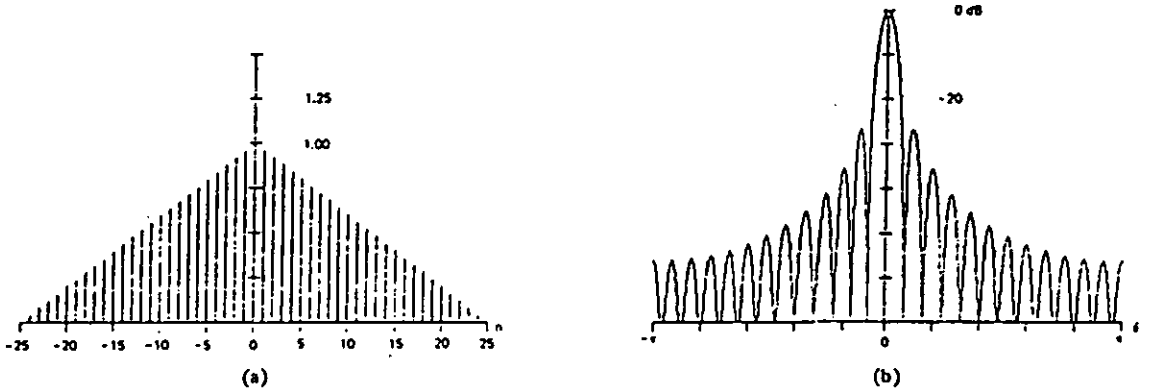
(i)



(a) (b)

(a) Rectangle window. (b) Log-magnitude of transform.

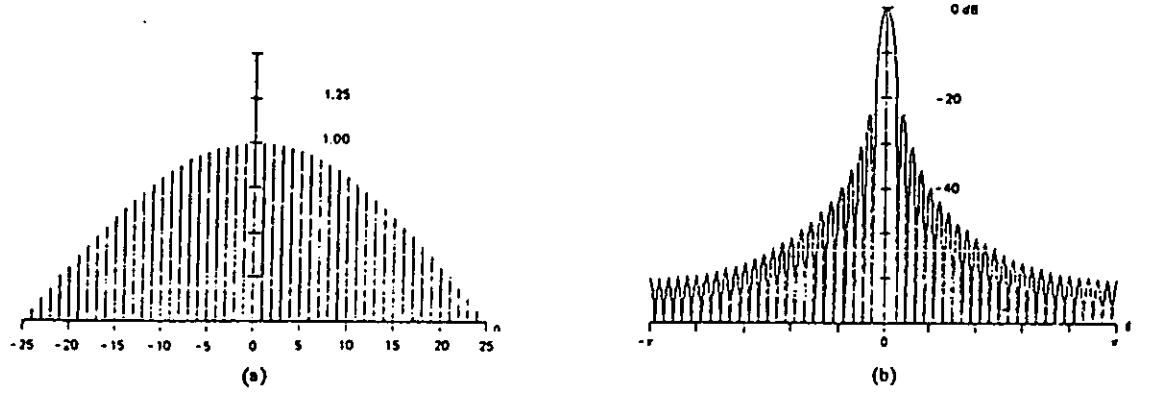
(ii)



(a) (b)

(a) Triangle window. (b) Log-magnitude of transform.

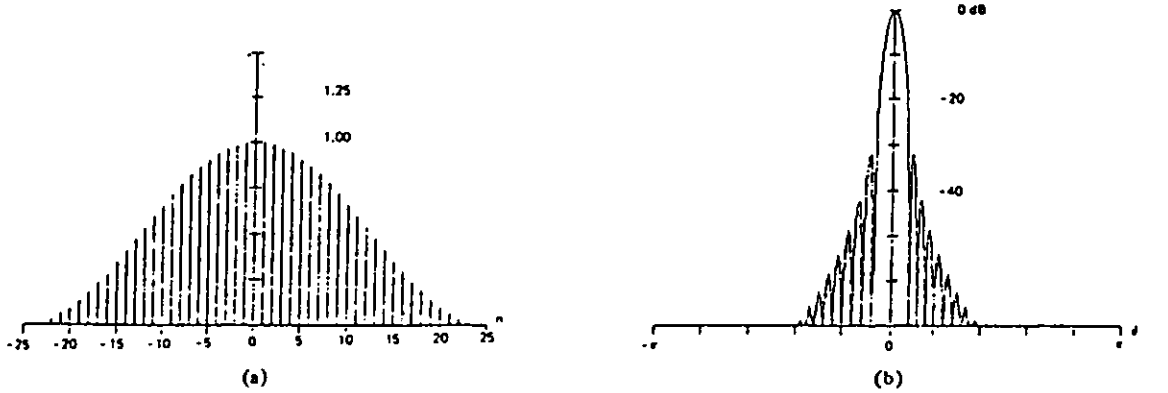
(iii)



(a) (b)

(a) Cos $(n\pi/N)$ window. (b) Log-magnitude of transform.

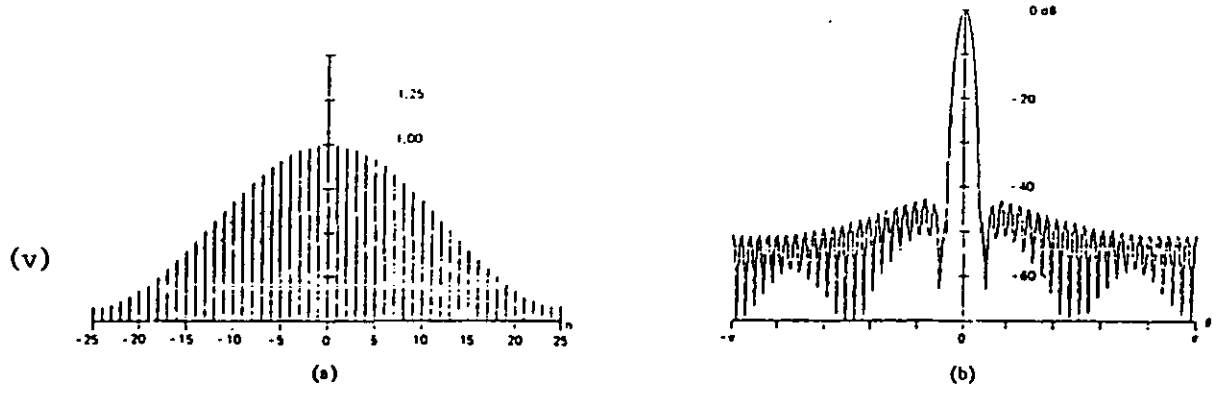
(iv)



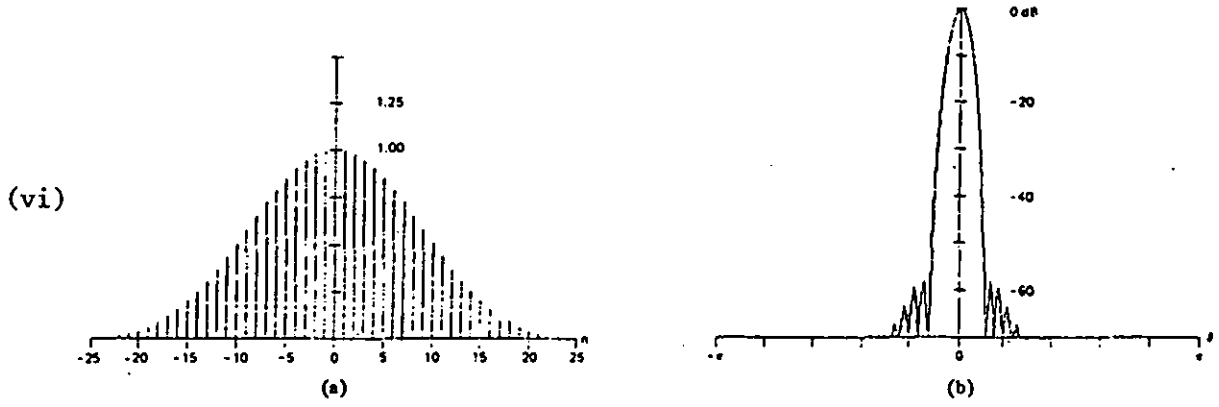
(a) (b)

(a) Cos^2 $(n\pi/N)$ window. (b) Log-magnitude of transform.

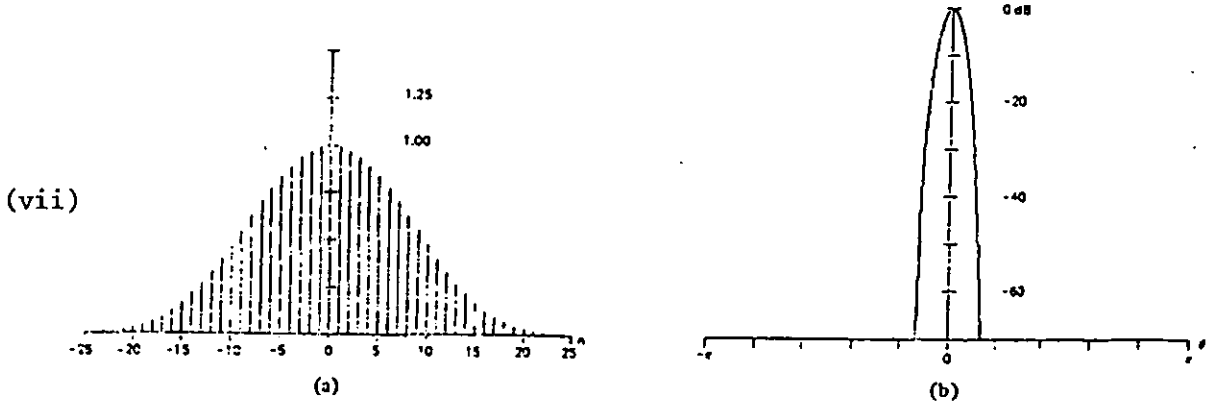
Fig 6.9 (ref.84)



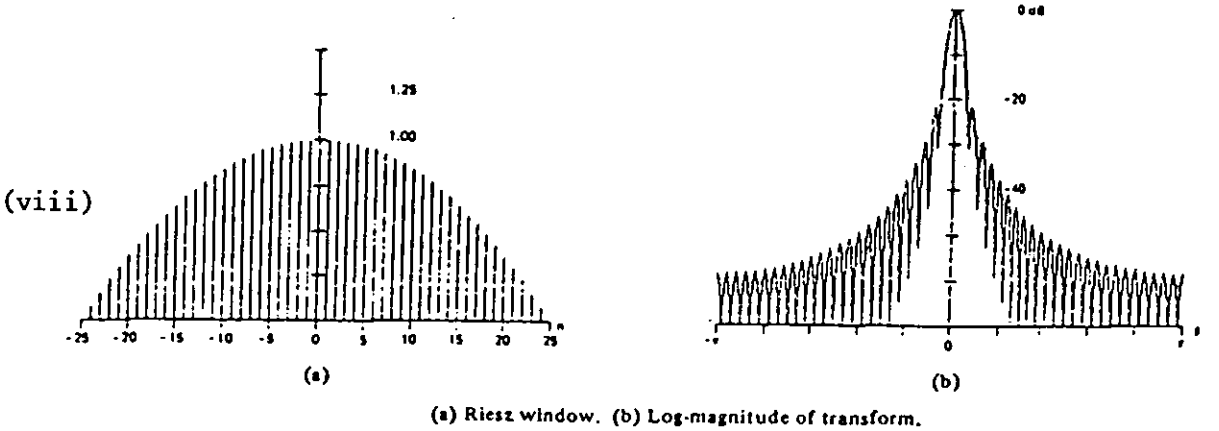
(a) Hamming window. (b) Log-magnitude of Fourier transform.



(a) Blackman window. (b) Log-magnitude of transform.

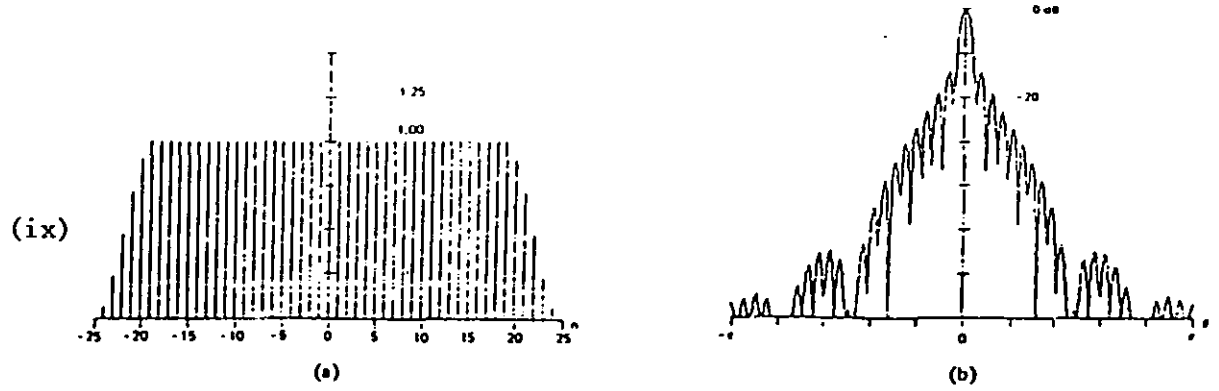


(a) 4-term Blackman-Harris window. (b) Log-magnitude of transform.

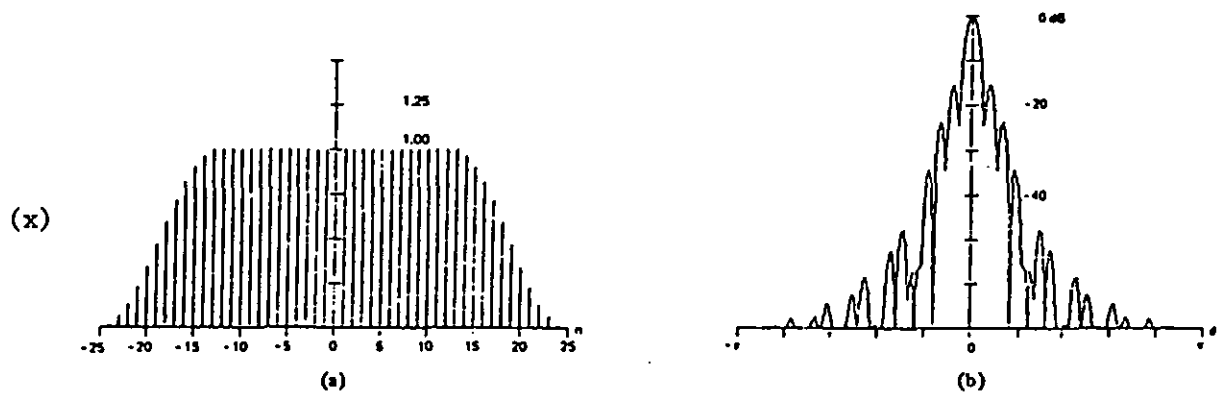


(a) Riesz window. (b) Log-magnitude of transform.

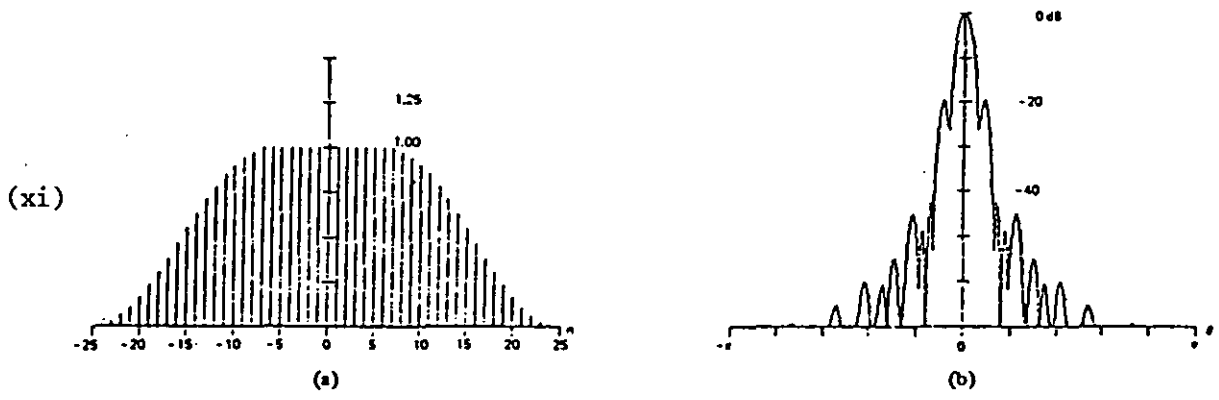
Fig 6.9 contd. (ref.84)



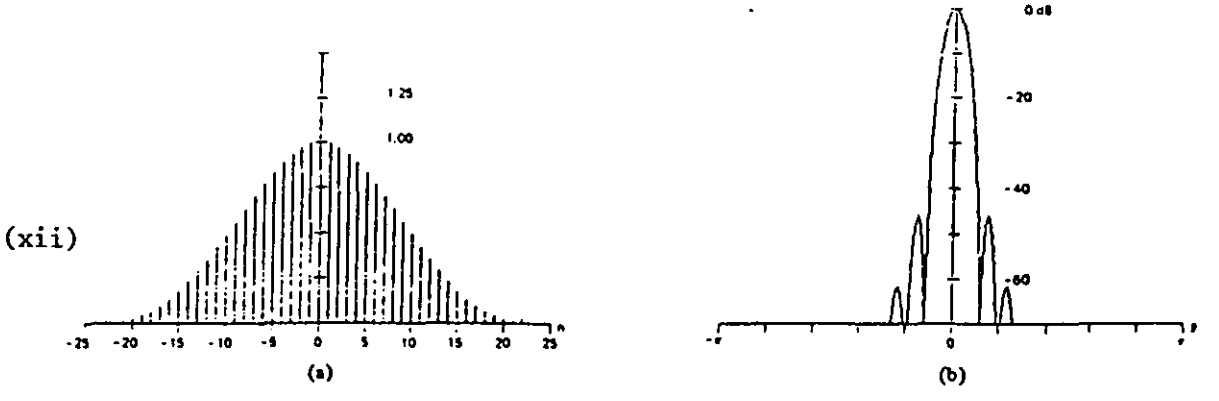
(a) 25-percent cosine taper (Tukey) window. (b) Log-magnitude of transform.



(a) 50-percent cosine taper (Tukey) window. (b) Log-magnitude of transform.



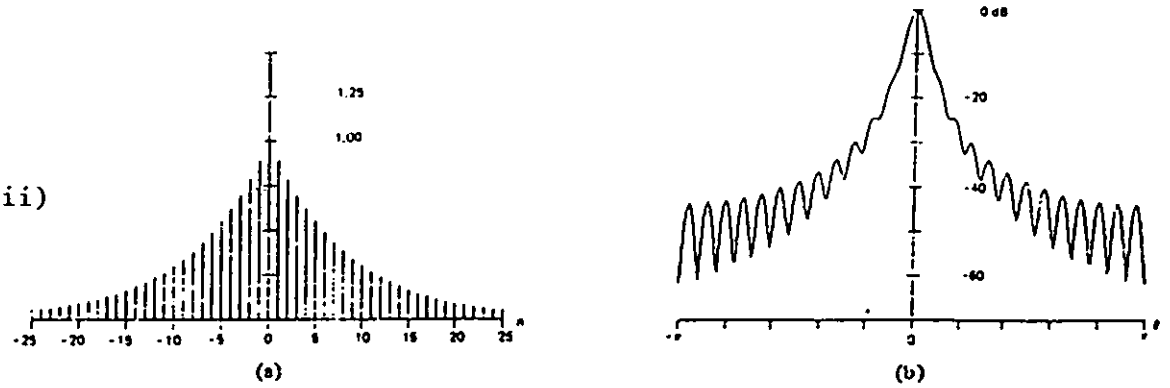
(a) 75-percent cosine taper (Tukey) window. (b) Log-magnitude of transform.



(a) Bohman window. (b) Log-magnitude of transform.

Fig 6.9 contd. (ref.84)

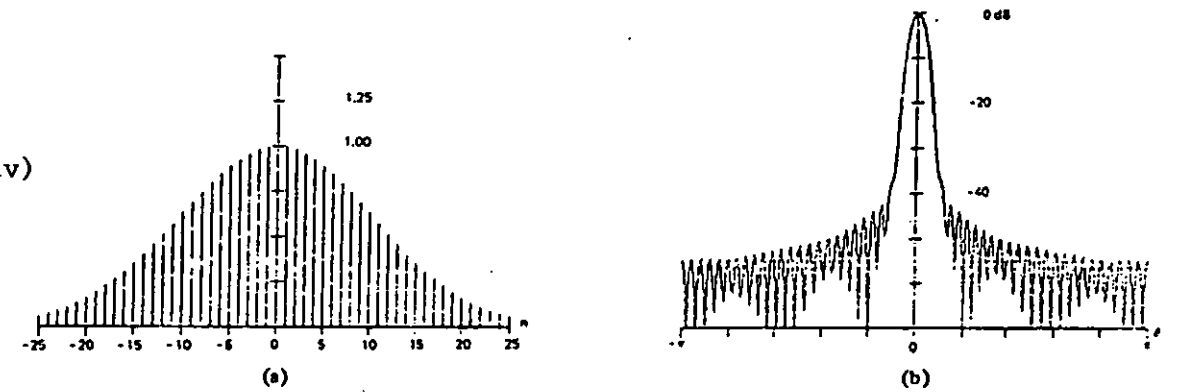
(xiii)



(a) (b)

(a) Poisson window. (b) Log-magnitude of transform (a = 3.0).

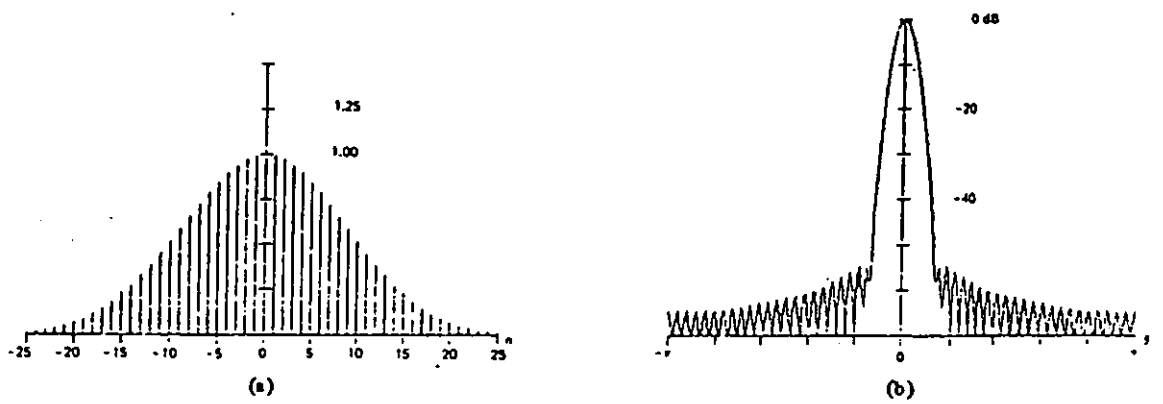
(xiv)



(a) (b)

(a) Gaussian window. (b) Log-magnitude of transform (a = 2.5).

(xv)



(a) (b)

(a) Gaussian window. (b) Log-magnitude of transform (a = 3.0).

Fig 6.9 contd. (ref.84)

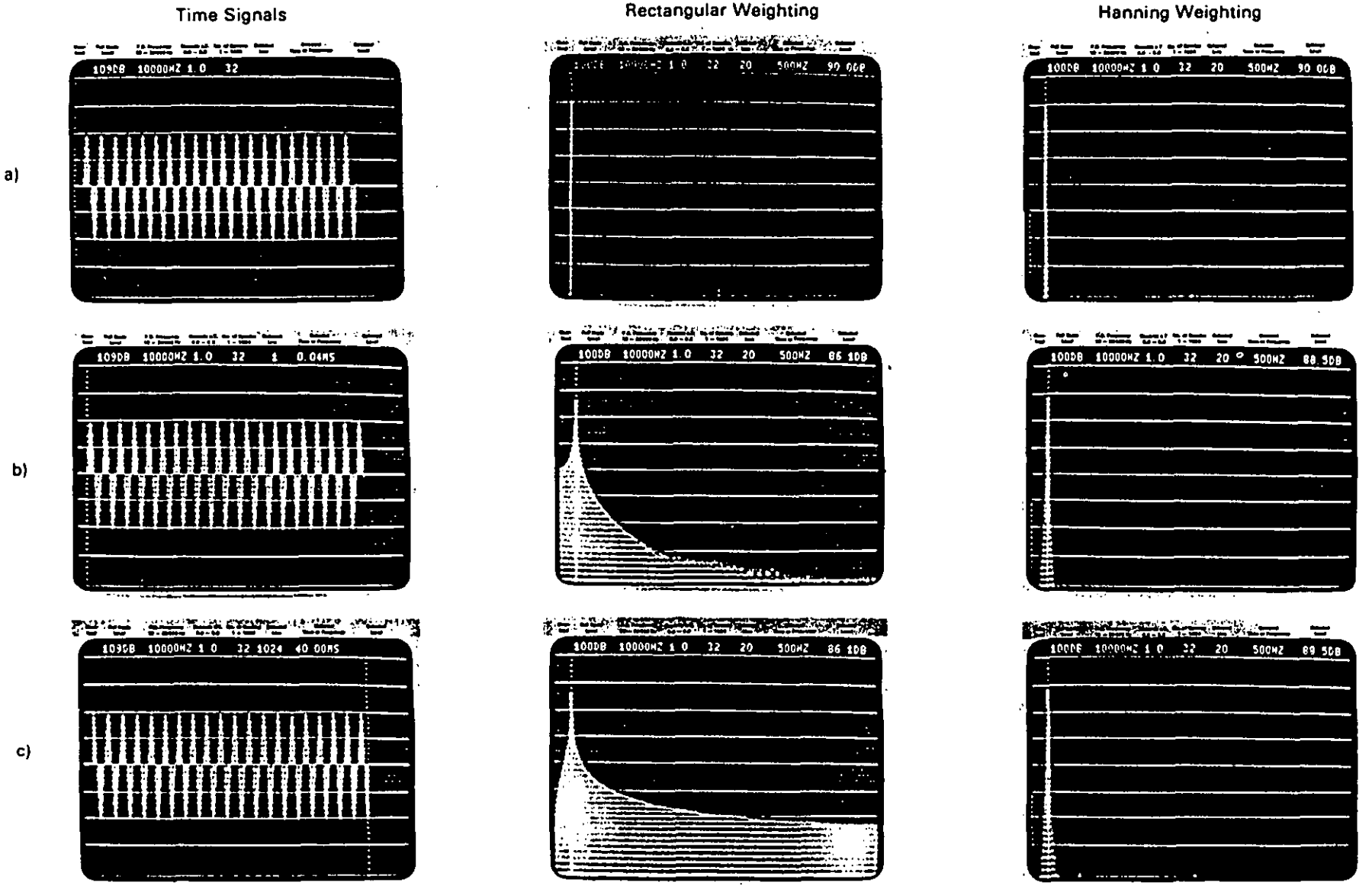
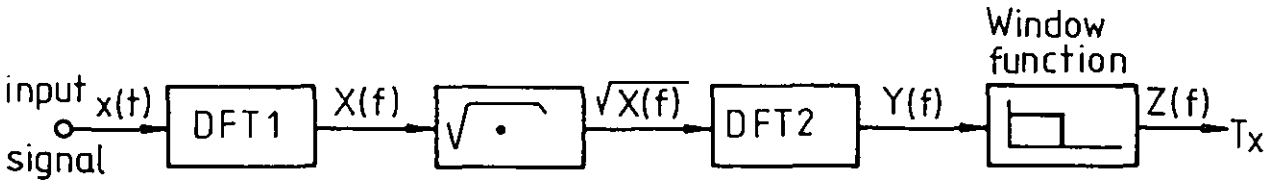
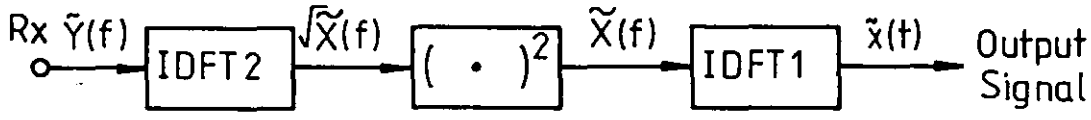


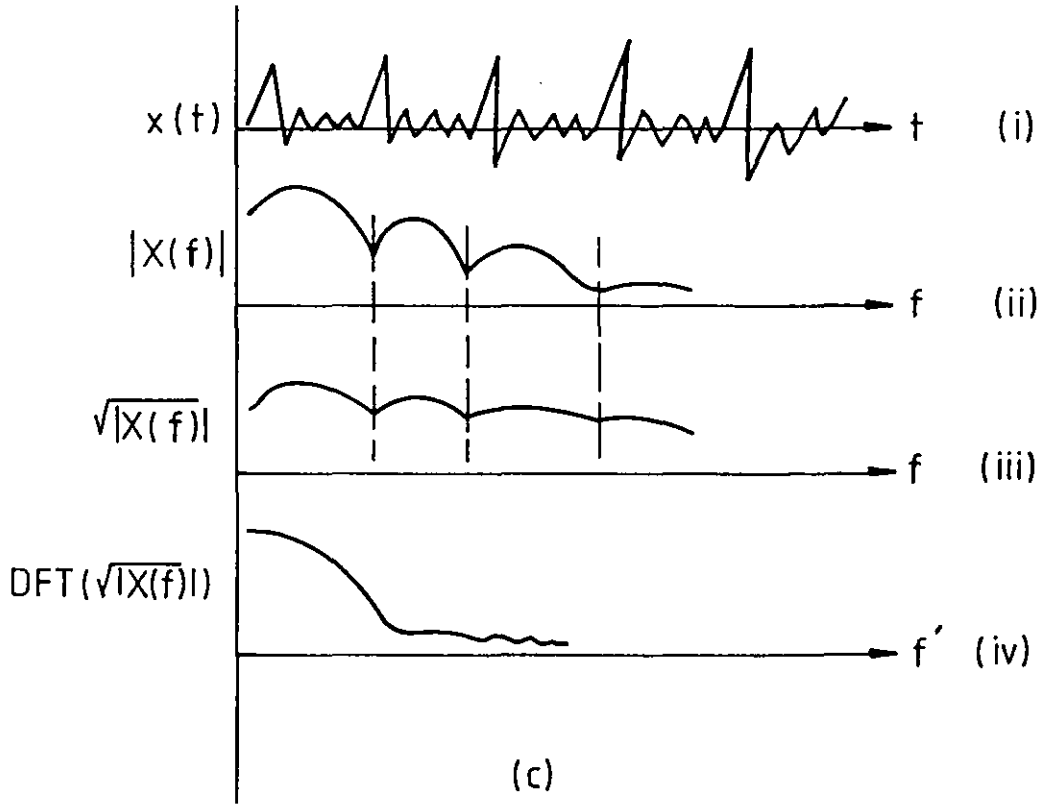
Fig 6.10 *FFT analysis of sinusoidal time signals using different window functions. Number of periods in time memory: a)integer, b) and c) half-integer but different phases* (ref.83)


input signal
x(t)
DFT1
X(f)
√ •
√X(f)
DFT2
Y(f)
Window function
Z(f)
Tx

(a)


Rx Ỹ(f)
IDFT2
√X̃(f)
(•)²
X̃(f)
IDFT1
x̃(t)
Output Signal

(b)


x(t)
t
(i)
|X(f)|
f
(ii)
√|X(f)|
f
(iii)
DFT(√|X(f)|)
f'
(iv)

(c)

FIG. 6·11

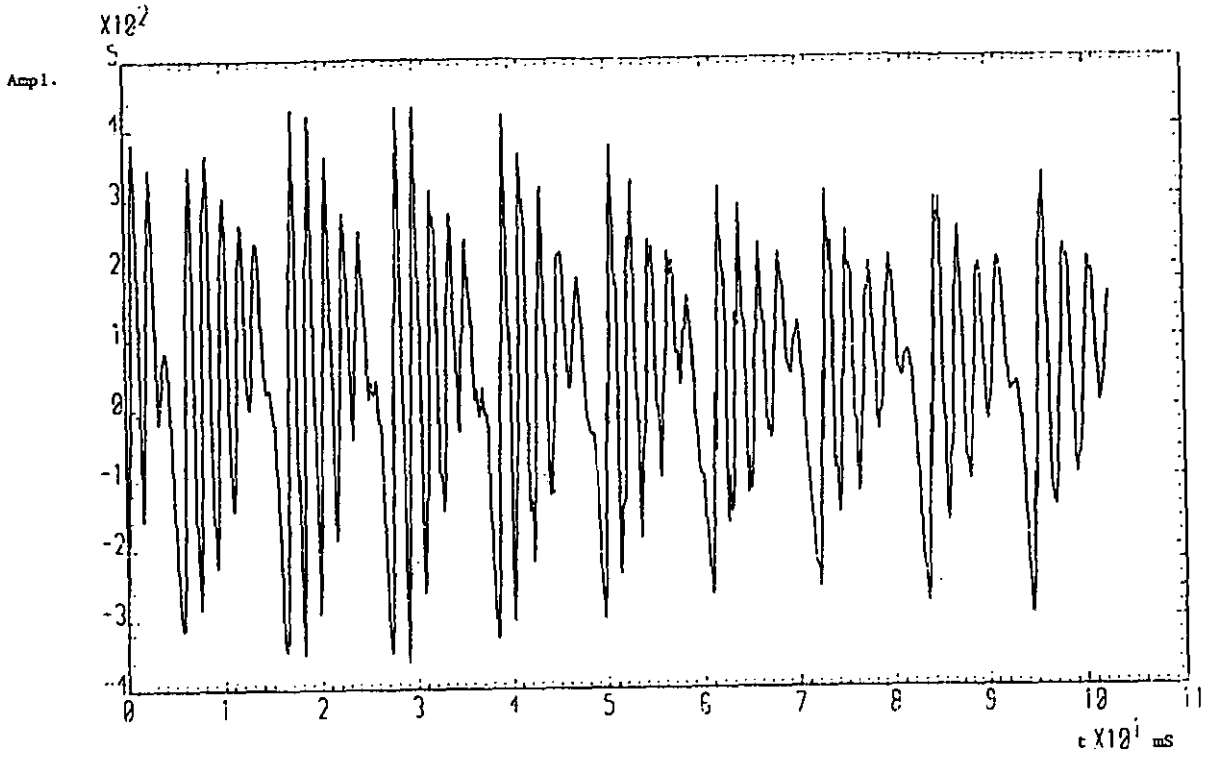


Fig 6.12(a) Time waveform of segment 1

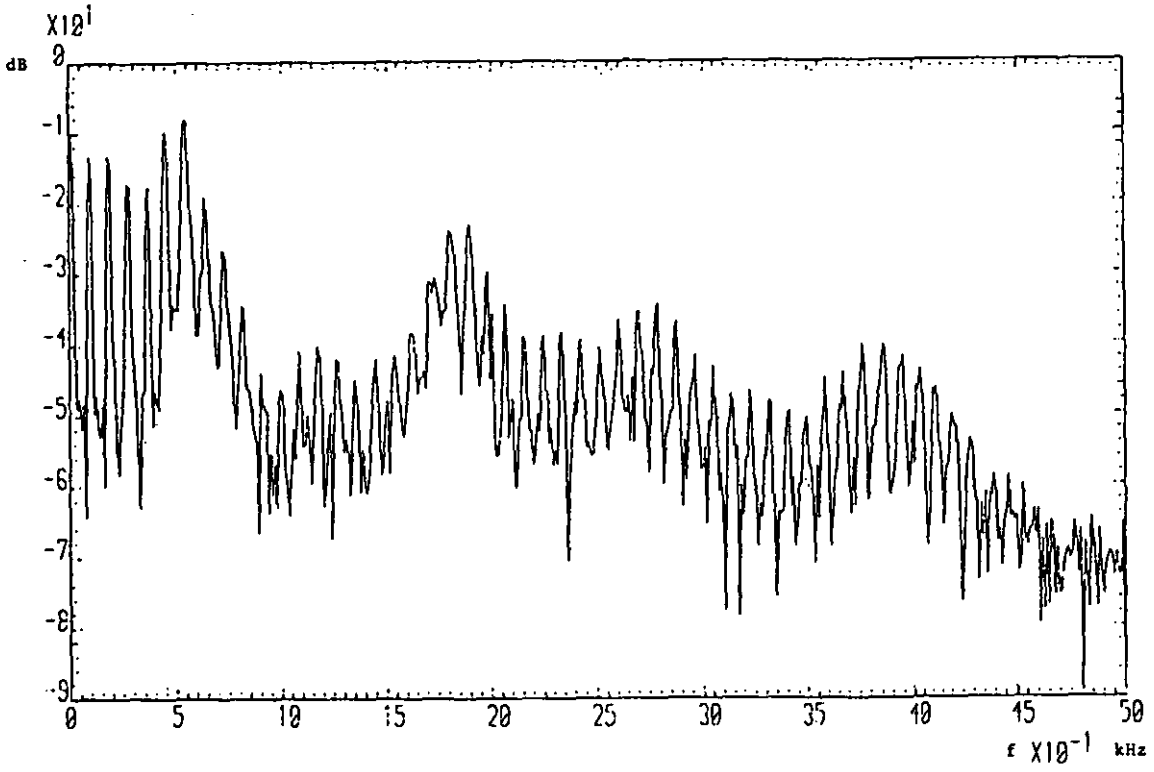


Fig 6.12(b) Spectrum of segment 1.

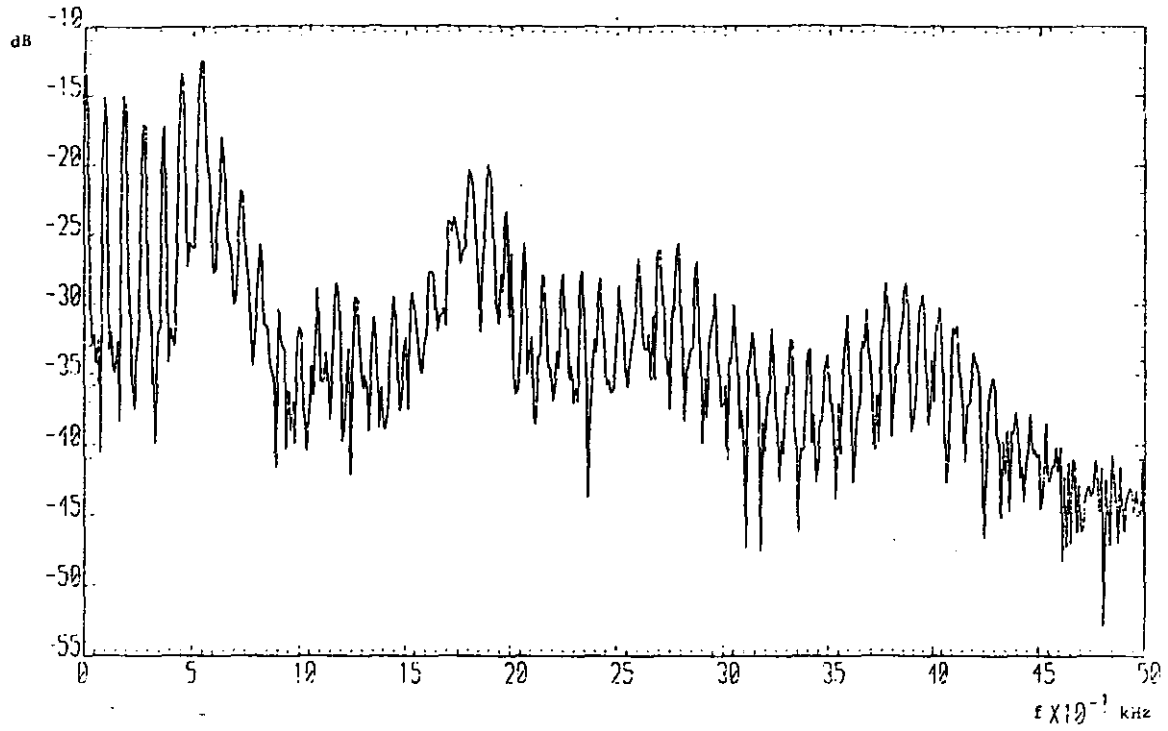


Fig 6.12(c) Square root (spectrum) of segment 1.

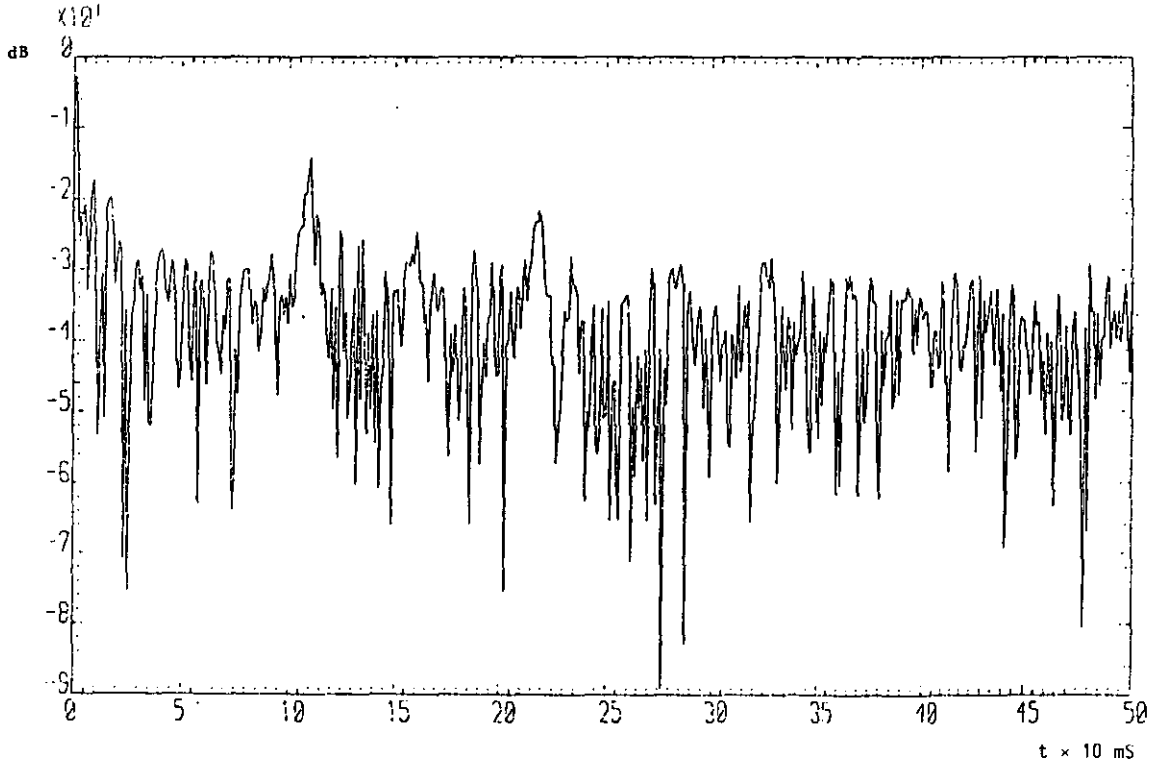


Fig 6.12(d) Second order spectrum of segment 1.

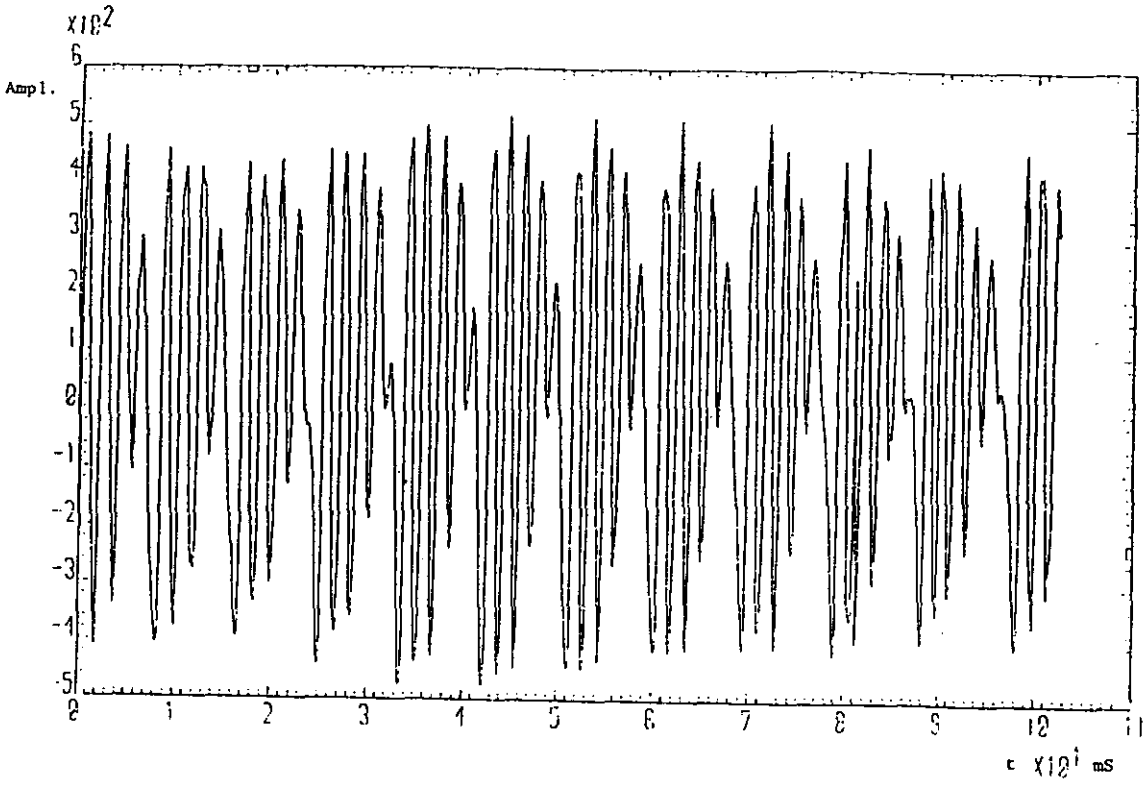


Fig 6.13(a) Time waveform of segment 2.

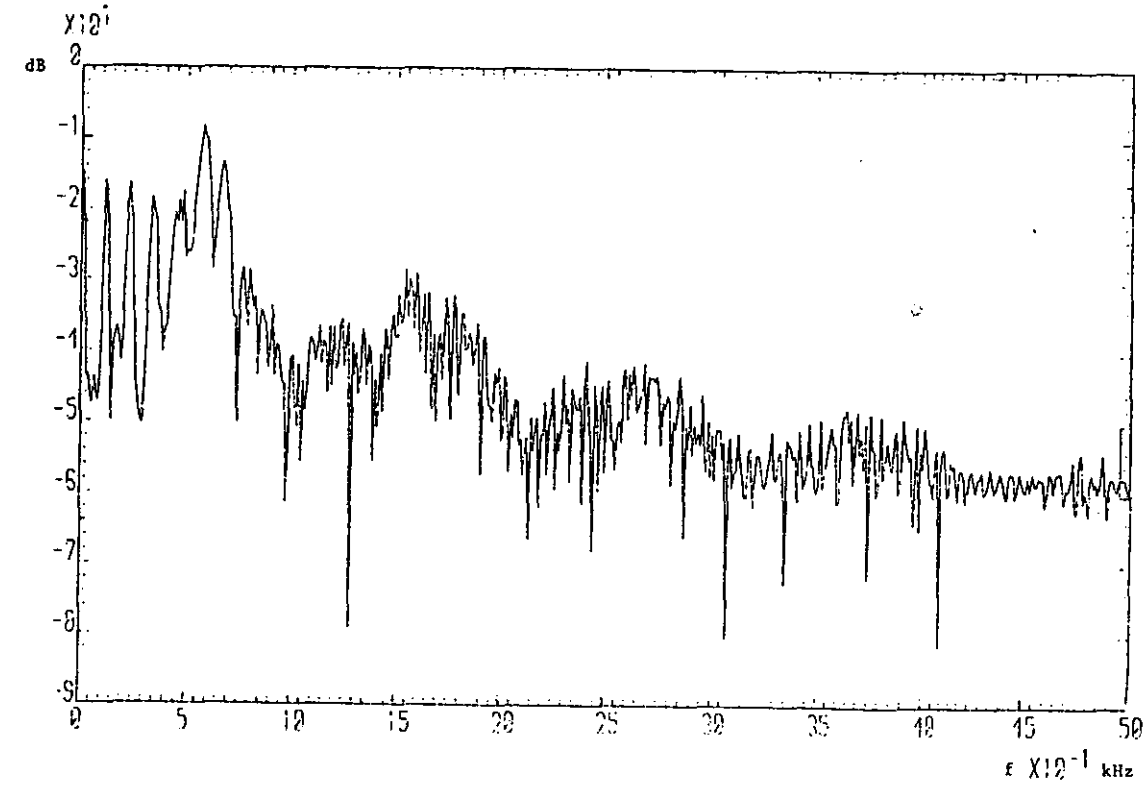


Fig 6.13(b) Spectrum of segment 2.

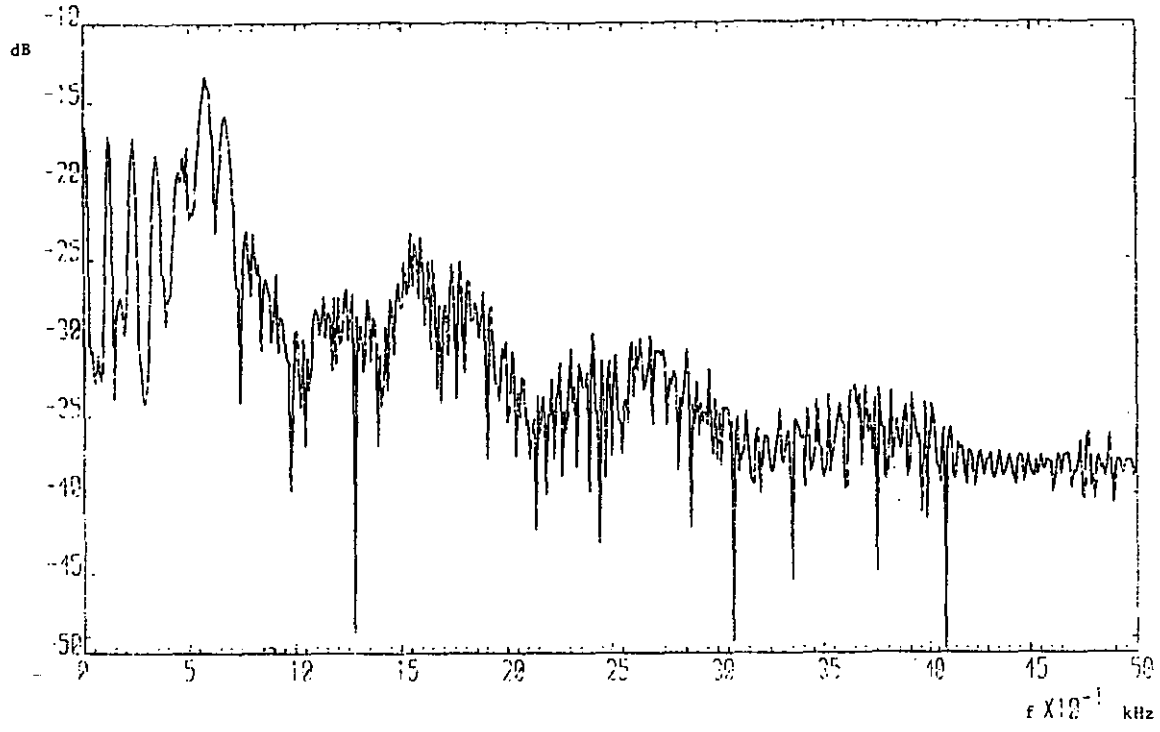


Fig 6.13(c) Square root (spectrum) of segment 2.

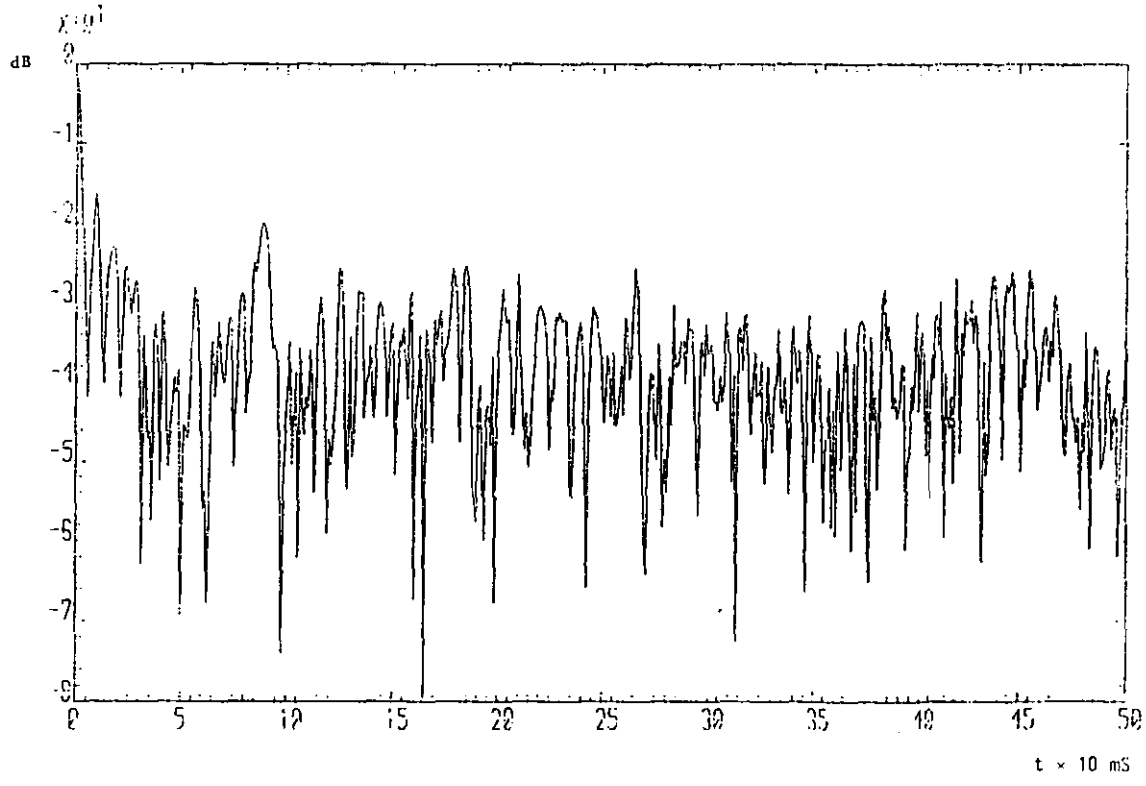


Fig 6.13(d) Second order spectrum of segment 2.

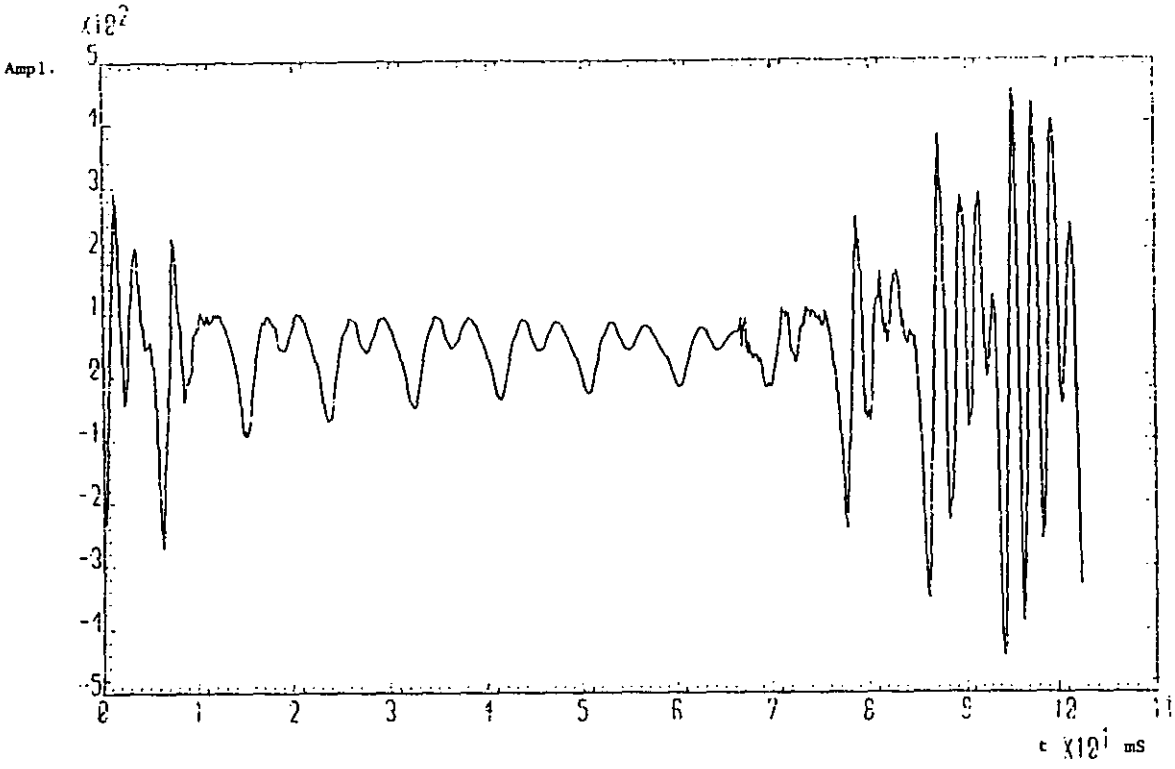


Fig 6.14(a) Time waveform of segment 3.

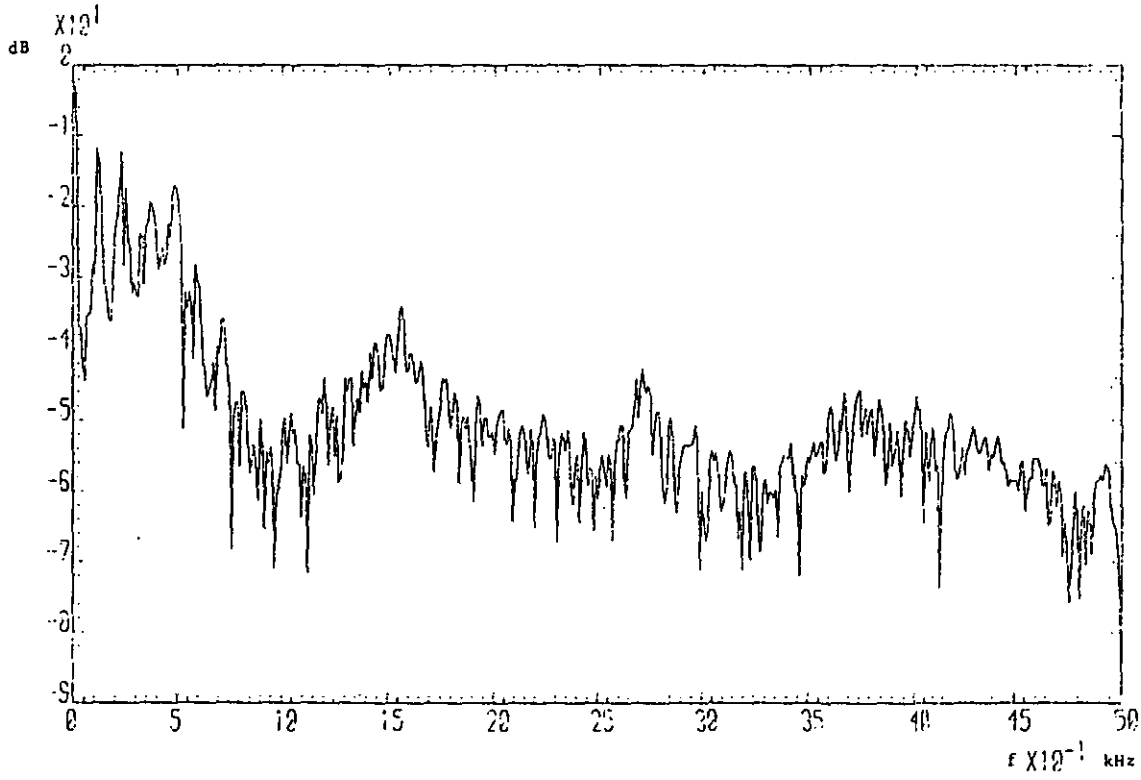


Fig 6.14(b) Spectrum of segment 3.

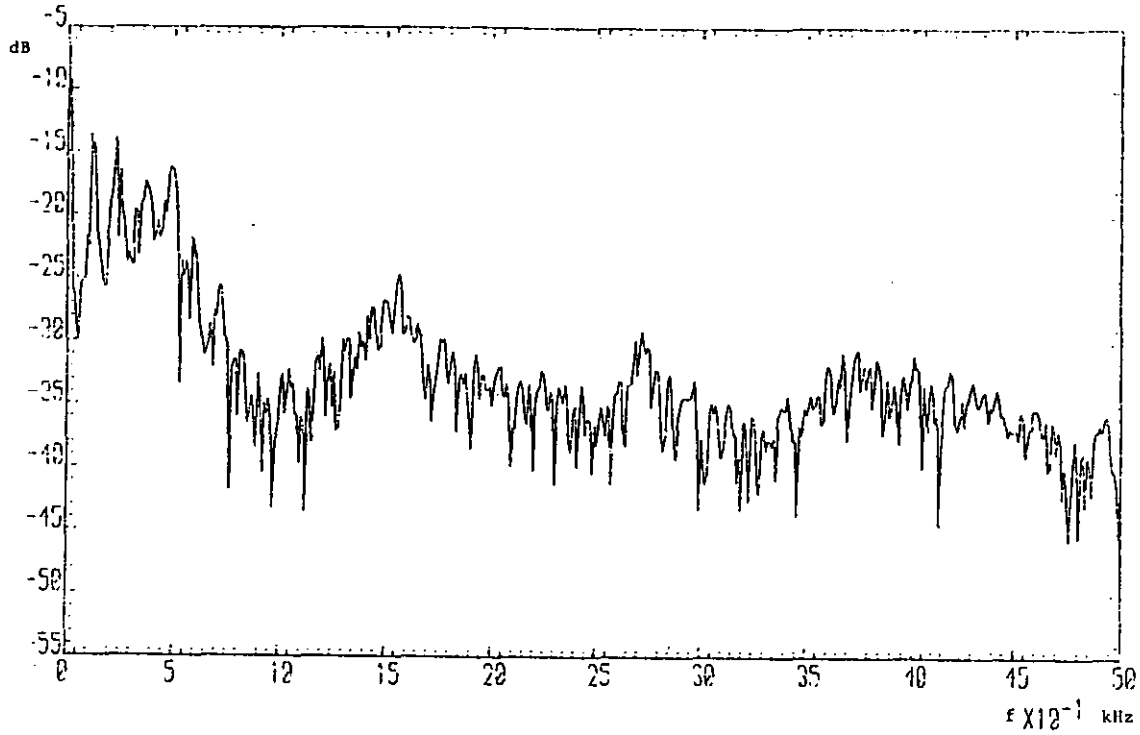


Fig 6.14(c) Square root (spectrum) of segment 3.

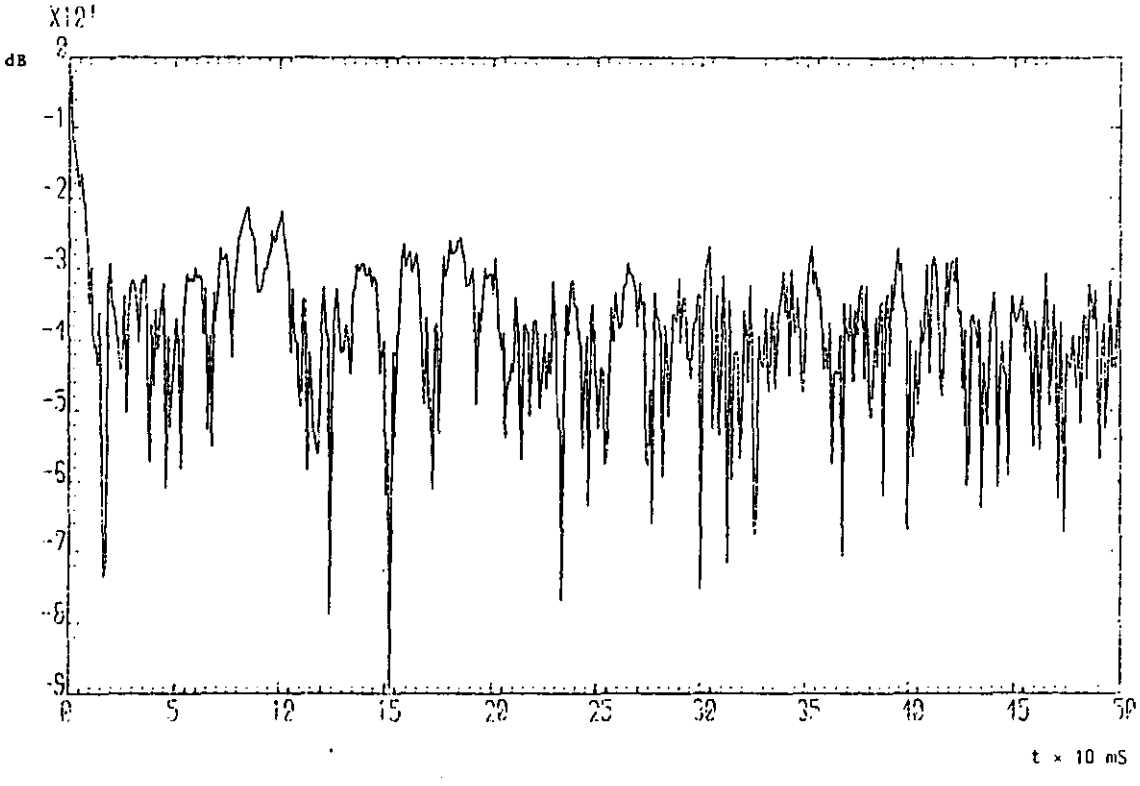


Fig 6.14(d) Second order spectrum of segment 3.

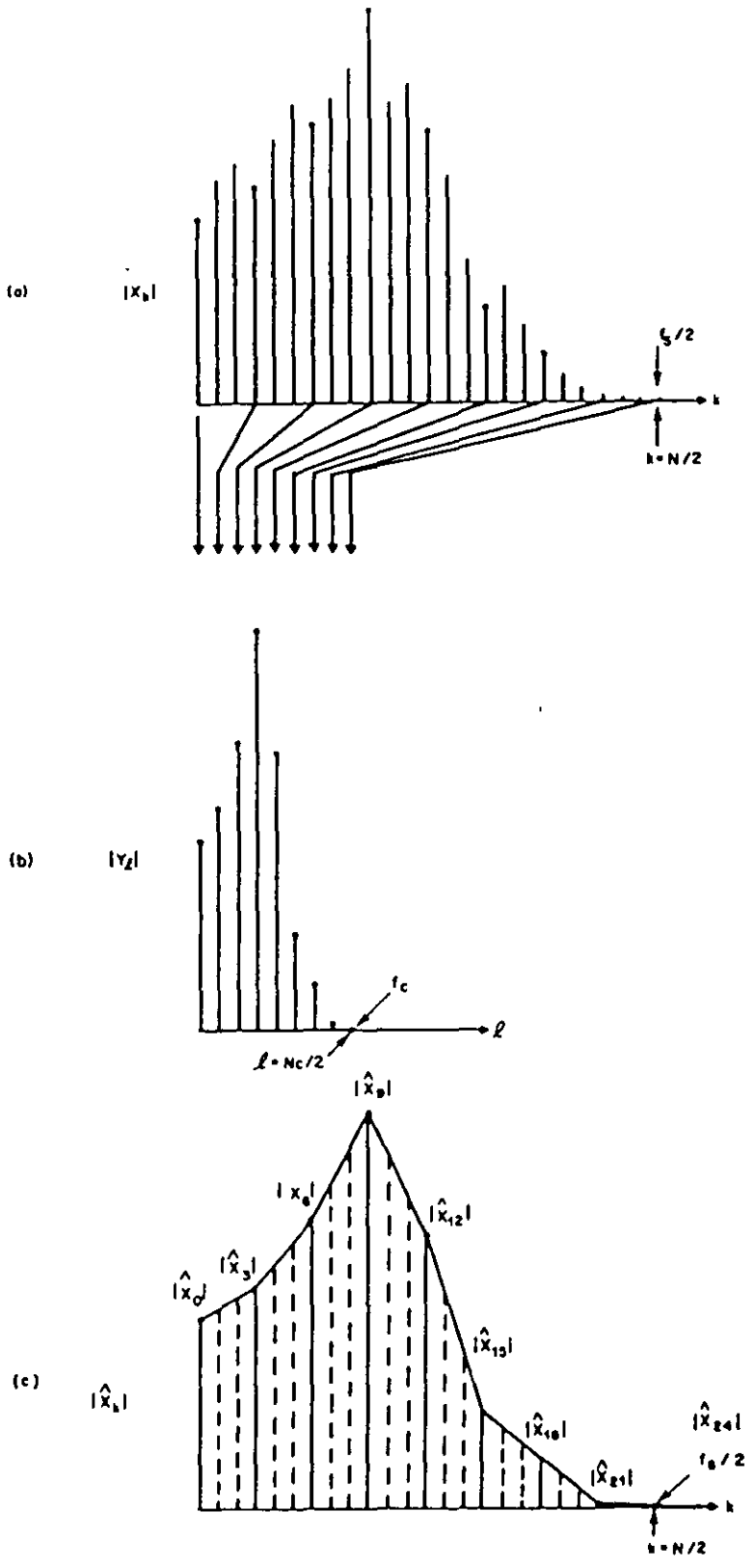


Fig 6.15 Frequency mapping the magnitude of the spectral components of an arbitrary signal.

(a) Components of the input speech signal

(b) Mapped components

(c) Frequency de-mapped components.

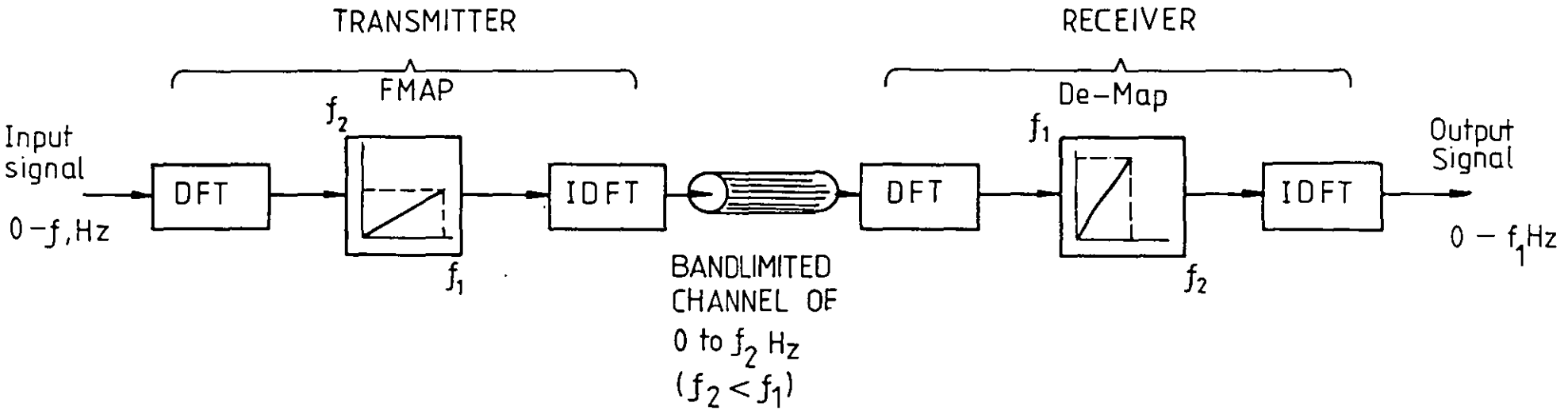

TRANSMITTER
RECEIVER
FMAP
De-Map
Input signal
0 – f, Hz
DFT
f2
f1
IDFT
BANDLIMITED CHANNEL OF
0 to f2 Hz
(f2 < f1)
DFT
f1
f2
IDFT
Output Signal
0 – f1 Hz

FIG. 6·16

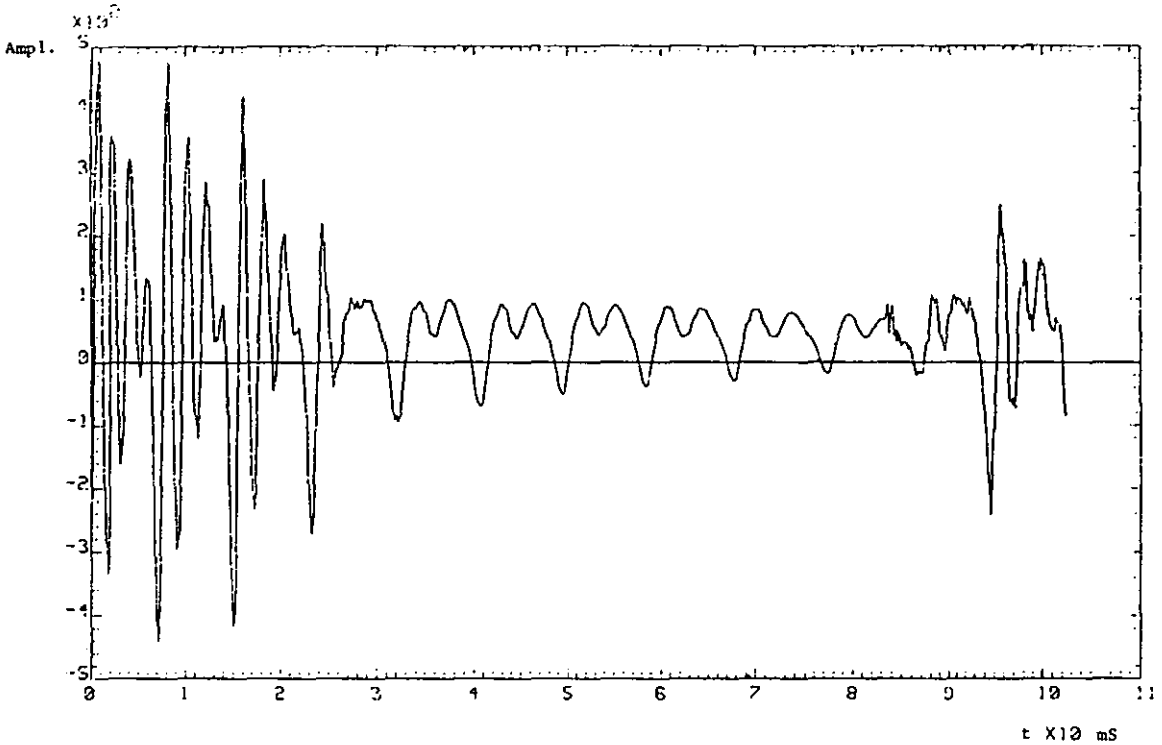


Fig 6.17(a) Input time waveform.

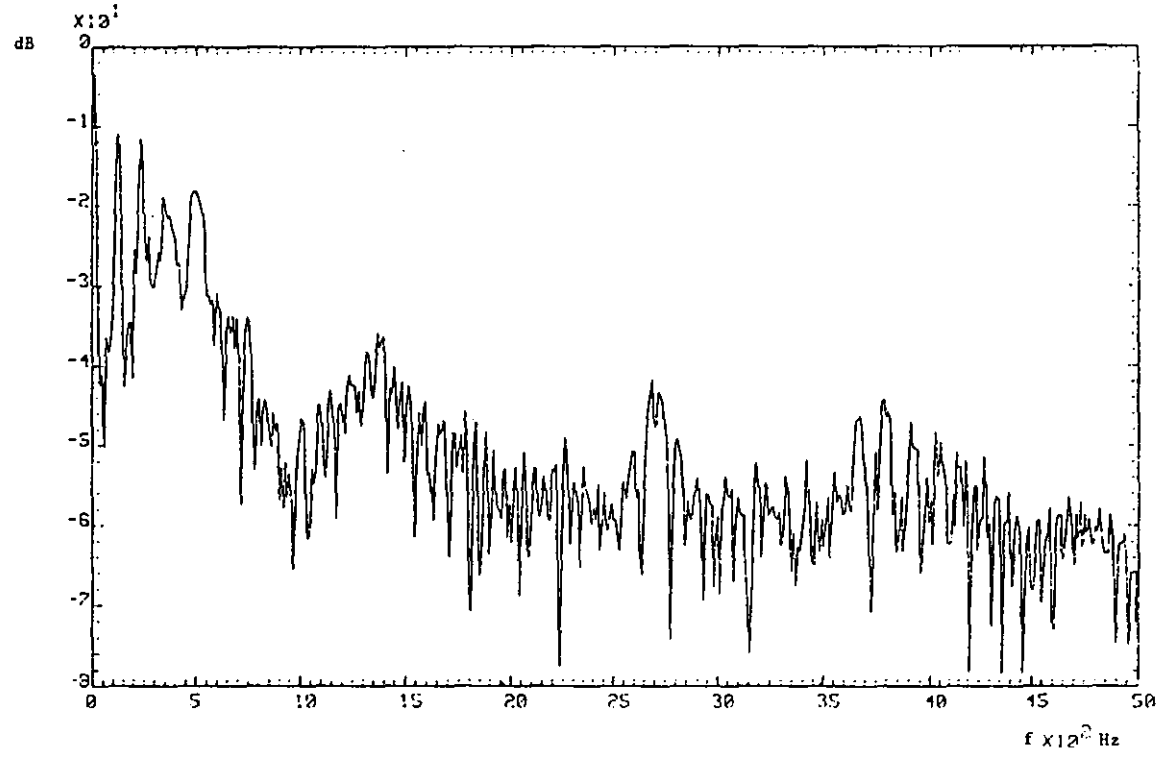


Fig 6.17(b) Input Spectrum.

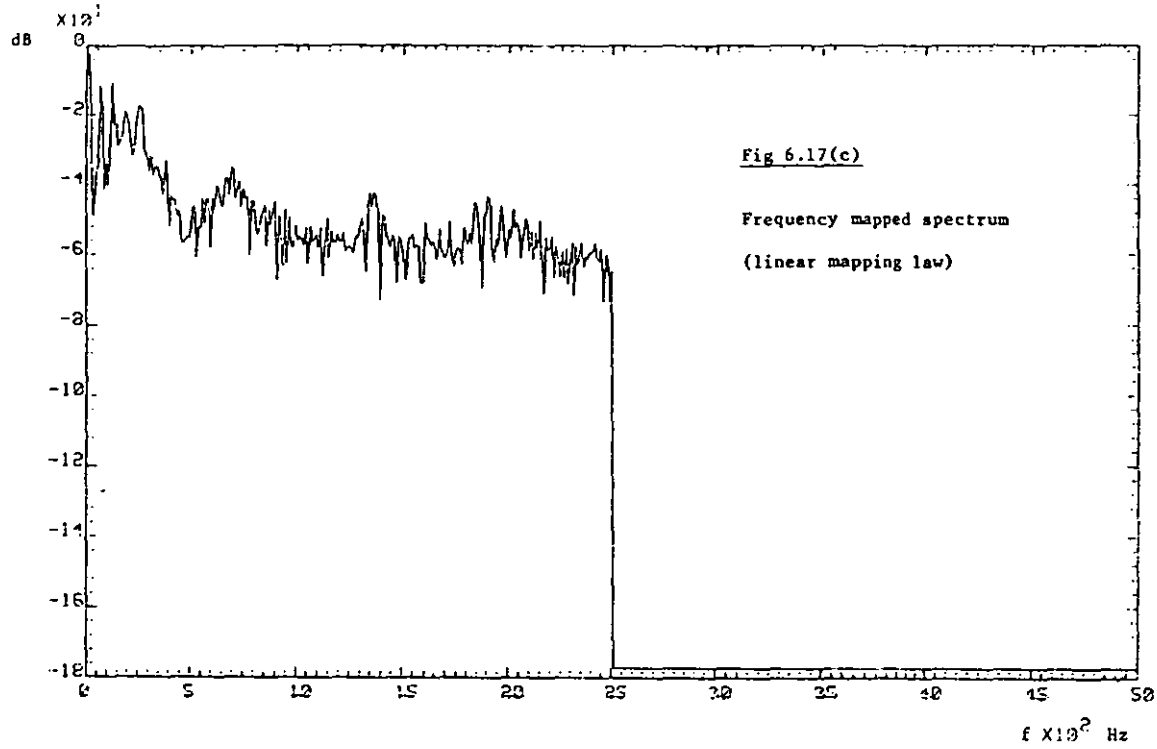


Fig 6.17(c)

Frequency mapped spectrum

(linear mapping law)

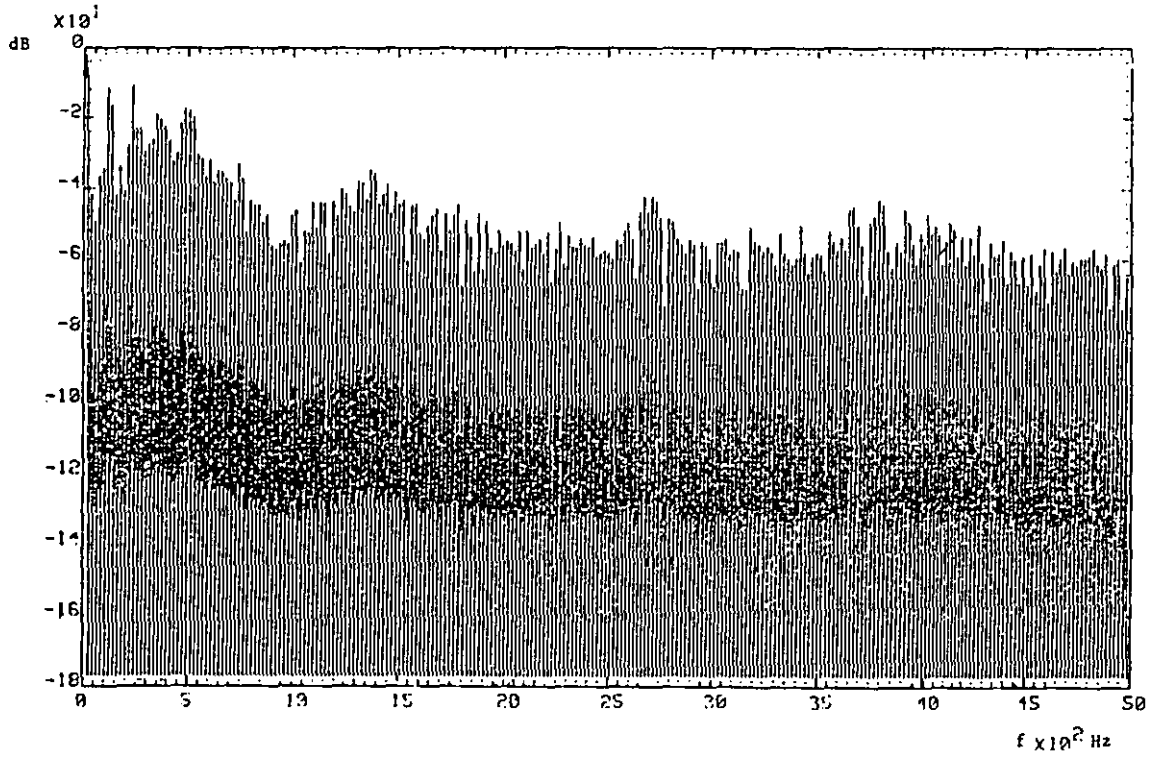


Fig 6.17(d) Inverse mapped spectrum.

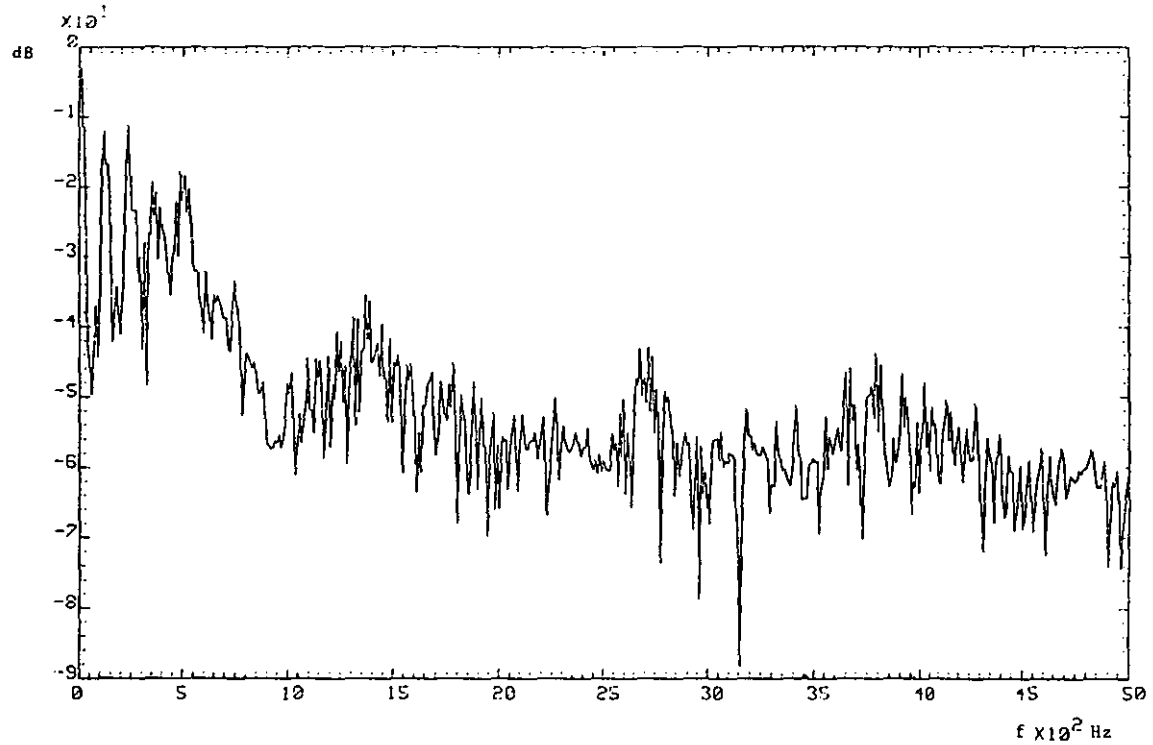


Fig 6.17(e) Interpolated spectrum.

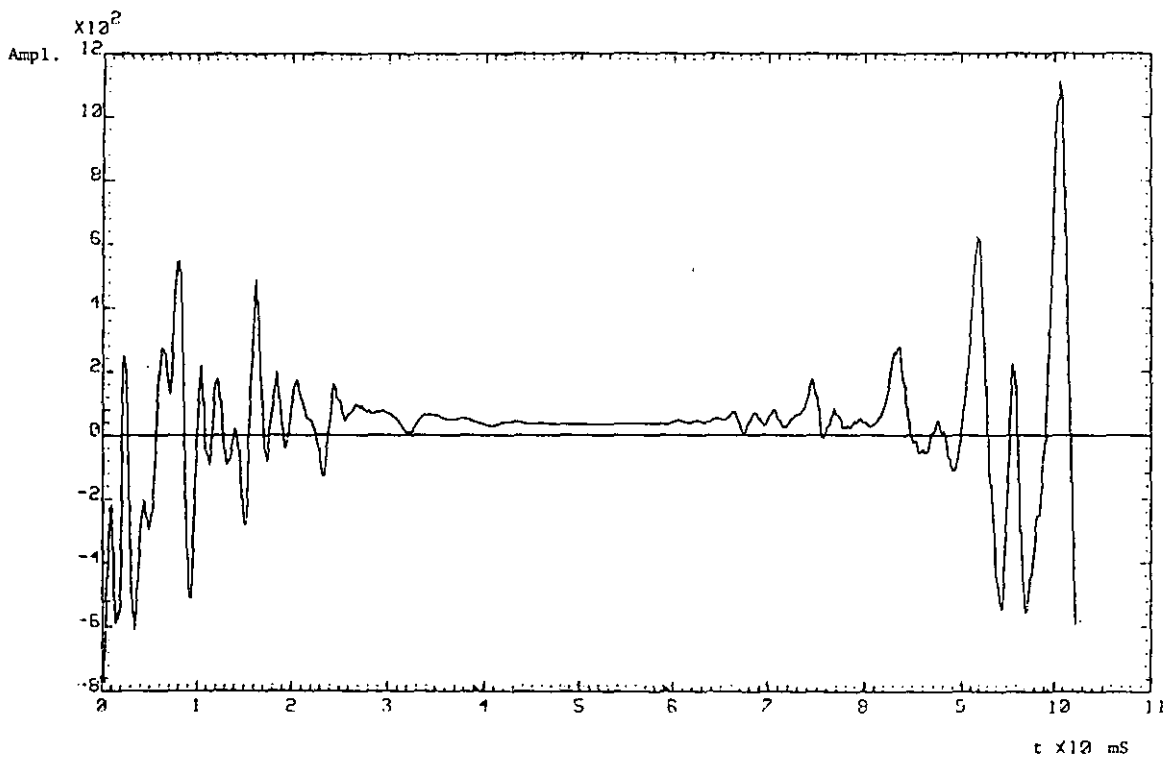


Fig 6.17(f) Output time waveform.

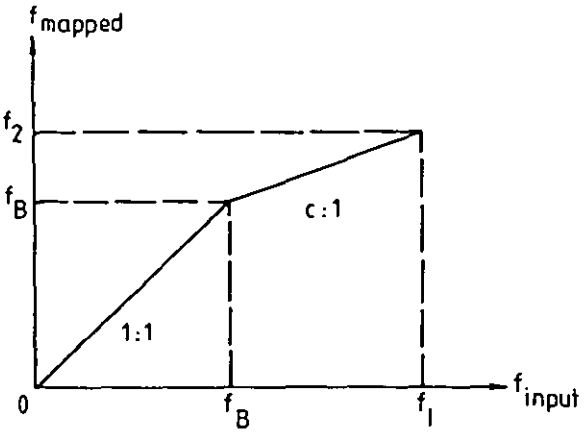


FIG. 6·18 PIECEWISE LINEAR MAPPING CHARACTERISTICS

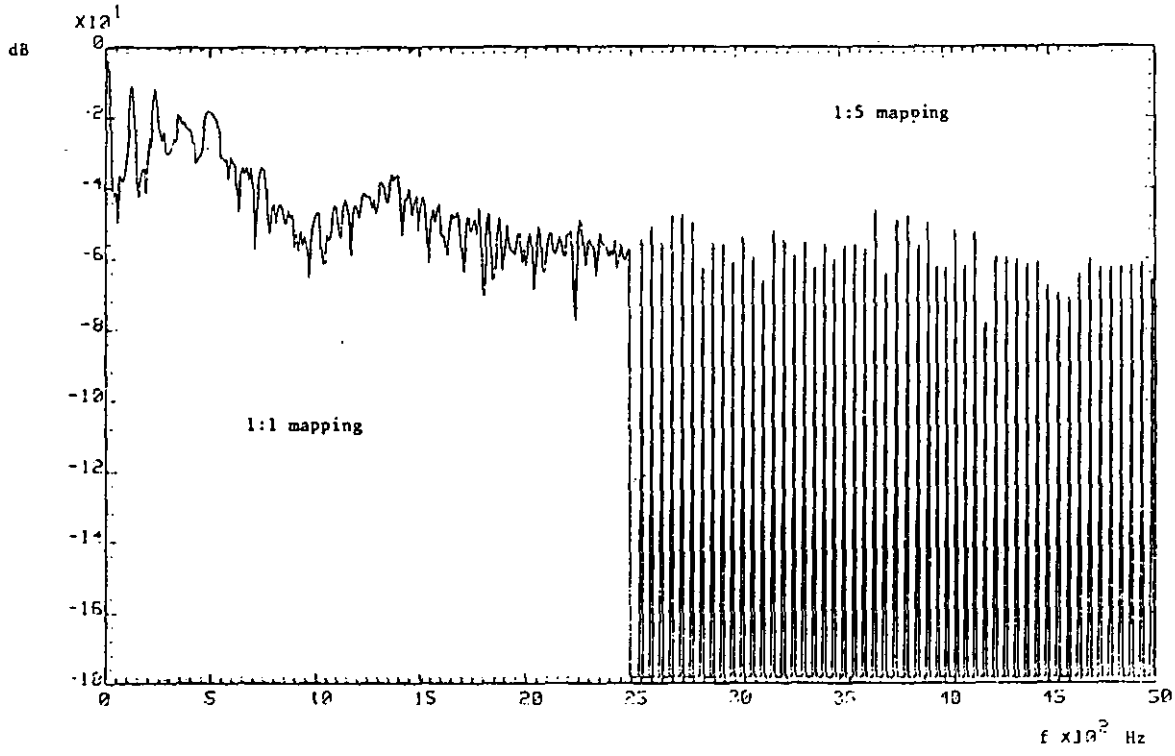


Fig 6.19(a) Inverse mapped spectrum
(piecewise linear law)

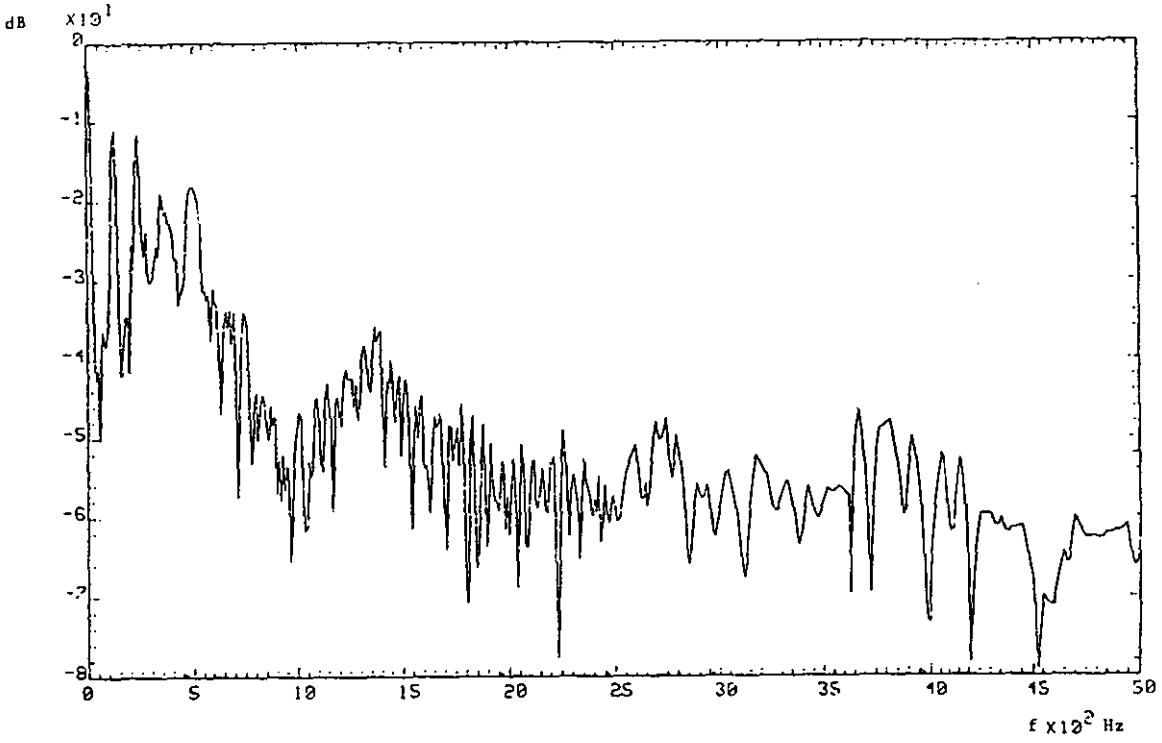


Fig 6.19(b) Interpolated spectrum.

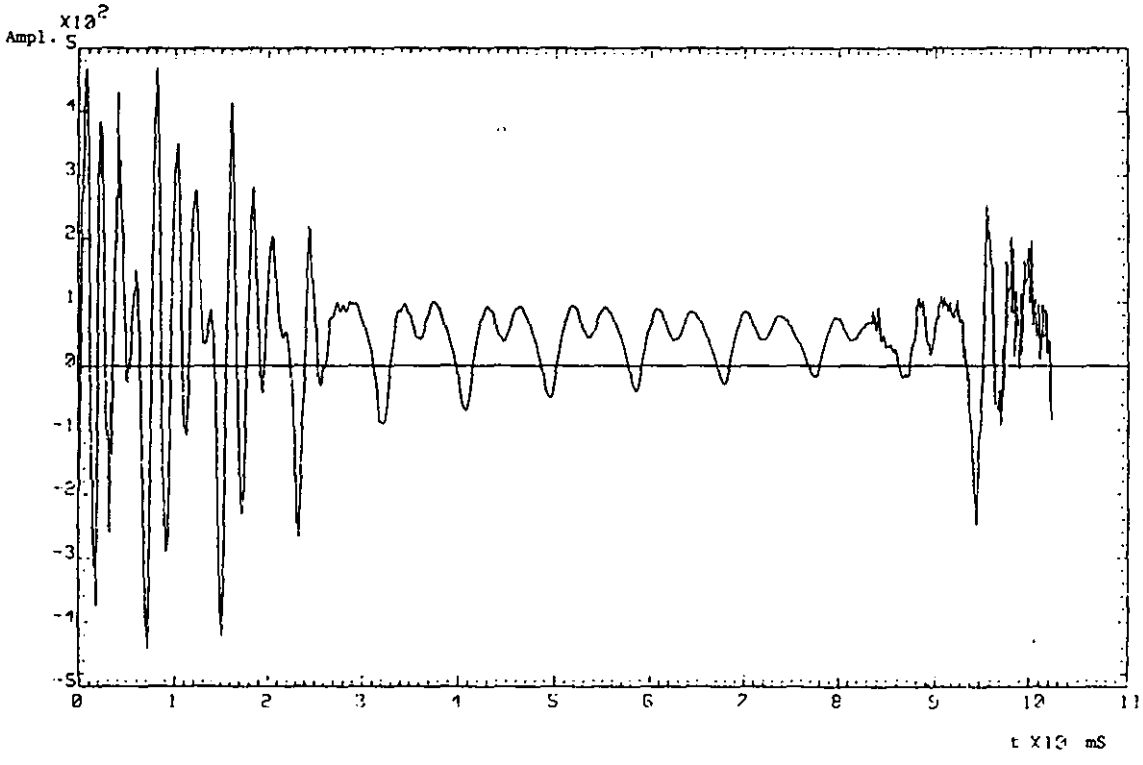


Fig 6.19(c) Output time waveform.

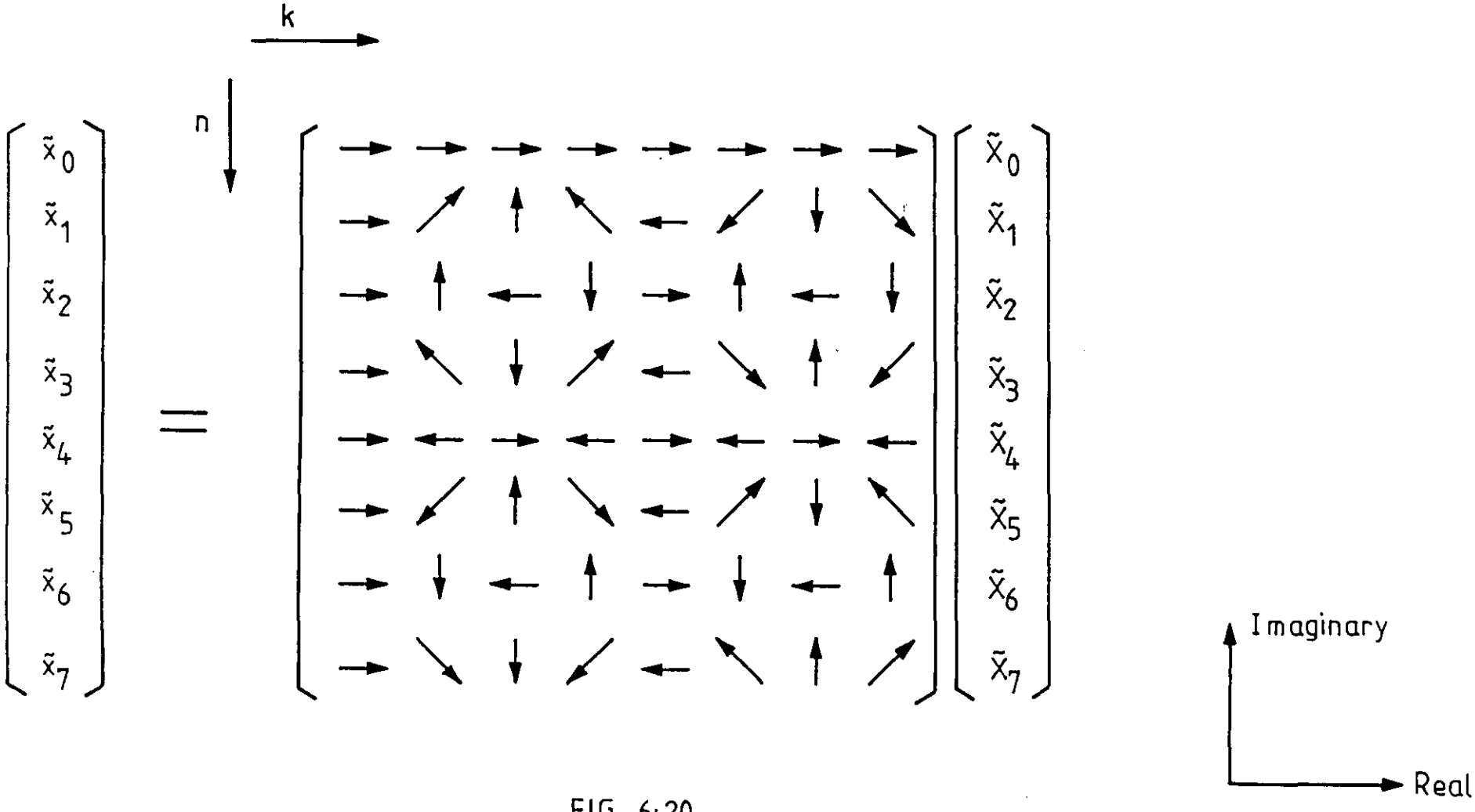

k
n
$\tilde{x}_0$
$\tilde{x}_1$
$\tilde{x}_2$
$\tilde{x}_3$
$\tilde{x}_4$
$\tilde{x}_5$
$\tilde{x}_6$
$\tilde{x}_7$
=
$\tilde{X}_0$
$\tilde{X}_1$
$\tilde{X}_2$
$\tilde{X}_3$
$\tilde{X}_4$
$\tilde{X}_5$
$\tilde{X}_6$
$\tilde{X}_7$
Imaginary
Real

FIG. 6·20

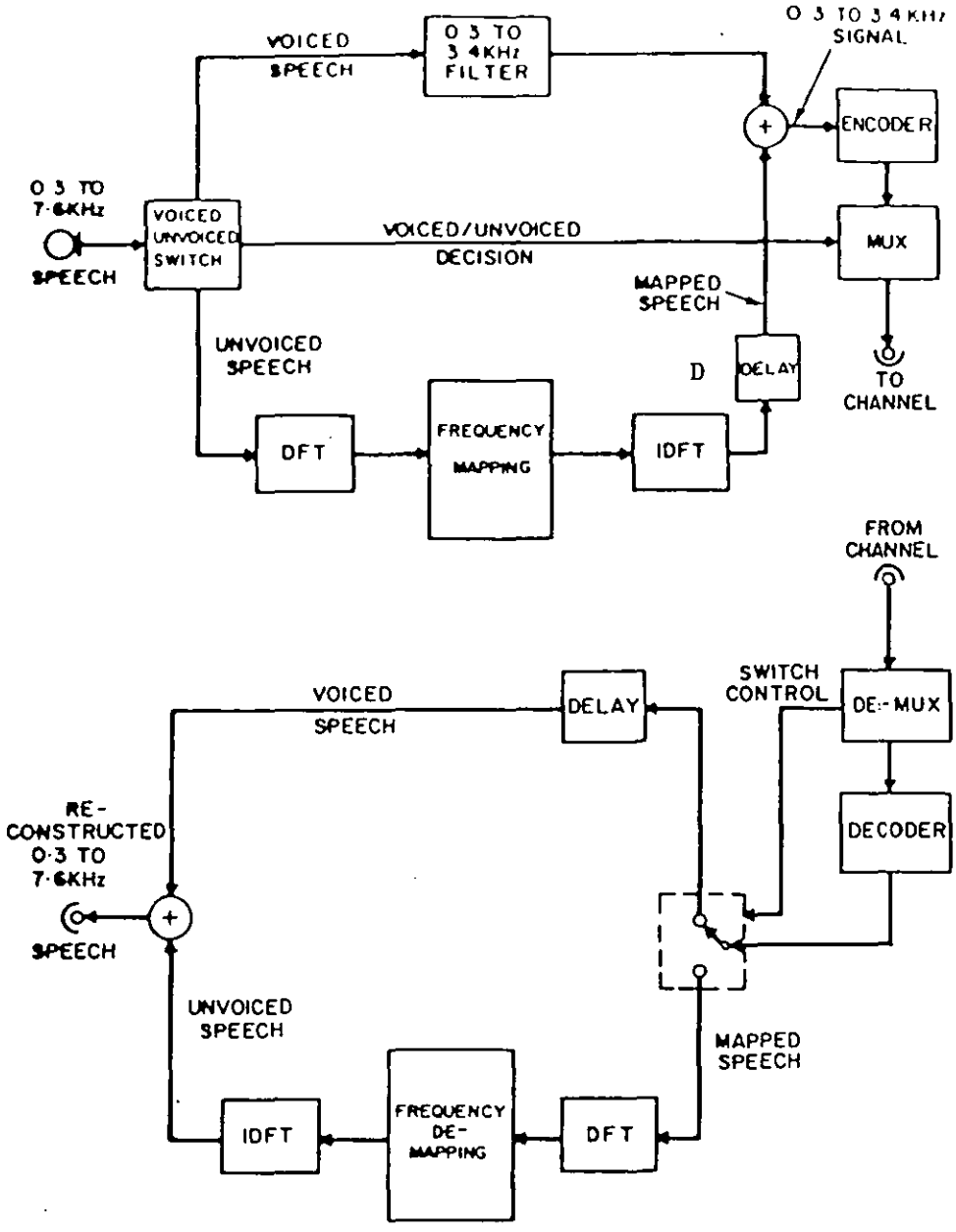


Fig 6.21

Frequency mapping system applied to unvoiced speech only.

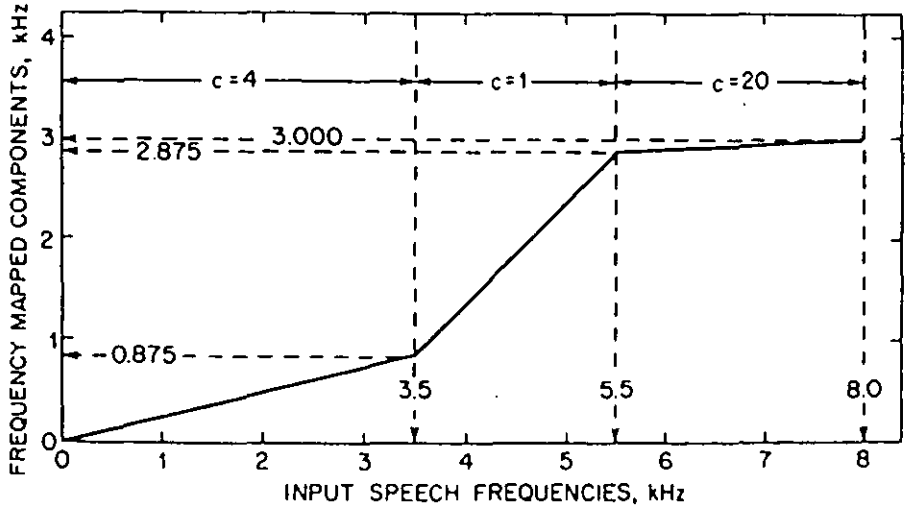
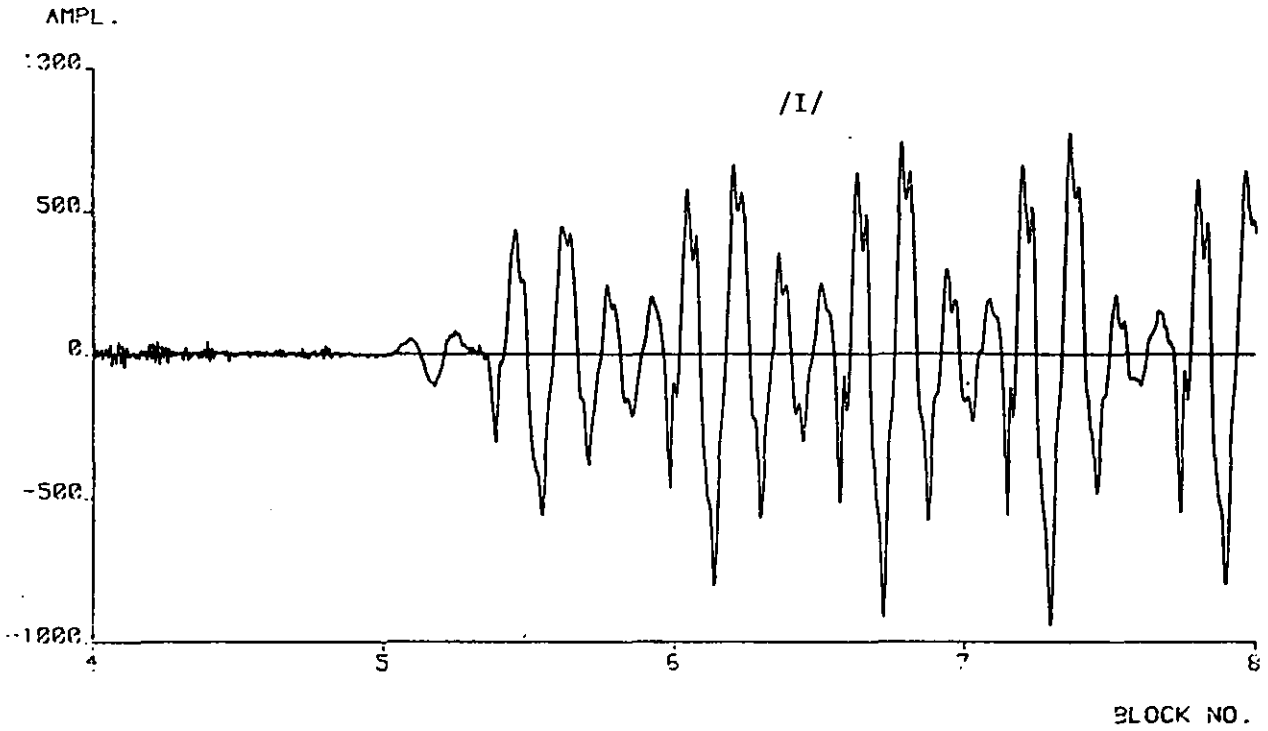


Fig 6.22 The mapping-law for /s/.

Fig 6.23(a) Original /sI/ signal
300-7600 kHz.



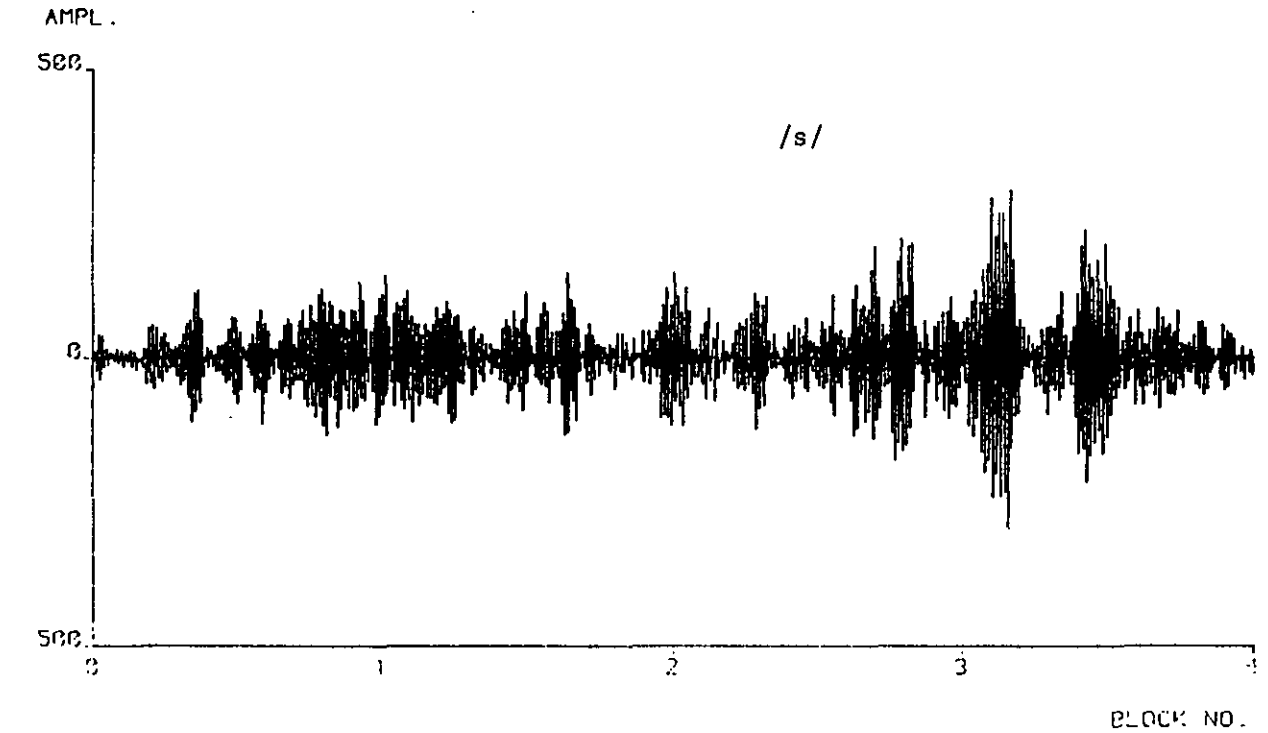
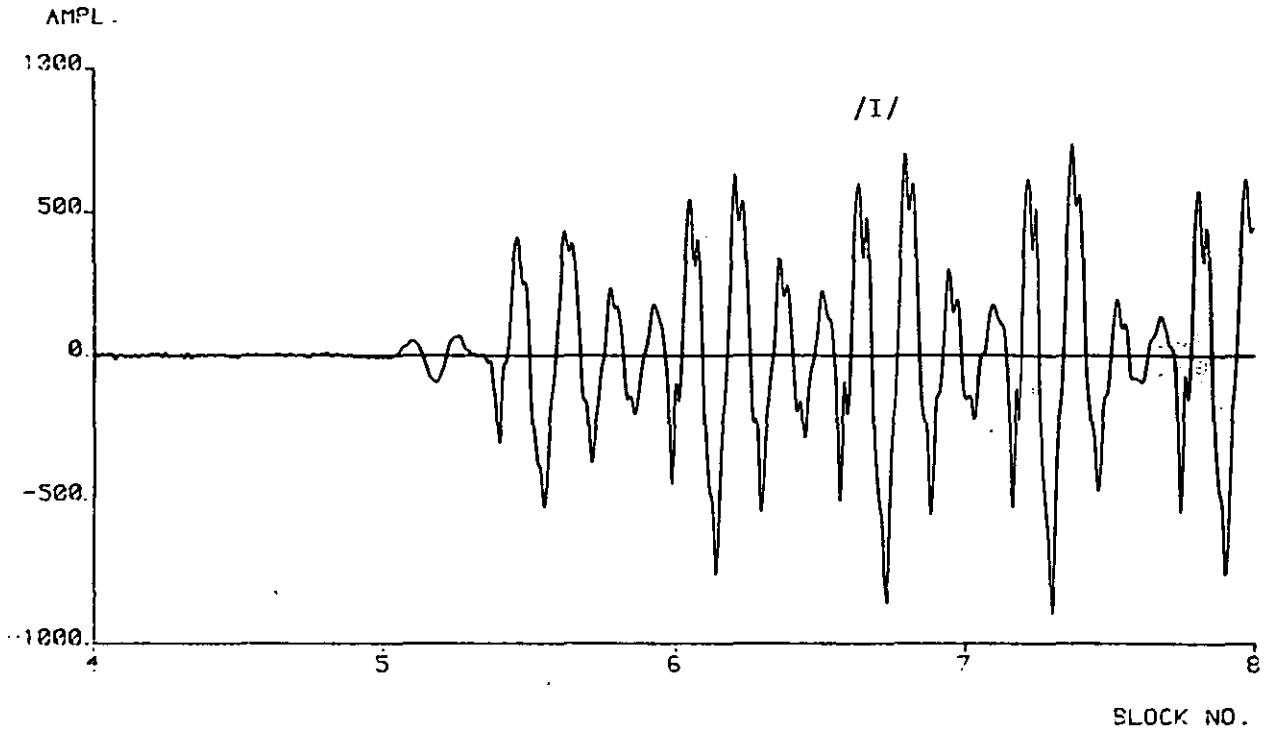


Fig 6.23(b) Bandpass filtered /sI/ signal, 300-3400 Hz.



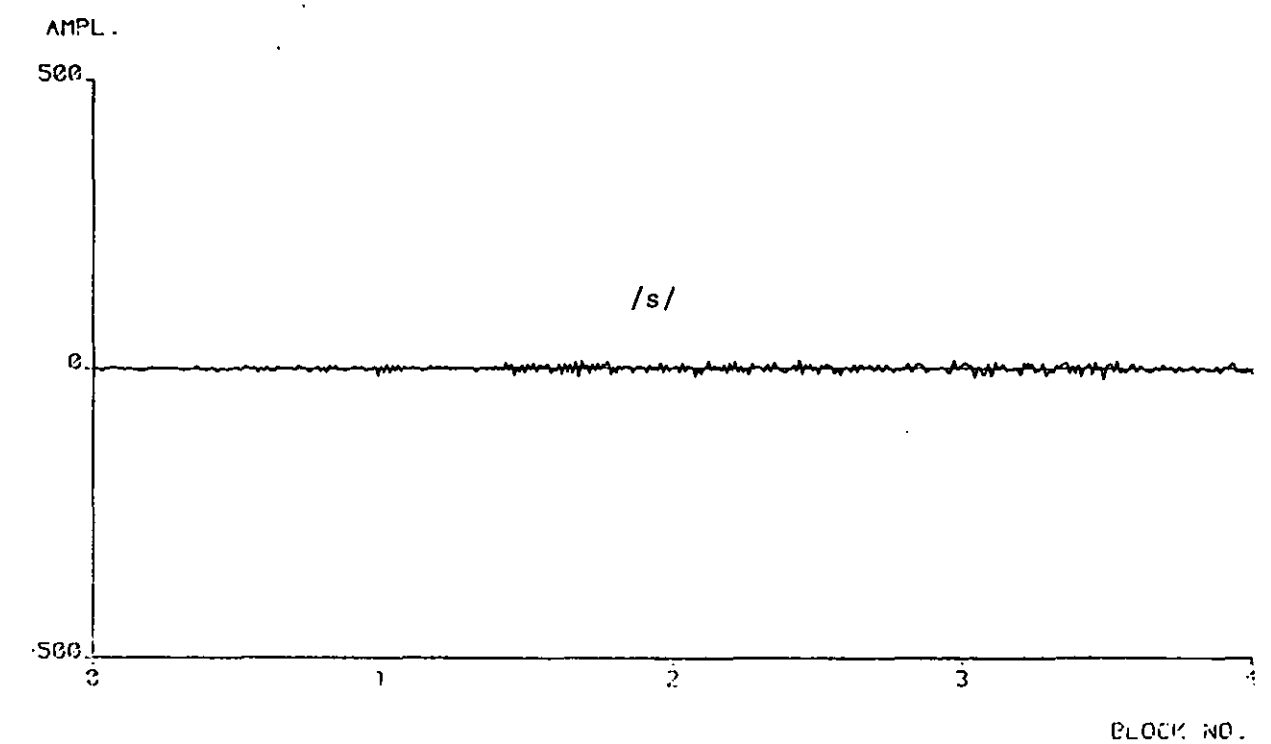
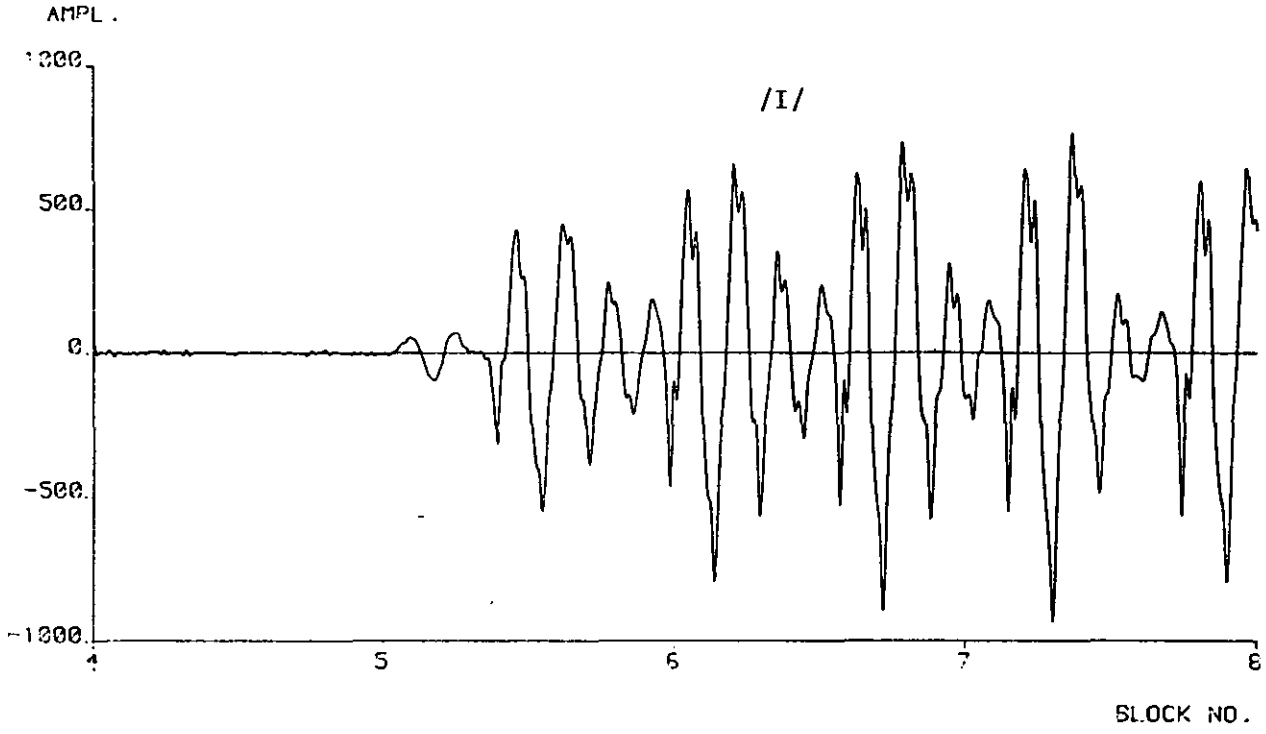


Fig 6.23(c) Frequency mapped 300-7600 Hz speech into a 300-3400 Hz signal.



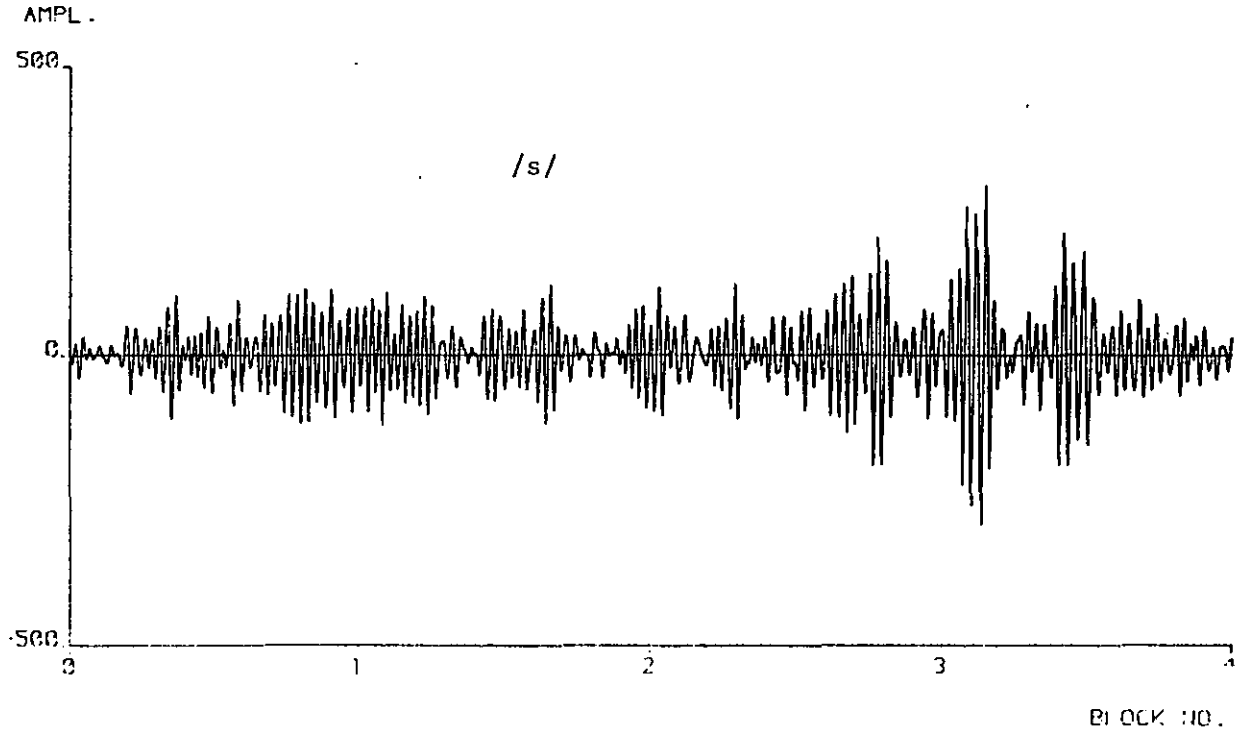
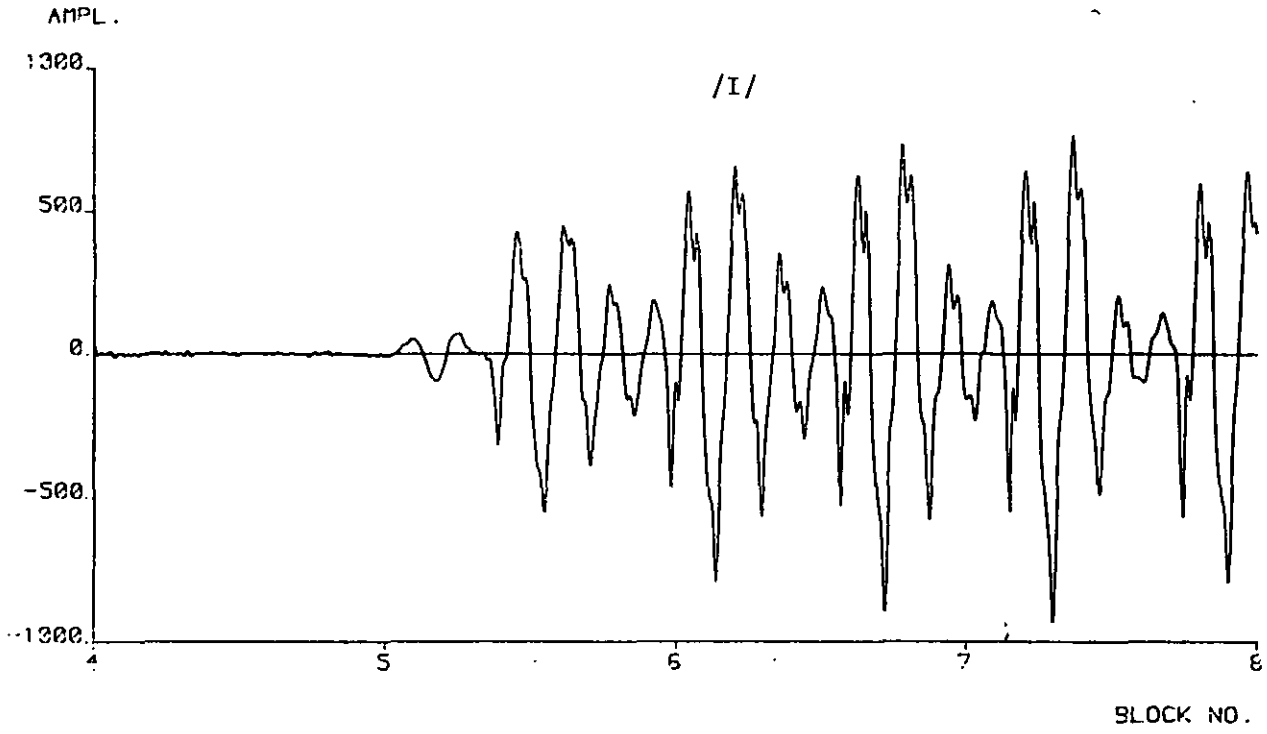
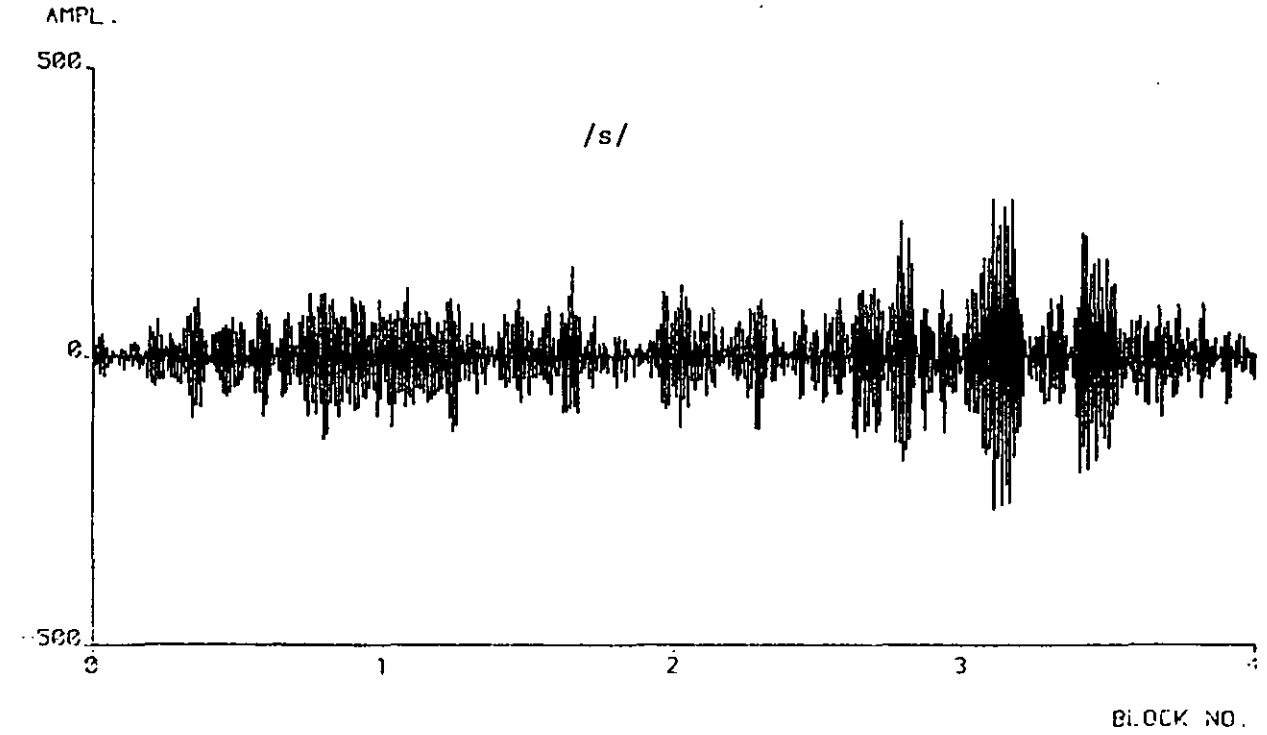


Fig 6.23(d) Recovered frequency mapped speech.





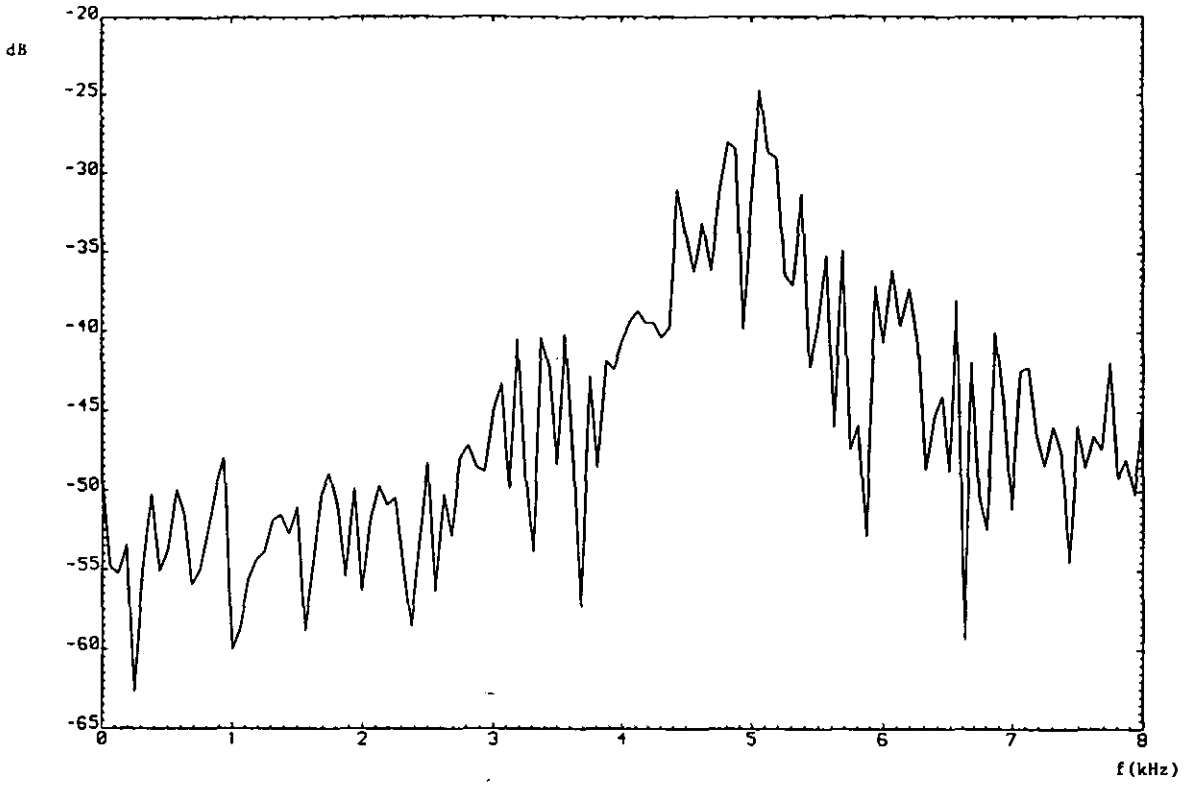


Fig 6.24(a) Input spectrum of an /s/ sound in S.K. Harvey.

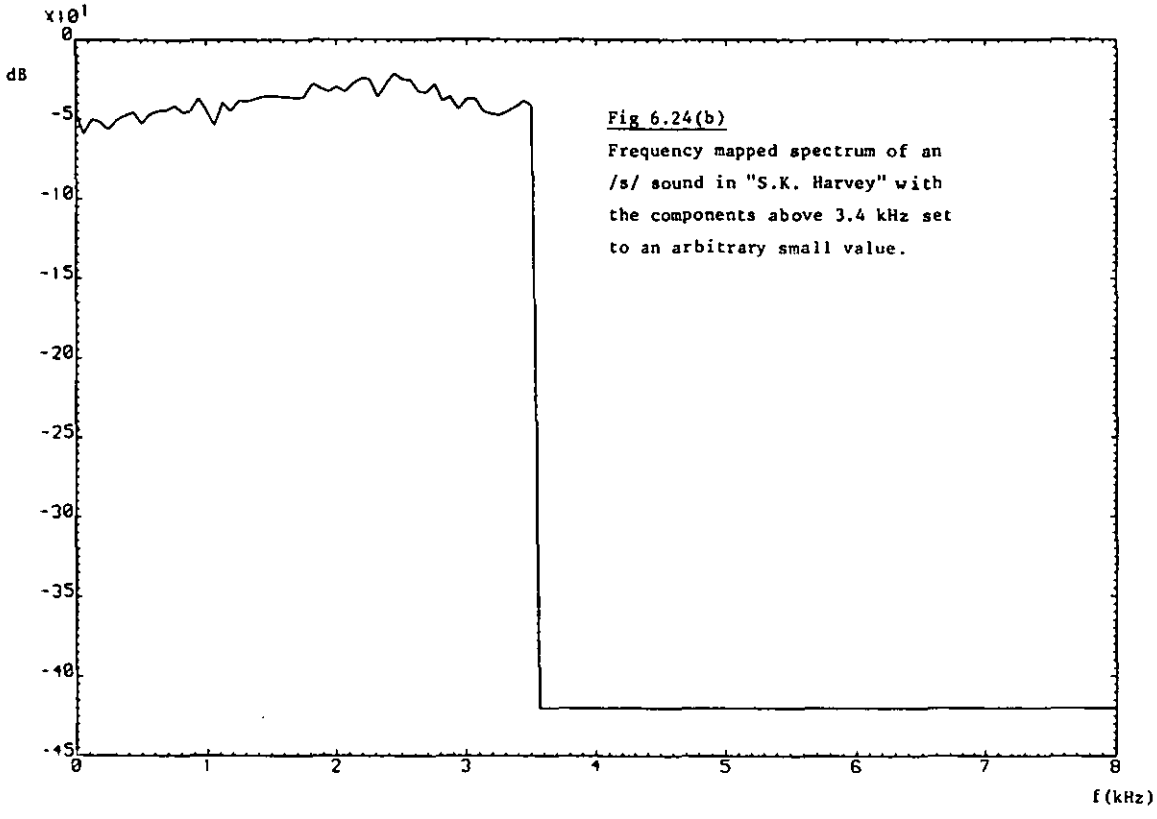


Fig 6.24(b)
Frequency mapped spectrum of an /s/ sound in "S.K. Harvey" with the components above 3.4 kHz set to an arbitrary small value.

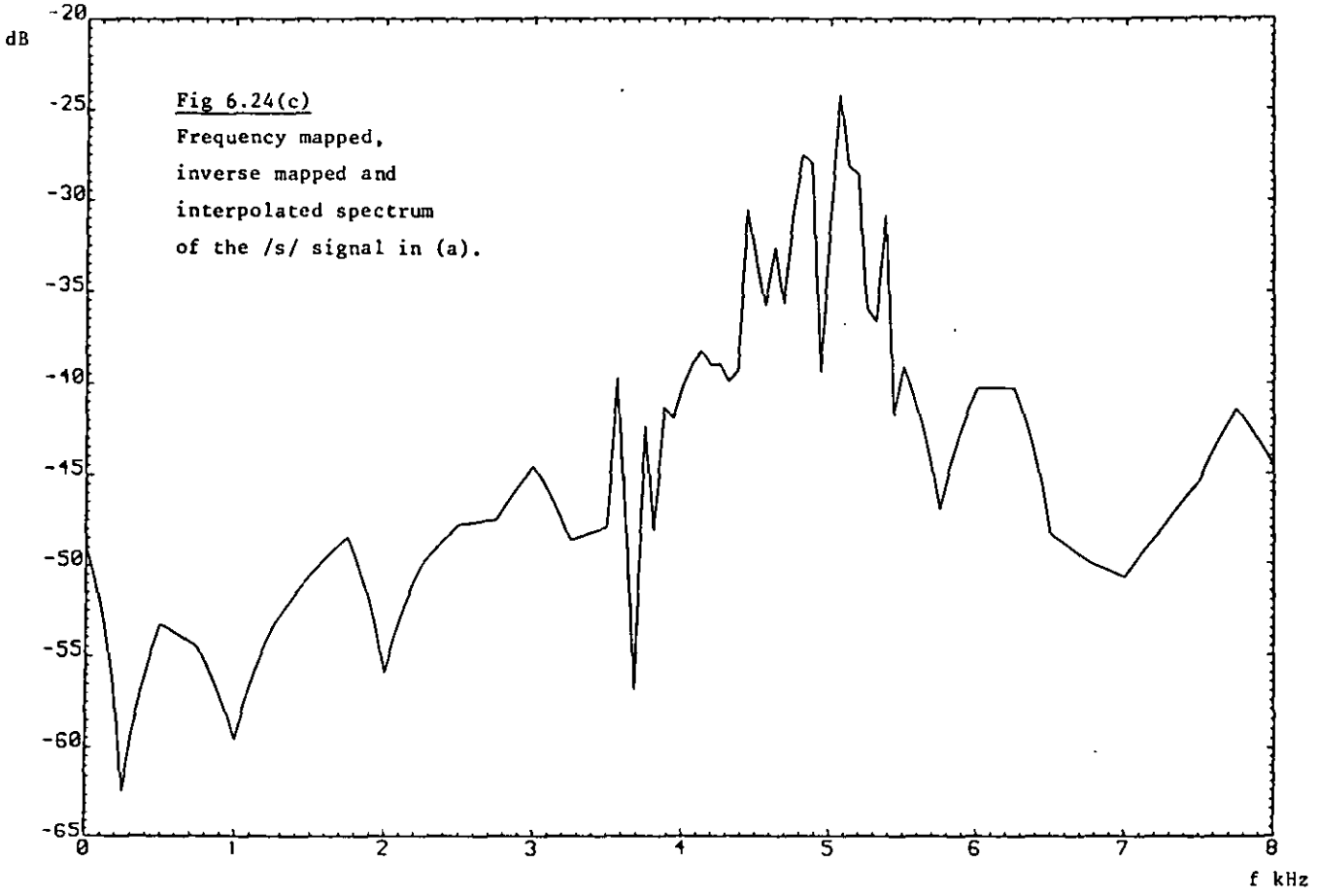
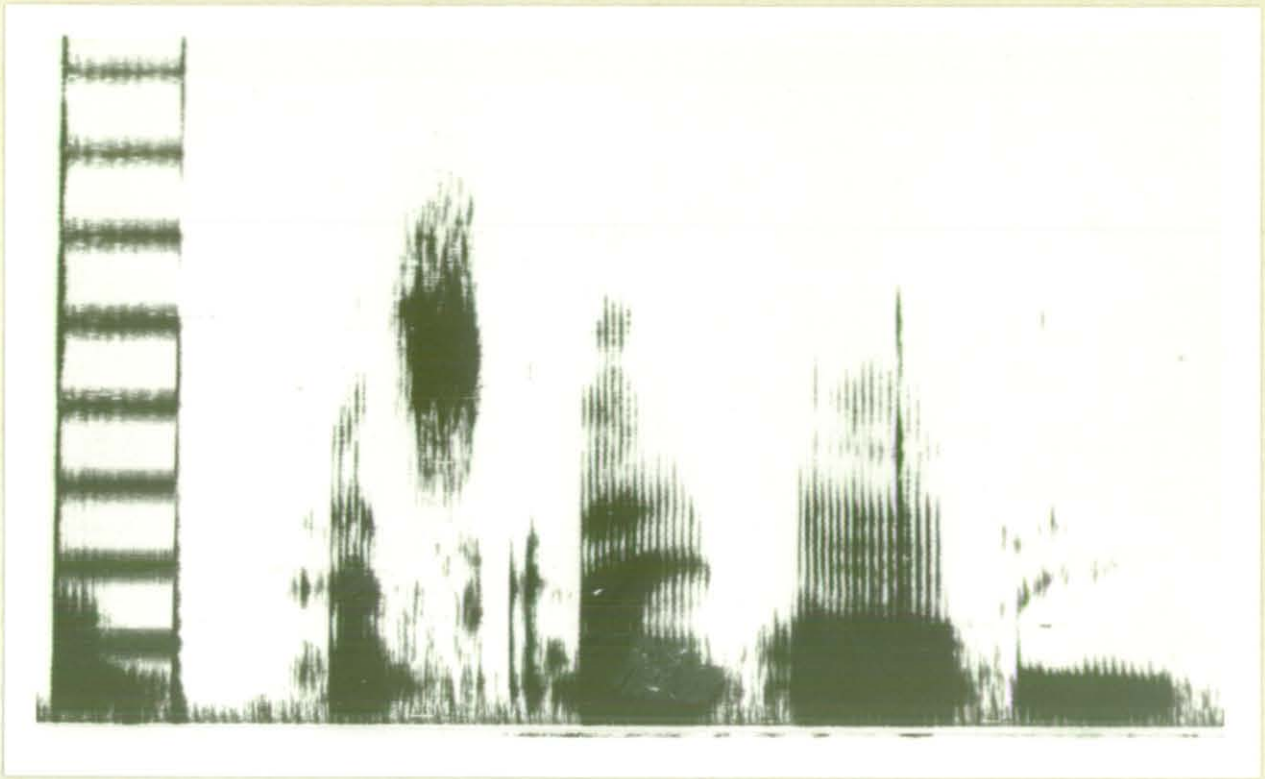
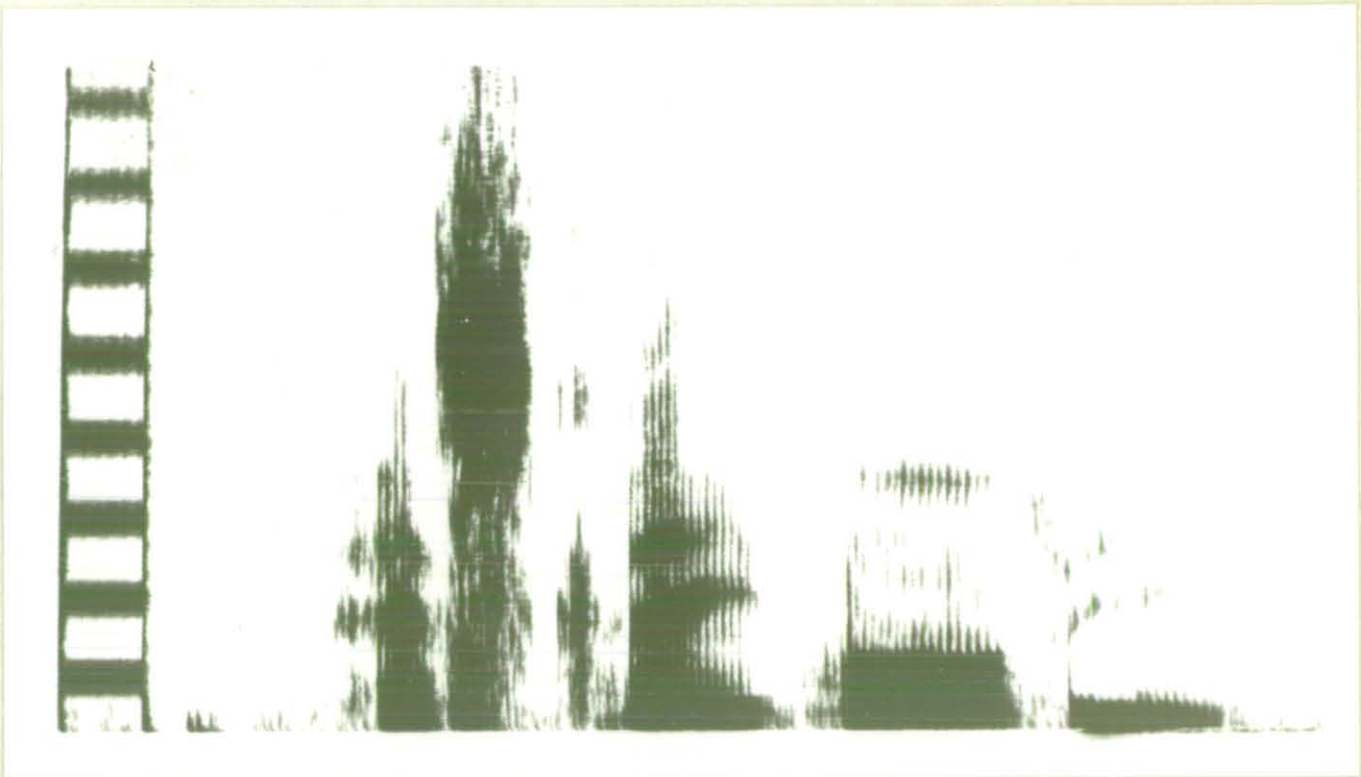


Fig 6.24(c)
Frequency mapped,
inverse mapped and
interpolated spectrum
of the /s/ signal in (a).



(a) 0.3 - 7.6 kHz original speech



(b) Recovered frequency mapped speech

Fig. 6.25 Spectrograms for the words "sister" and "father"

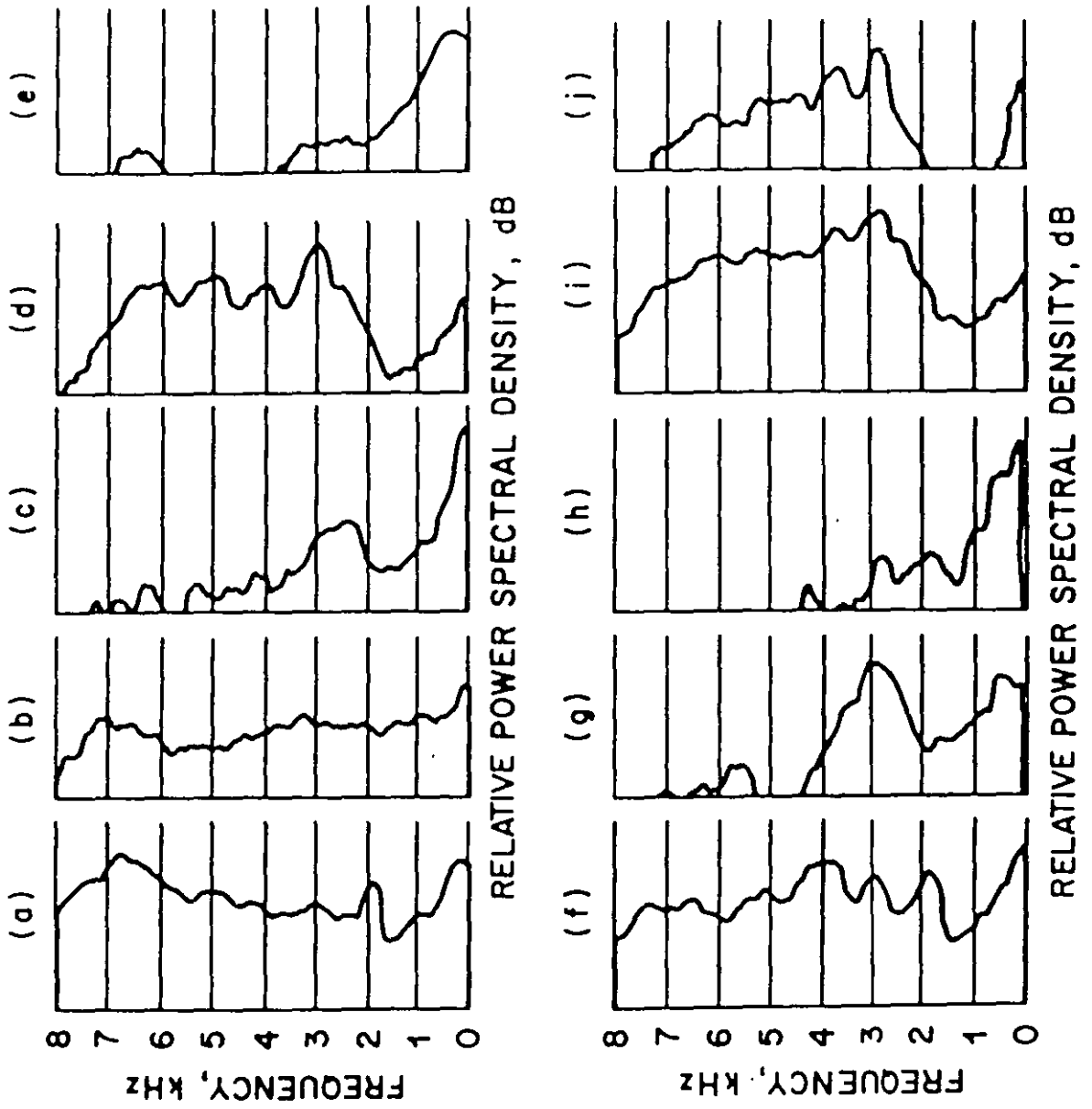


Fig 6.26 Some spectral density functions of unvoiced sounds.

(a) /s/ as in see
(b) /f/ as in for
(c) /h/ as in he
(d) /ʃ/ as in she
(e) /θ/ as in thin
(f) /t/ as in to
(g) /k/ as in key
(h) /p/ as in pay
(i) /tʃ/ as in chew
(j) /dz/ as in jar

A dc marker is present on all the spectra from 0-300 Hz.

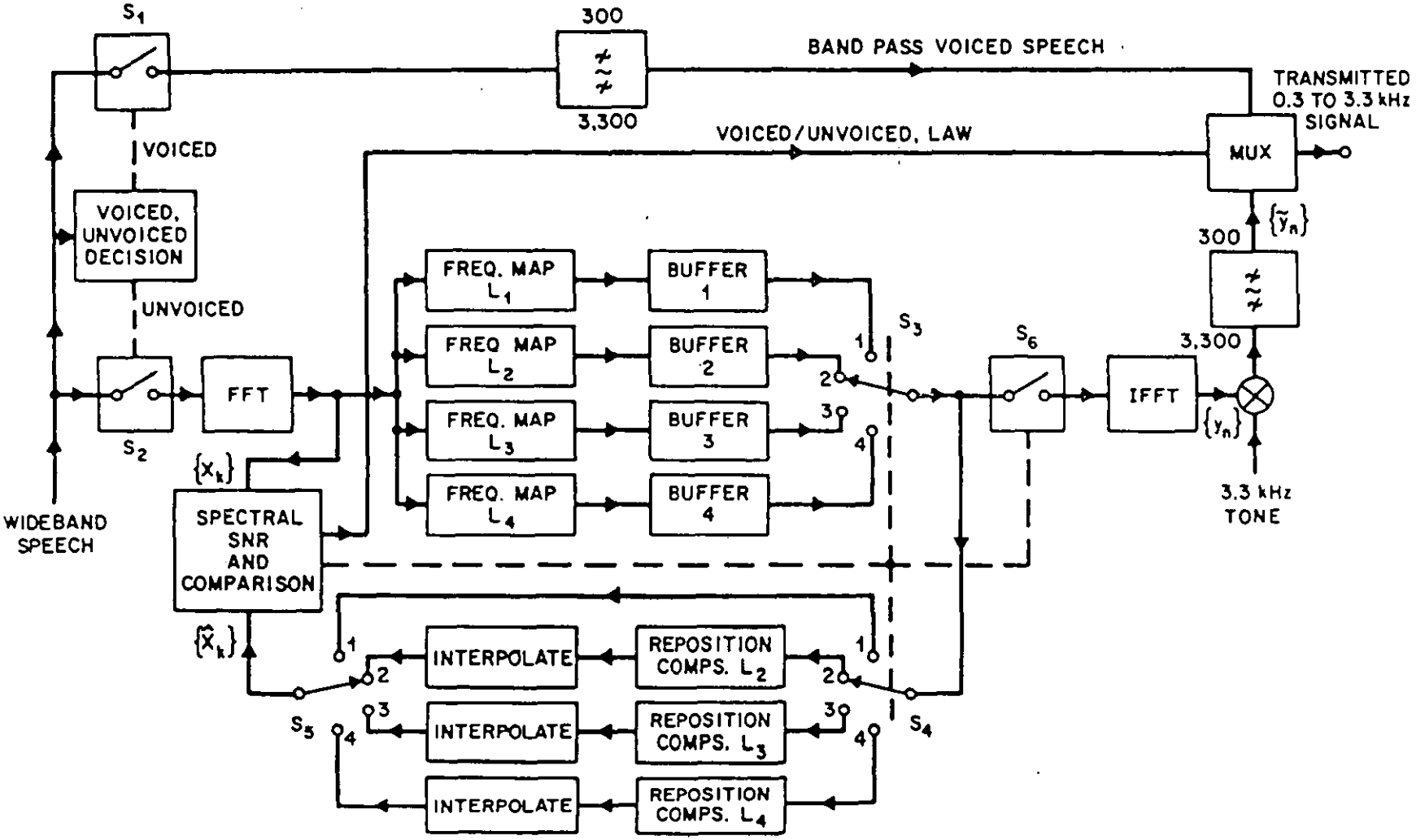


Fig 6.27 Adaptive frequency mapping arrangement.

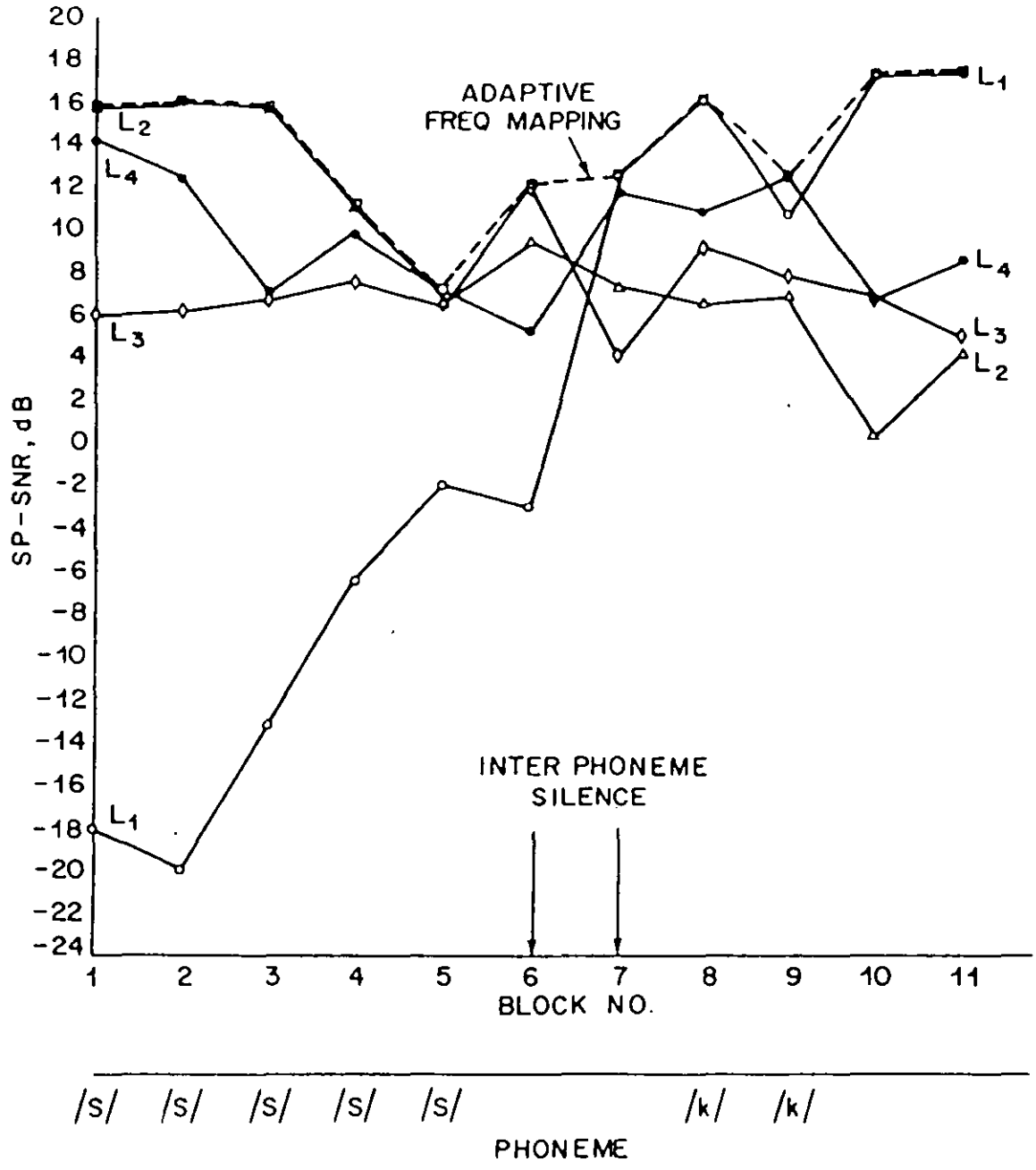


Fig 6.28 Variation of spectral SNR against block number for the four mapping laws and the selection by the adaptive frequency mapping system.

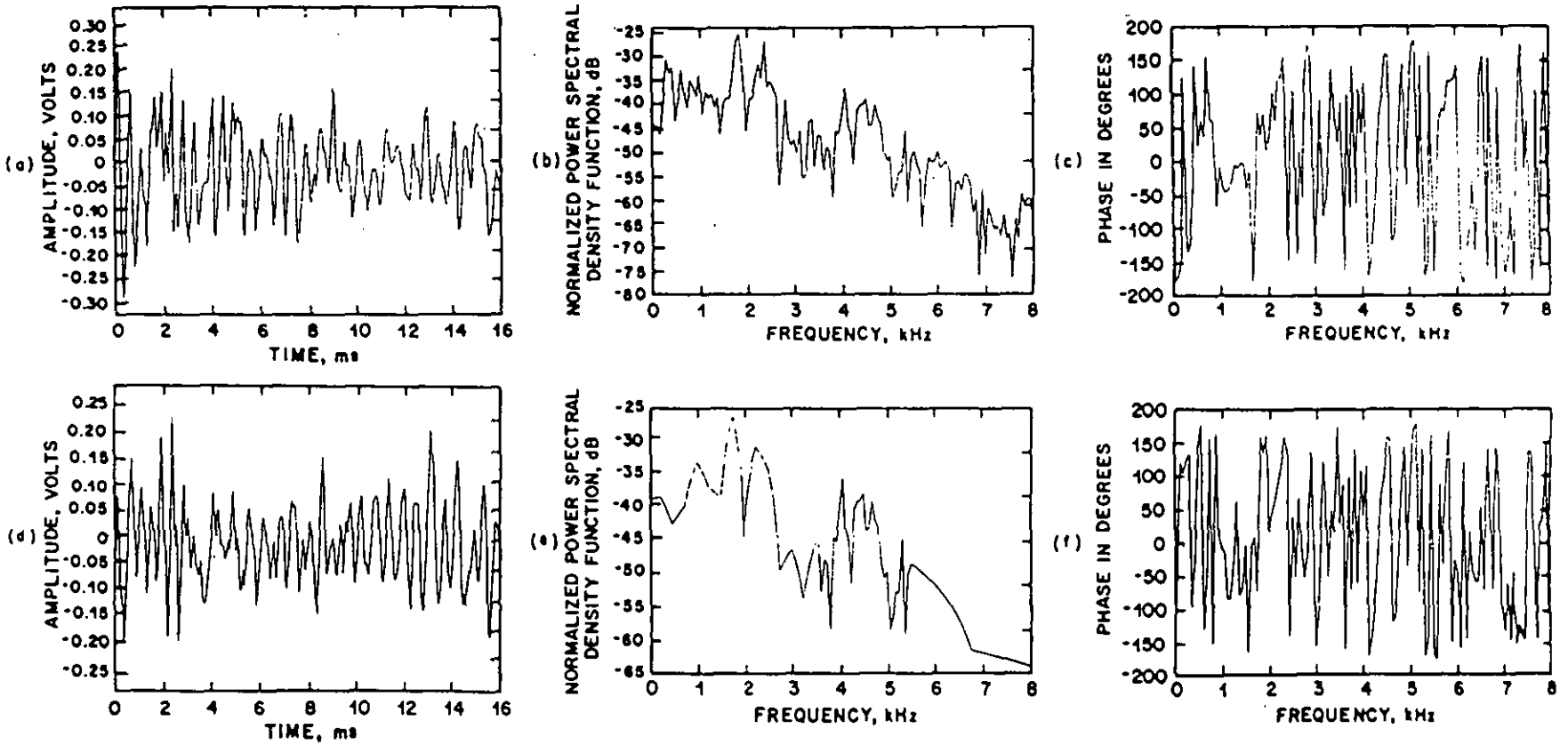


Fig 6.29 Time waveforms, power spectral density and phase spectra for the second 16 mS block of /k/ in the utterance "S.K. Harvey". For the wideband signal, (a) is the time function, (b) is the power spectrum and (c) is the phase spectrum. For the frequency de-mapped signals using laws L_2, L_3 and L_4 respectively, (d), (g) and (j) are the time waveforms, (e), (h) and (k) are the power spectra and (f), (i) and (ℓ) are the phase spectra.

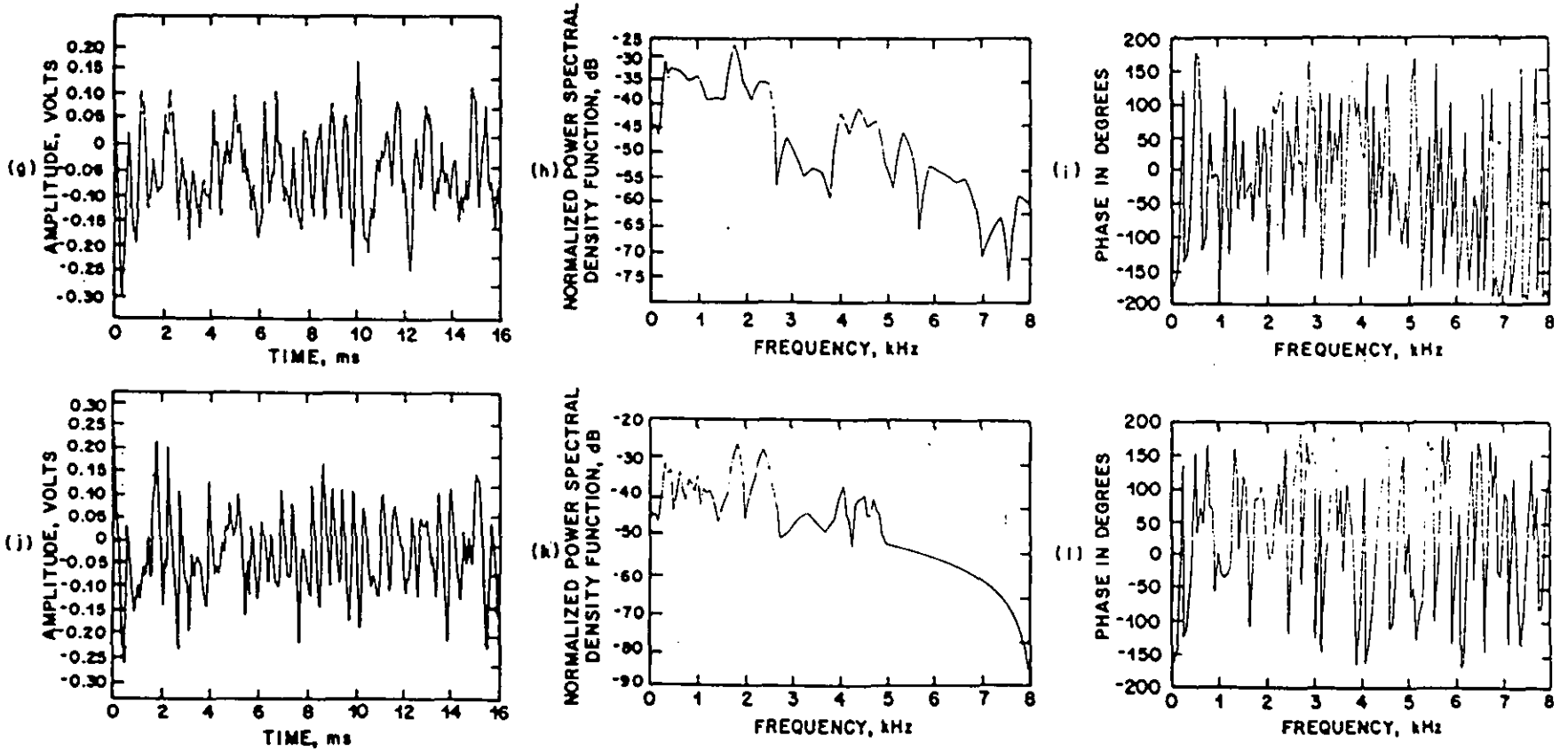


Fig 6.29 contd.

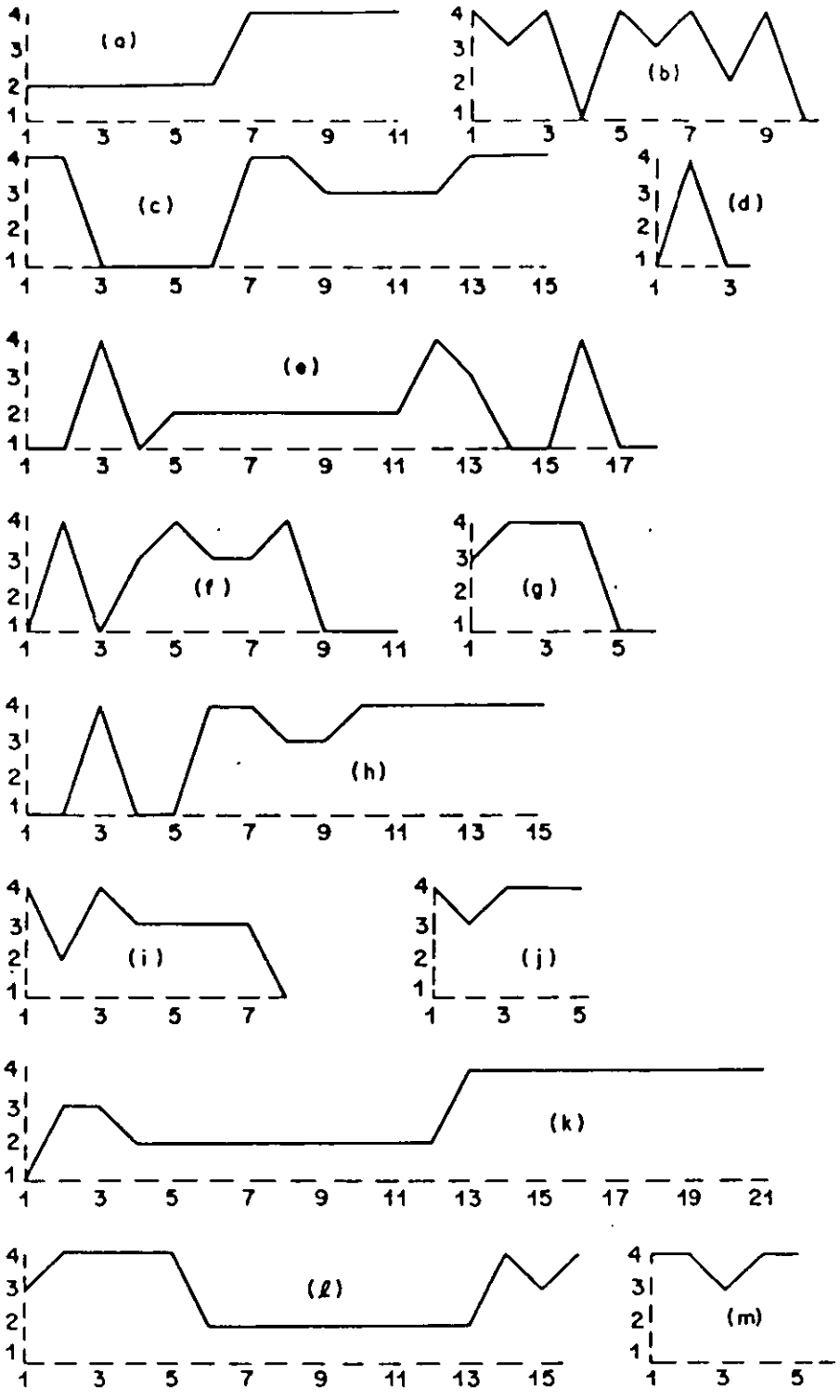


Fig 6.30 The mapping law used as a function of the block number measured from the beginning of the unvoiced sound.

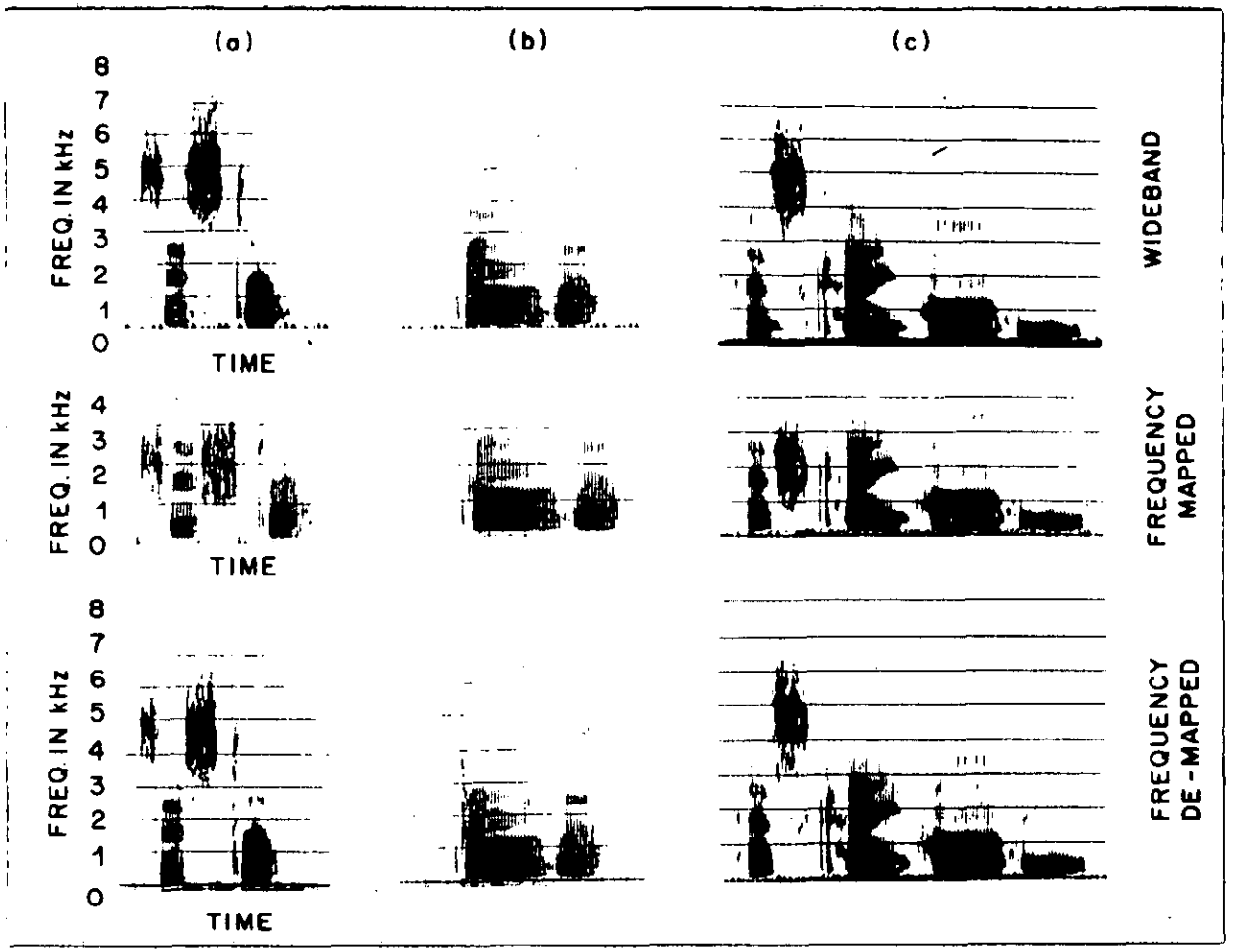


Fig. 6.31 Spectrograms of wideband frequency mapped and frequency demapped words.

(a) "sister",

(b) "father",

(c) "S.K.Harvey",

(d) "shift",

(e) "thick",

(f) "fist" and

(g) "talk".

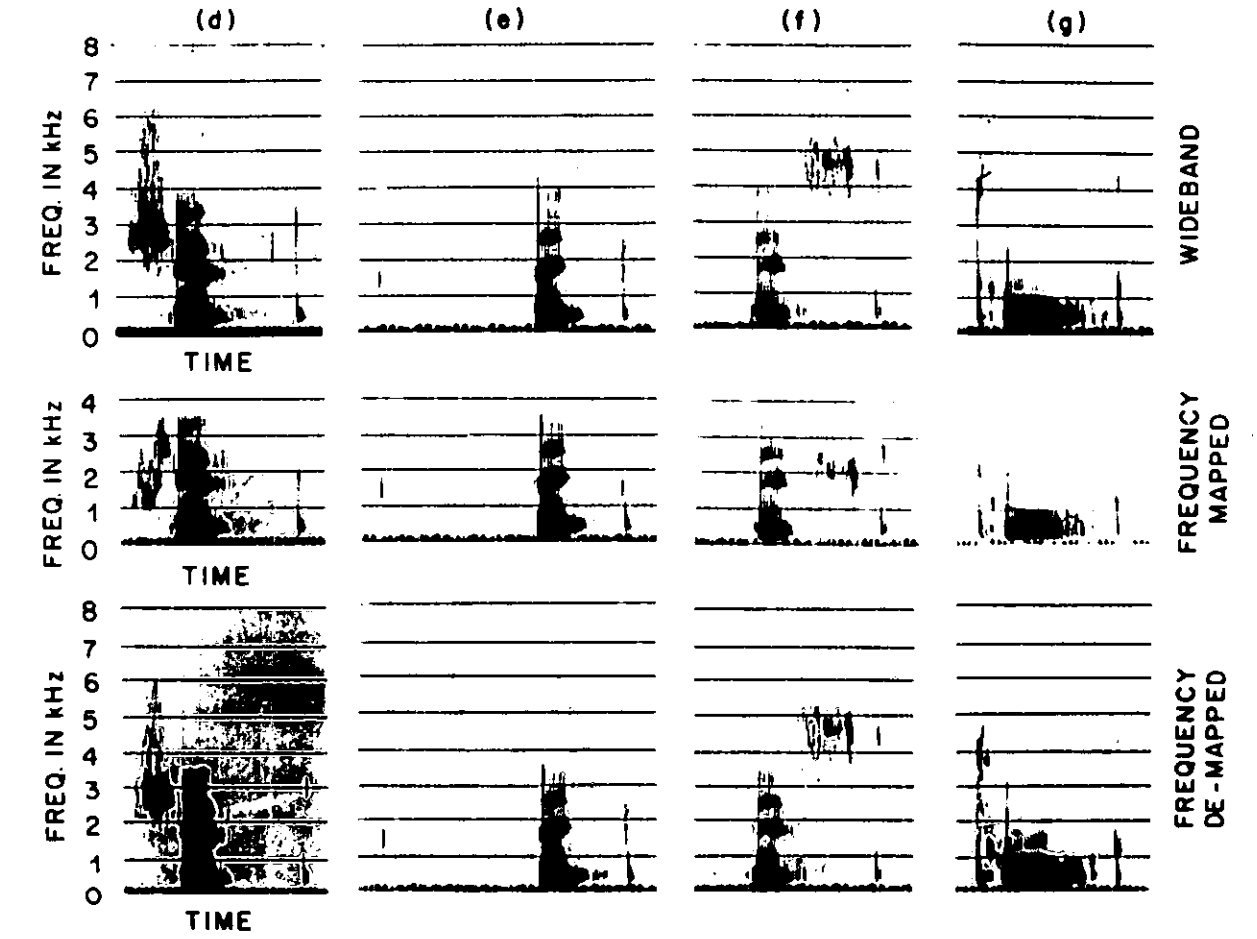


Fig. 6.31. (Continued)

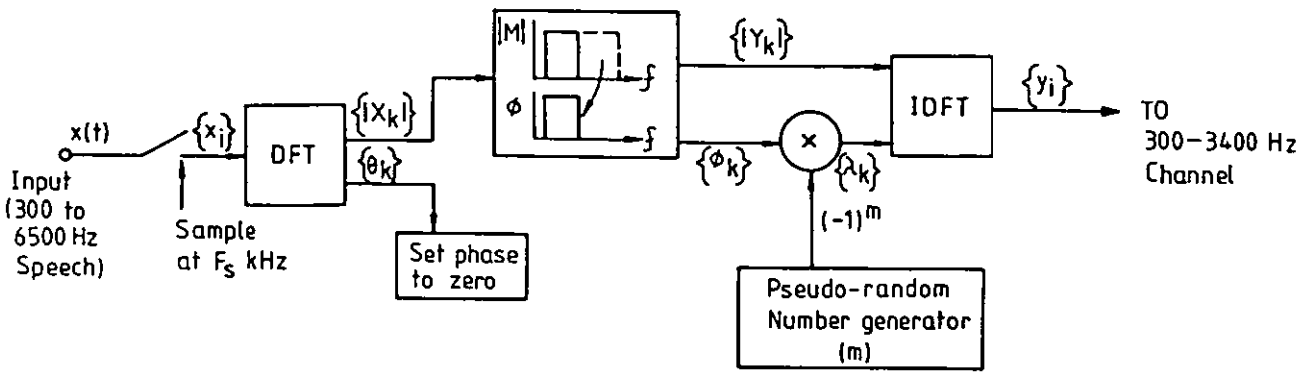


FIG. 6·32(a) RANDOM PHASE PROCESSOR TRANSMITTER

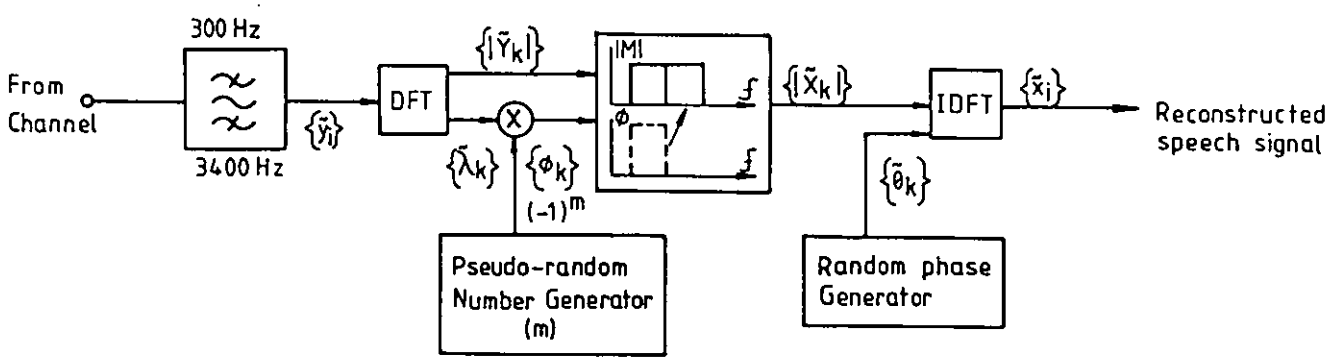


FIG. 6·32(b) RANDOM PHASE PROCESSOR RECEIVER

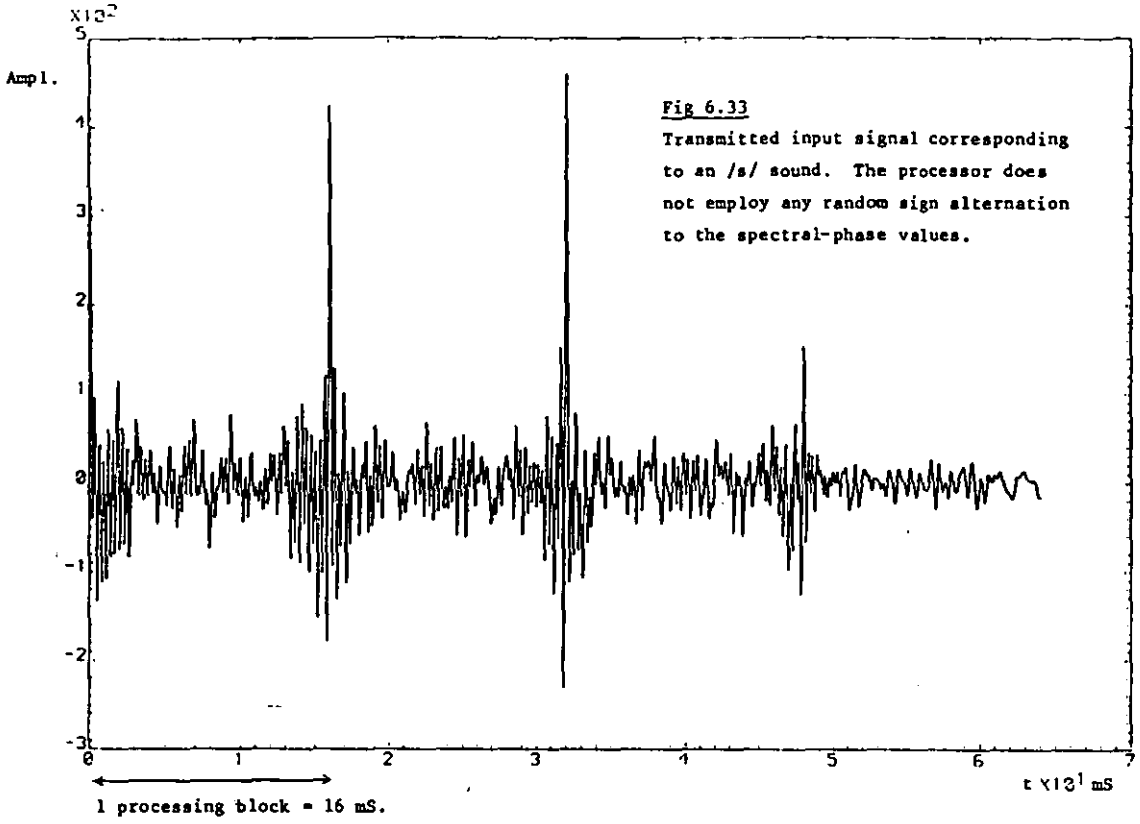


Fig 6.33
Transmitted input signal corresponding to an /s/ sound. The processor does not employ any random sign alternation to the spectral-phase values.

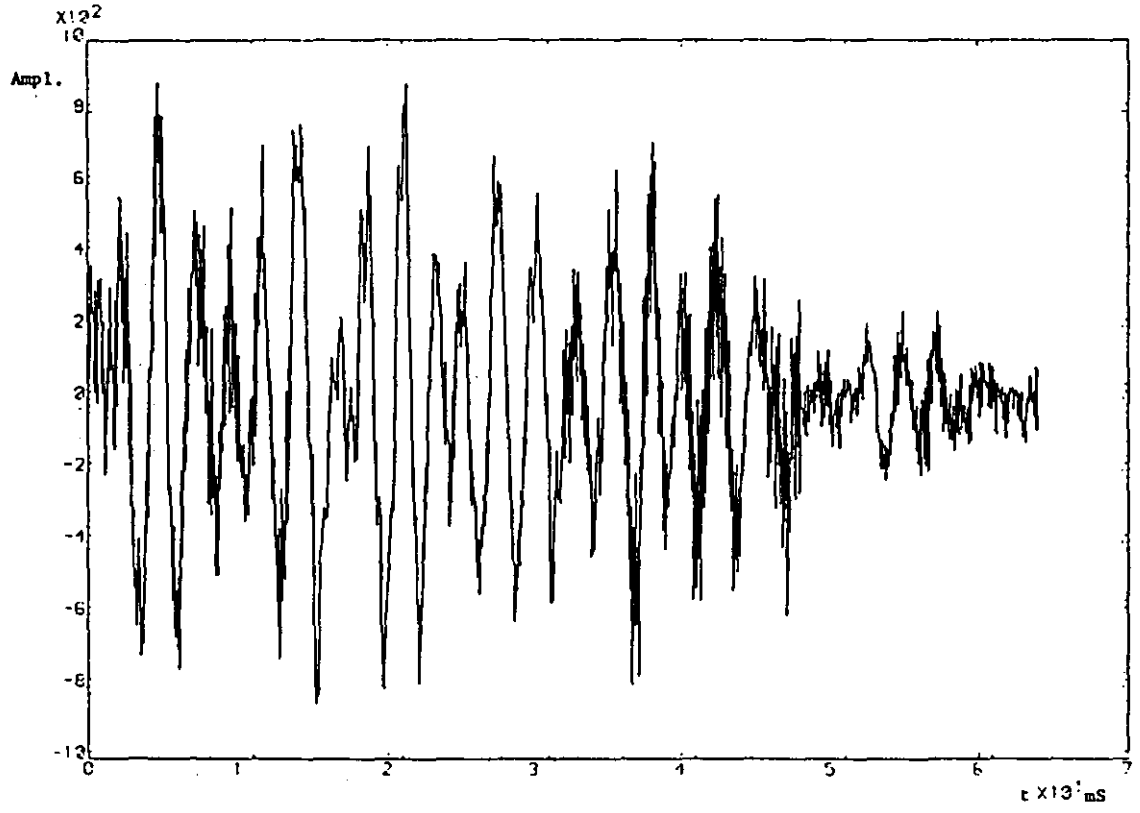


Fig 6.34 Voiced /I/ with random phase processing applied.

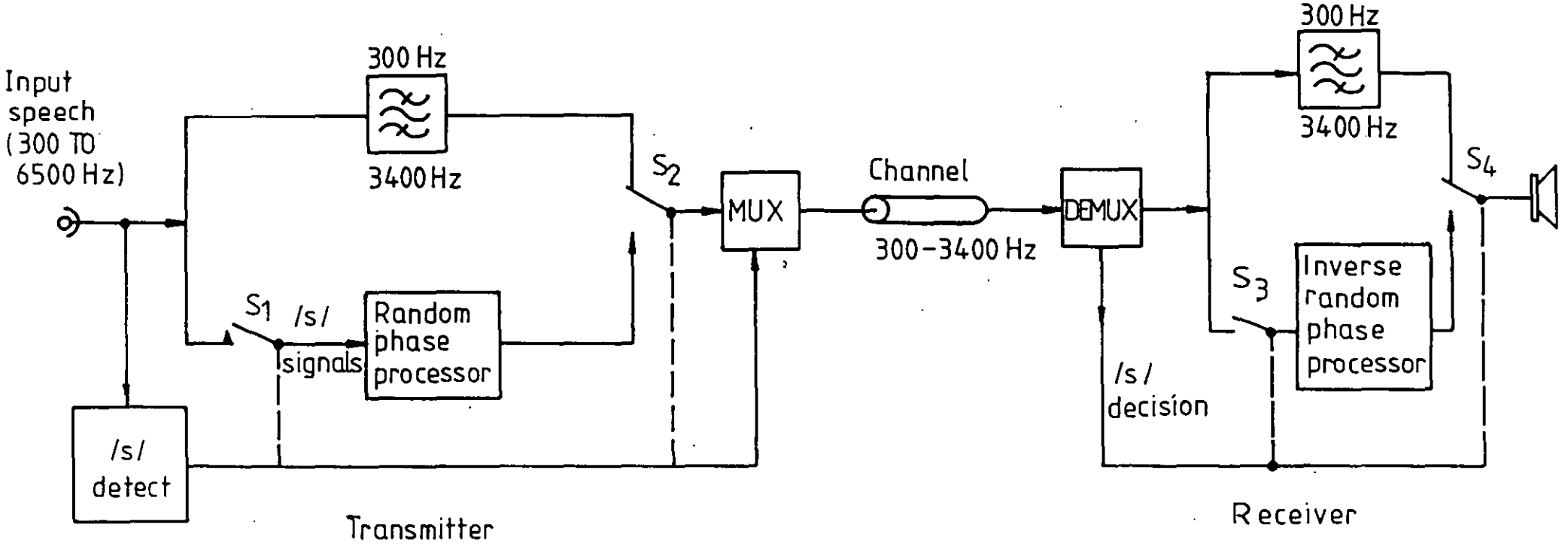
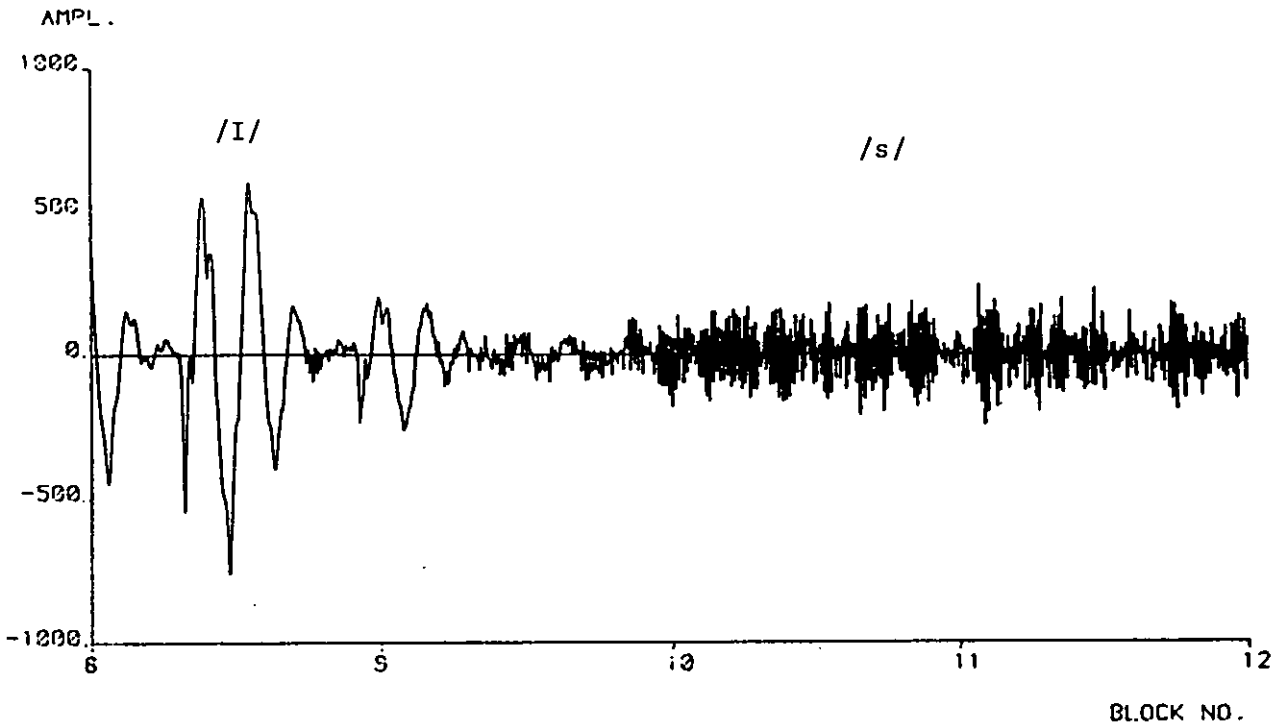


FIG.6·35 RANDOM PHASE PROCESSOR APPLIED TO /s/ SOUNDS ONLY

(a) Original 300-7600 Hz signal.



(b) BPF signal, 300-3400 Hz.

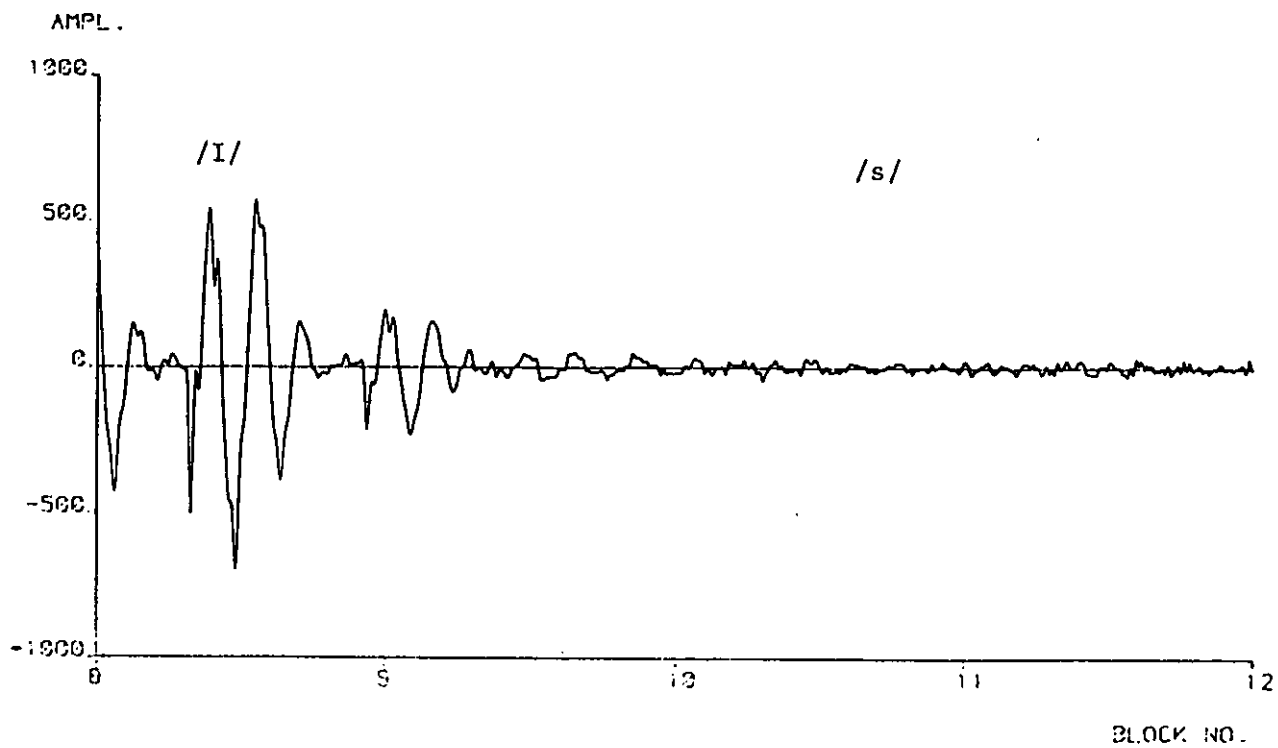
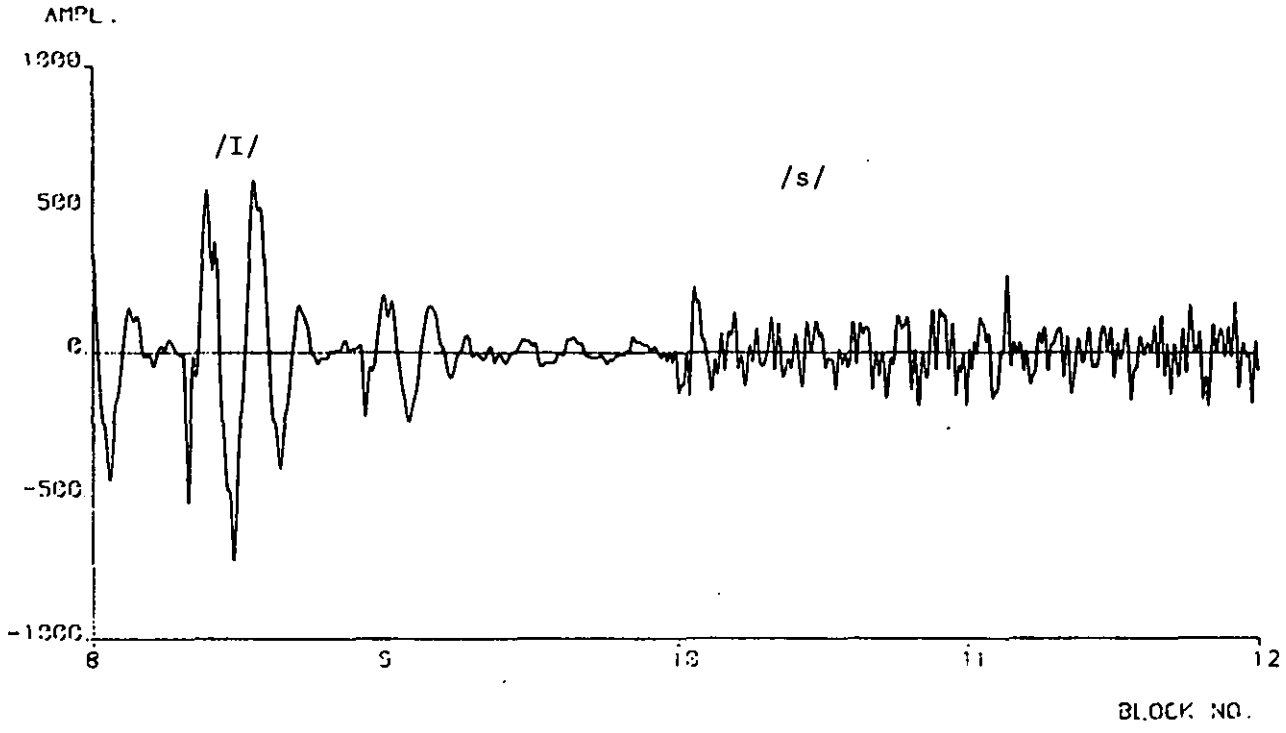


Fig 6.36

(c) Transmitted signal, 300-3400 Hz.



(d) Random phase processed signal, 300-6500 Hz.

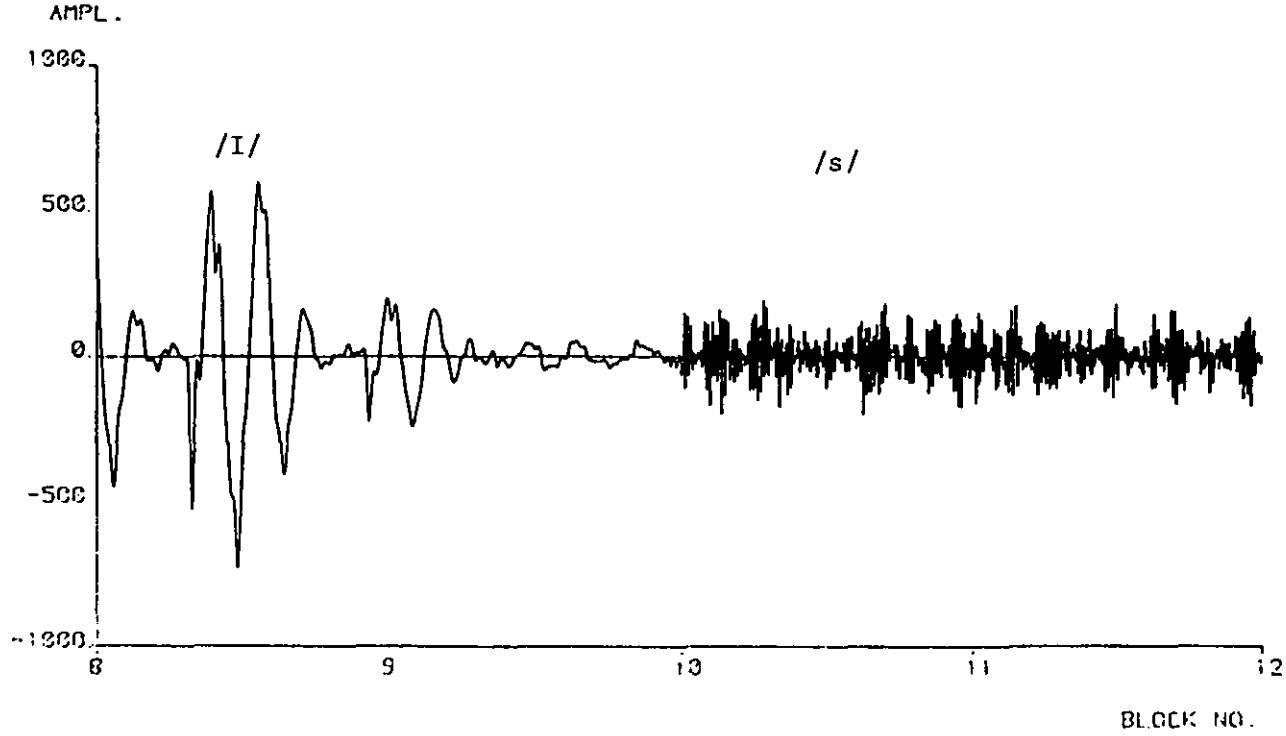


Fig 6.36

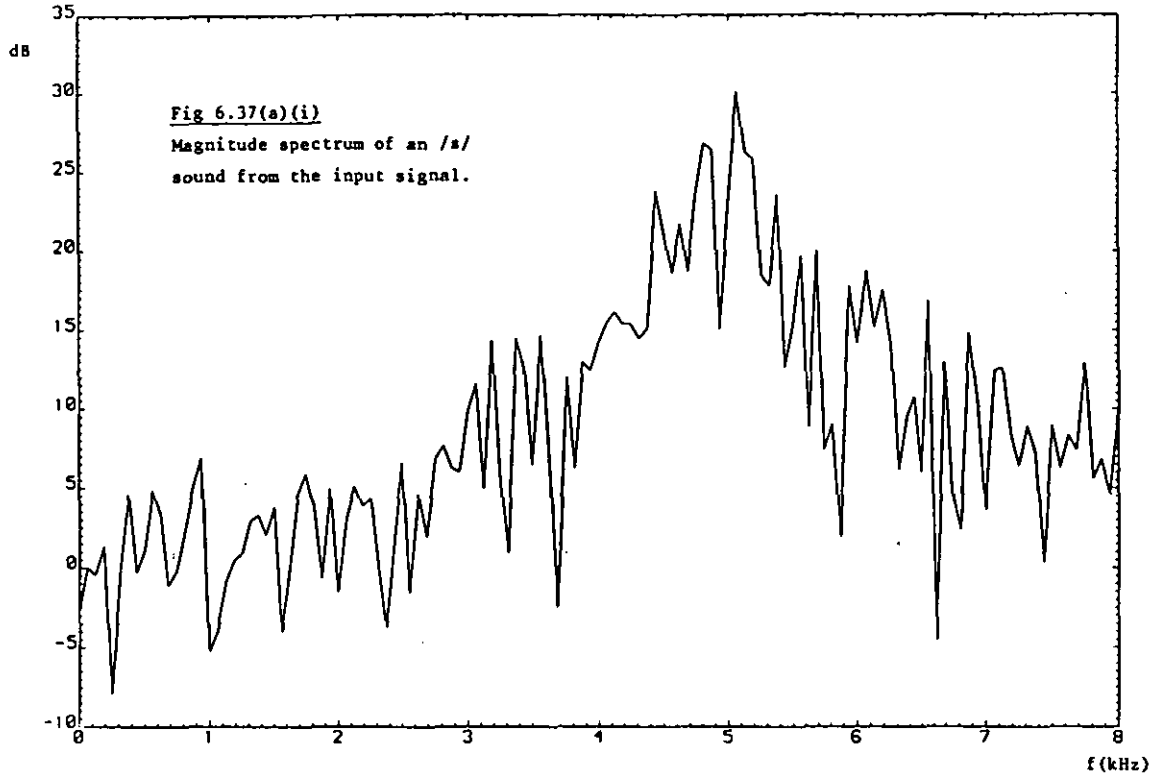


Fig 6.37(a)(i)
Magnitude spectrum of an /s/ sound from the input signal.

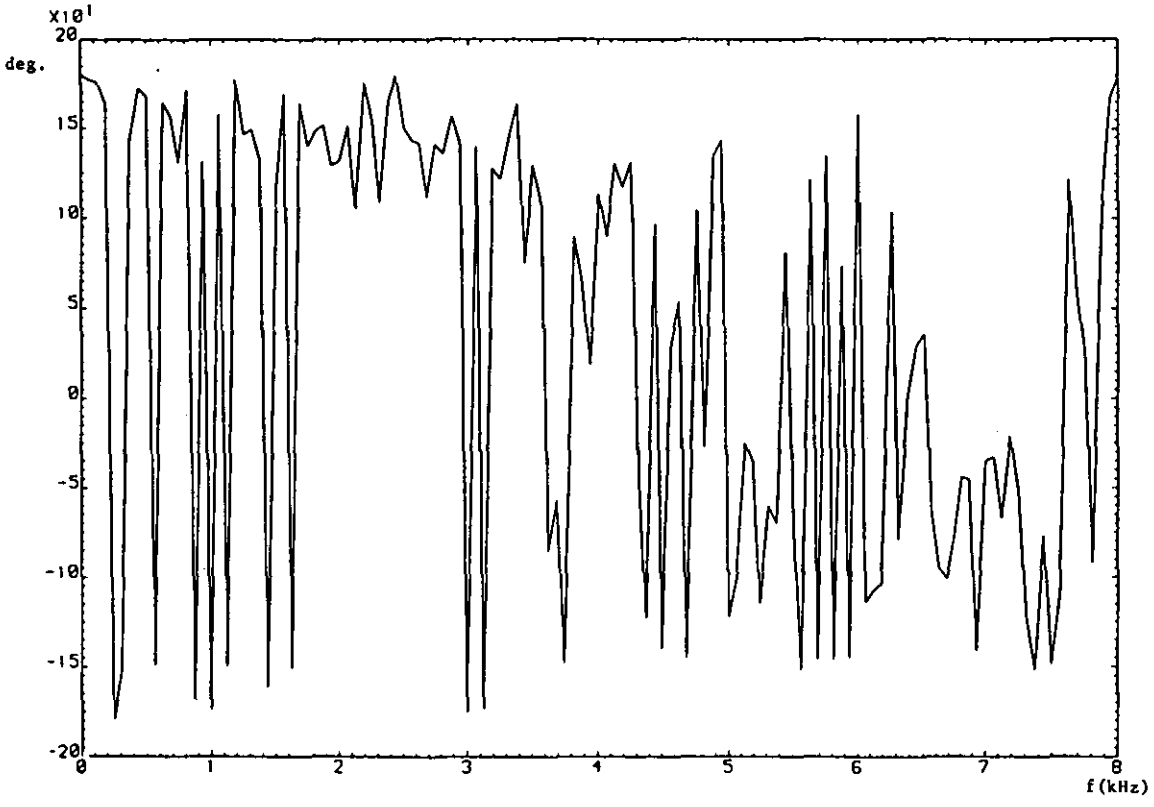


Fig 6.37(a)(ii) Phase spectrum of the /s/ sound in (i).

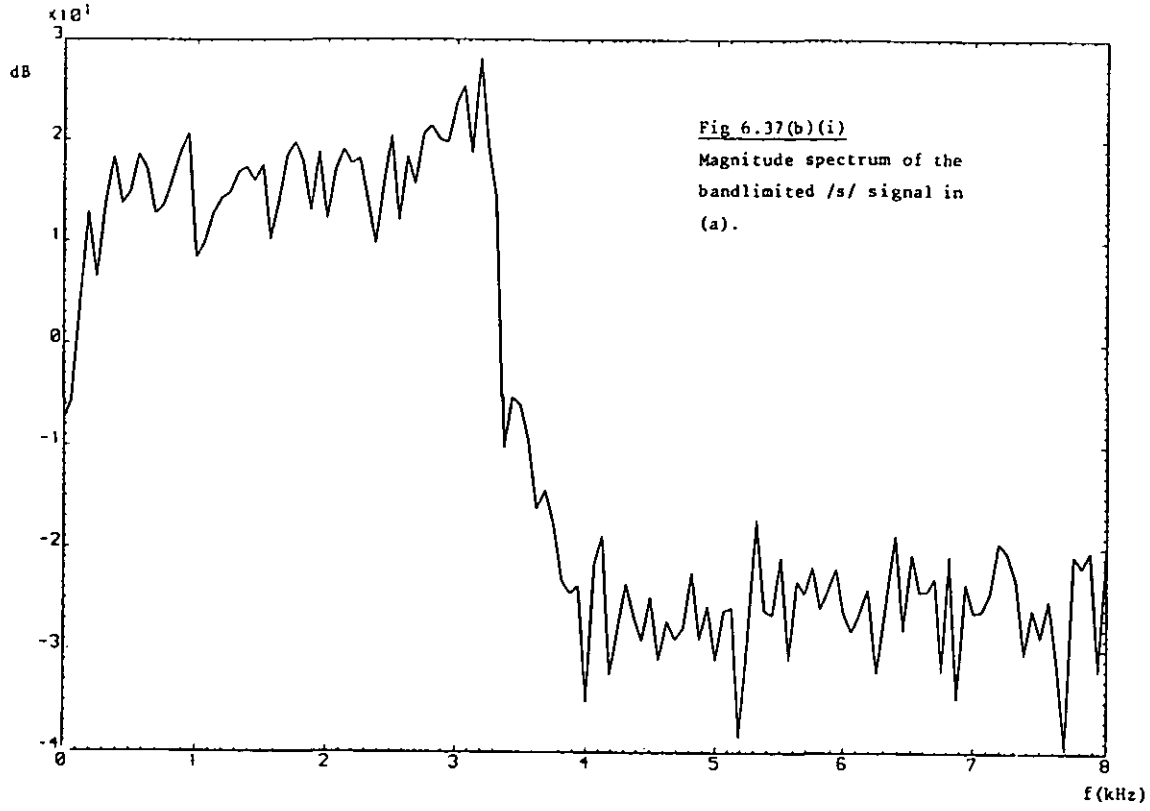


Fig 6.37(b)(i)
Magnitude spectrum of the bandlimited /s/ signal in (a).

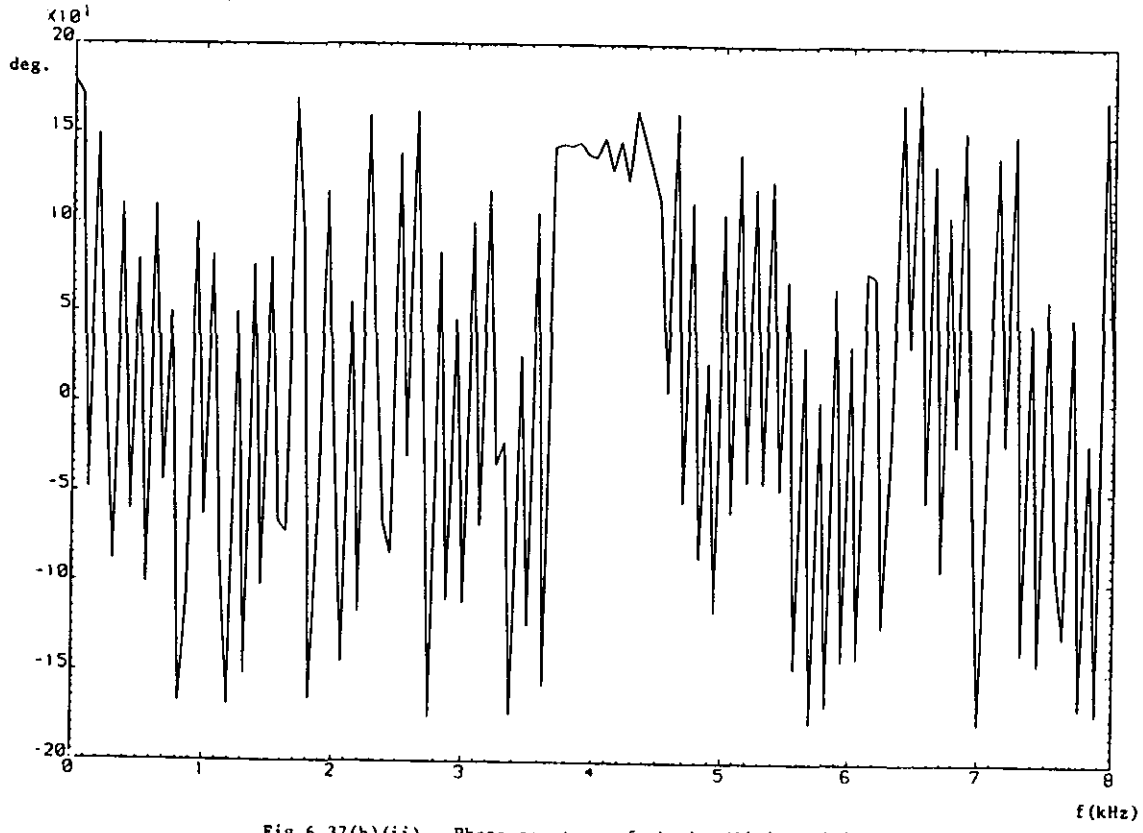


Fig 6.37(b)(ii) Phase spectrum of the bandlimited /s/ signal.

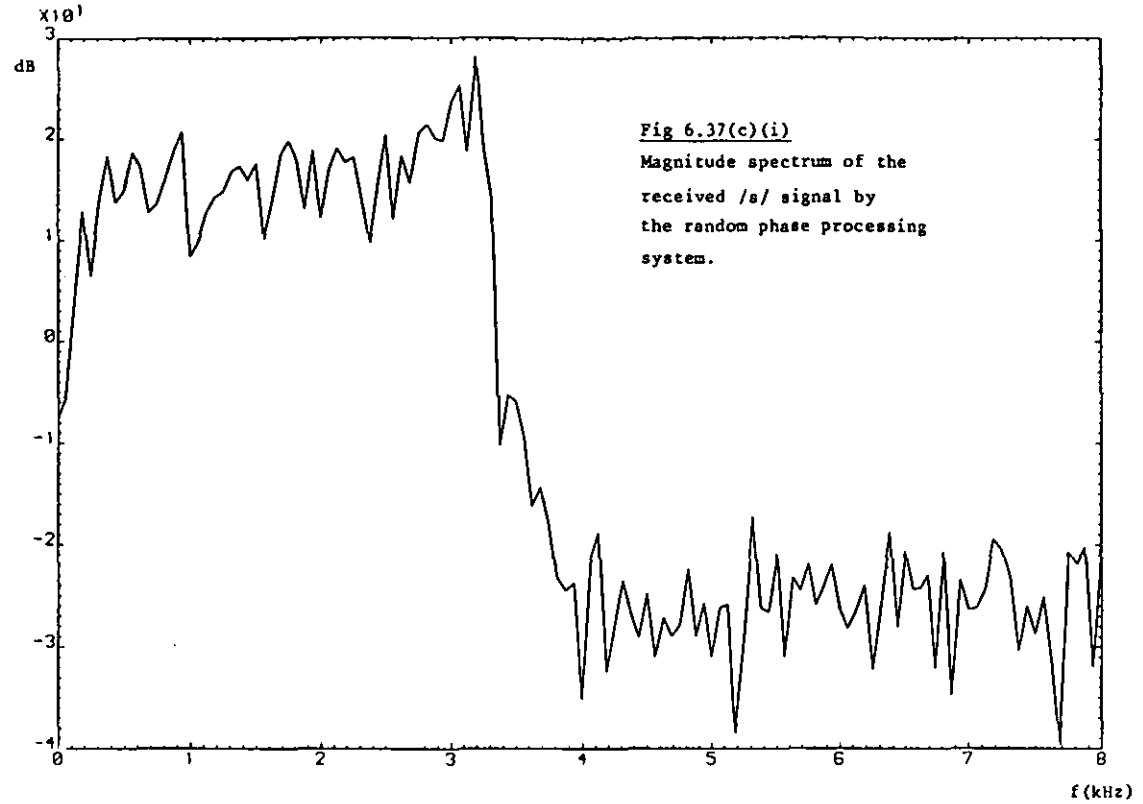


Fig 6.37(c)(i)
Magnitude spectrum of the received /s/ signal by the random phase processing system.

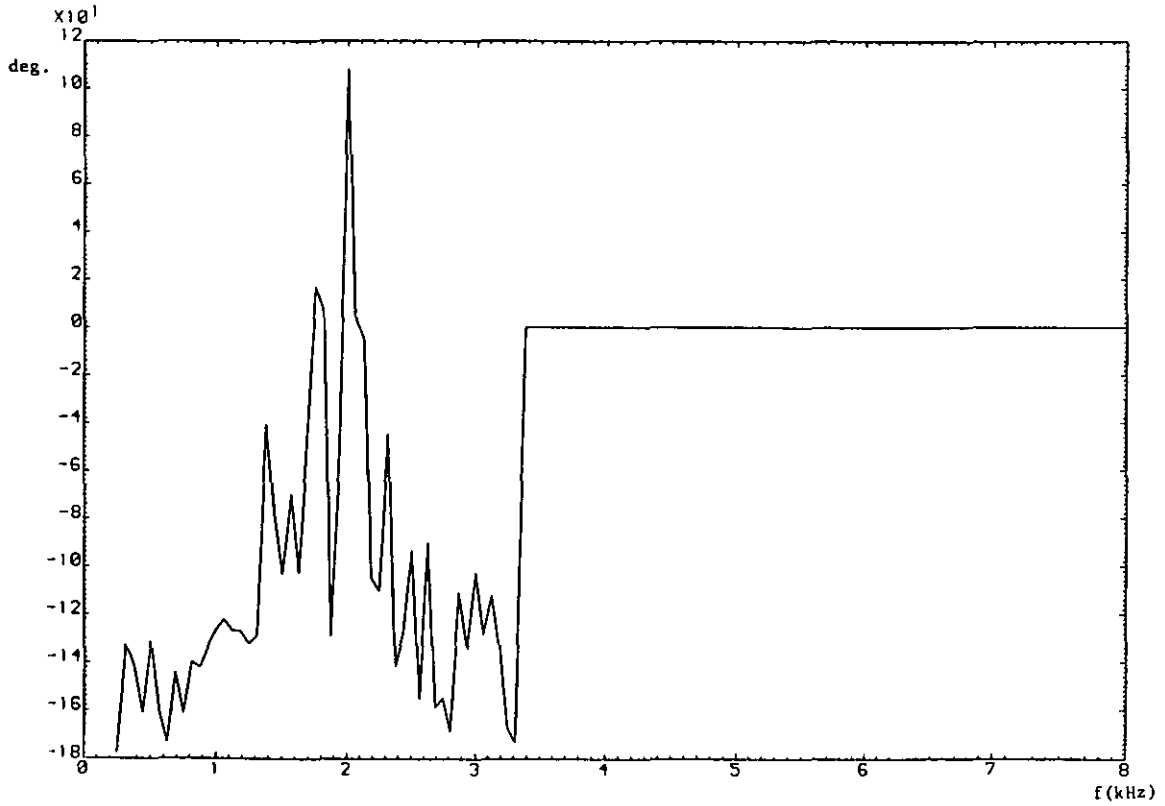


Fig 6.37(c)(ii) Phase spectrum of the received /s/ signal (after sign de-randomisation).

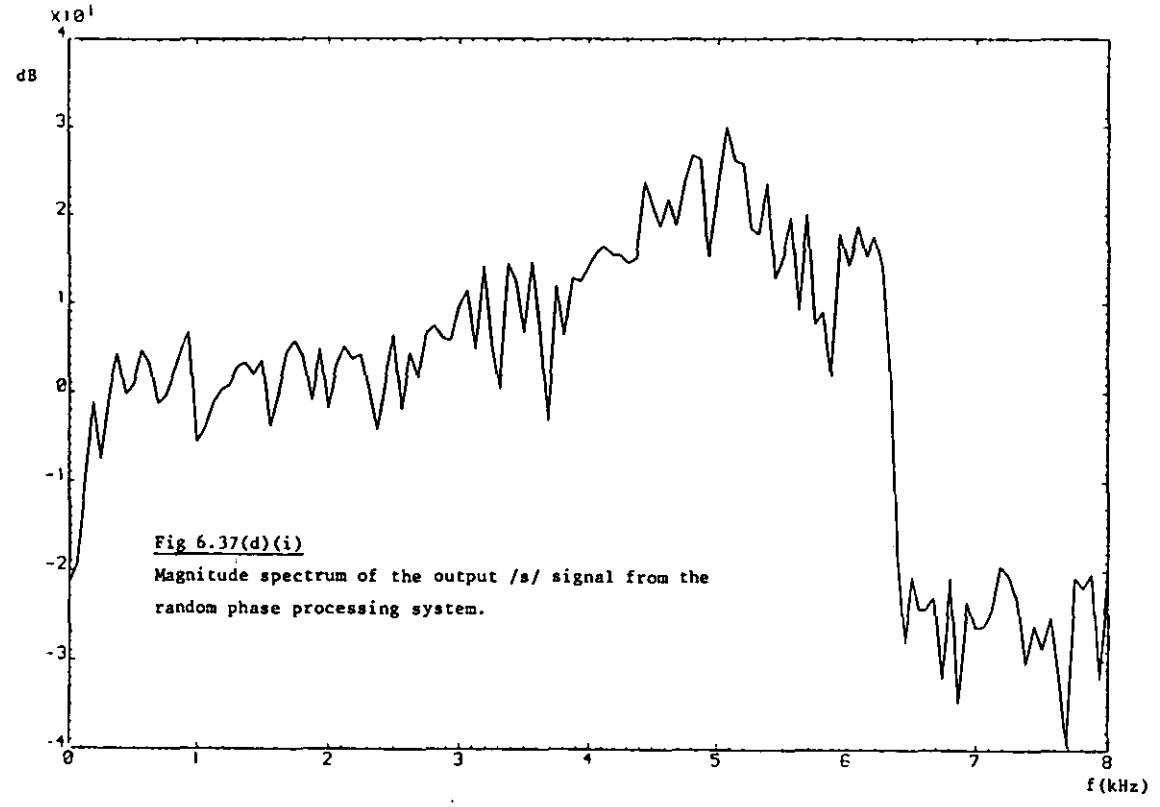


Fig 6.37(d)(i)
Magnitude spectrum of the output /s/ signal from the random phase processing system.

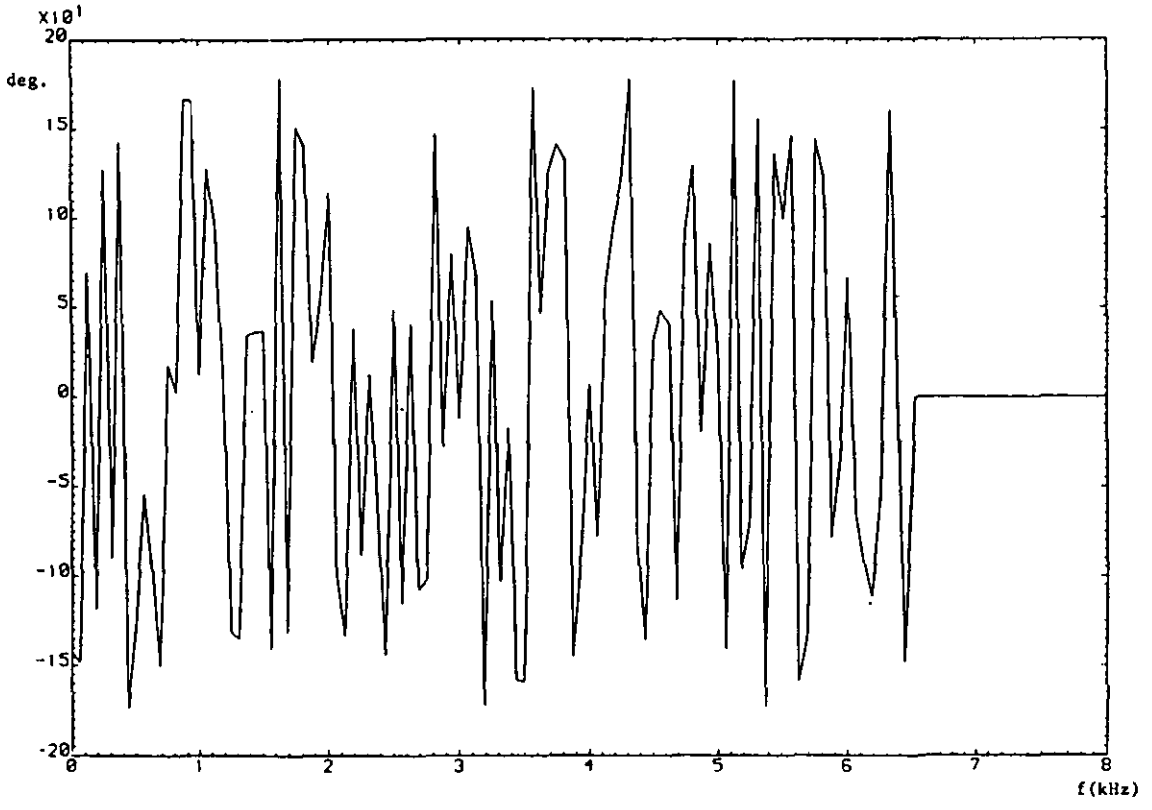


Fig 6.37(d)(ii) Random phase spectrum generated for the output /s/ signal.

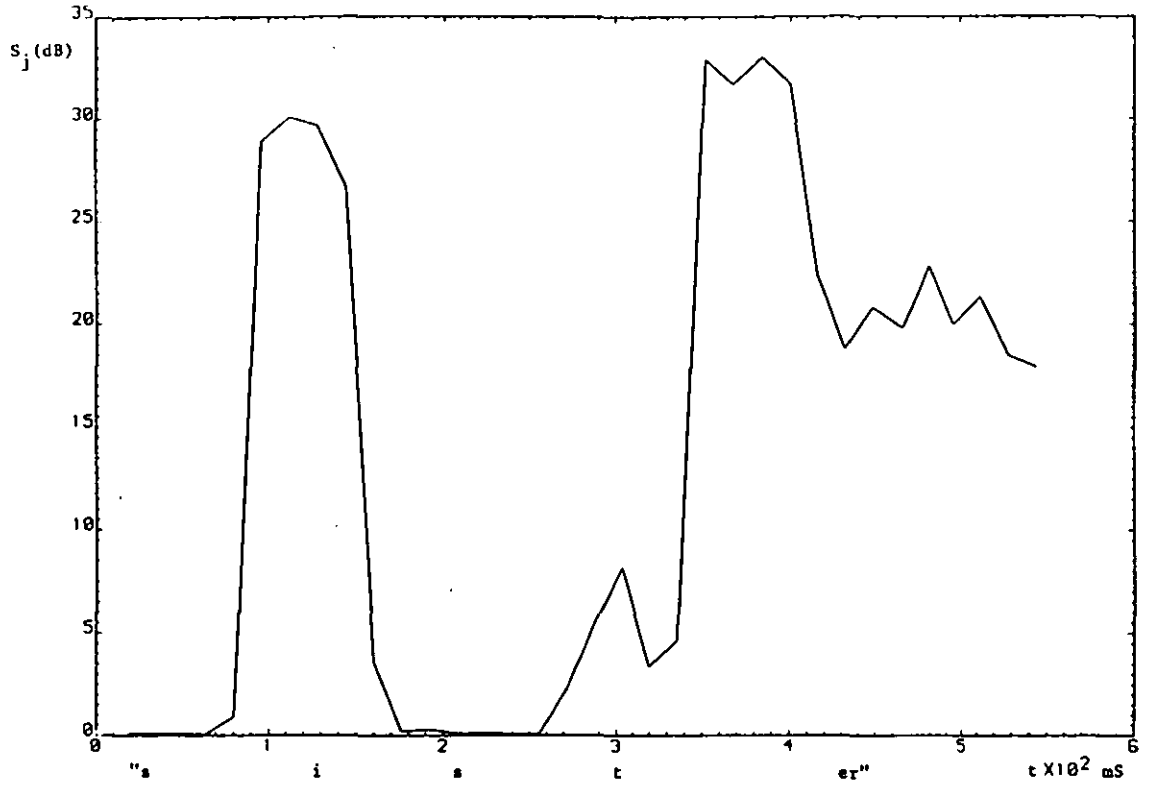


Fig 6.38(a) Spectral-SNR for a segment of bandlimited speech.

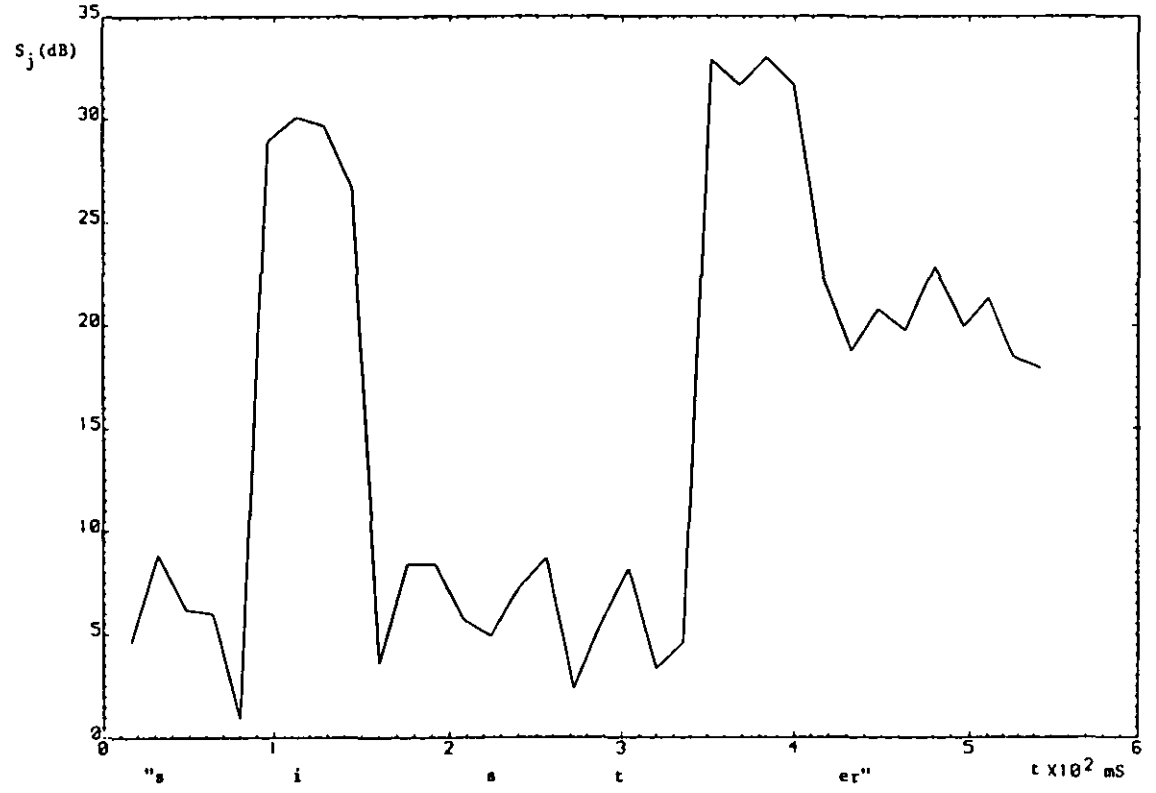


Fig 6.38(b) Spectral-SNR for a segment of phase processed speech.

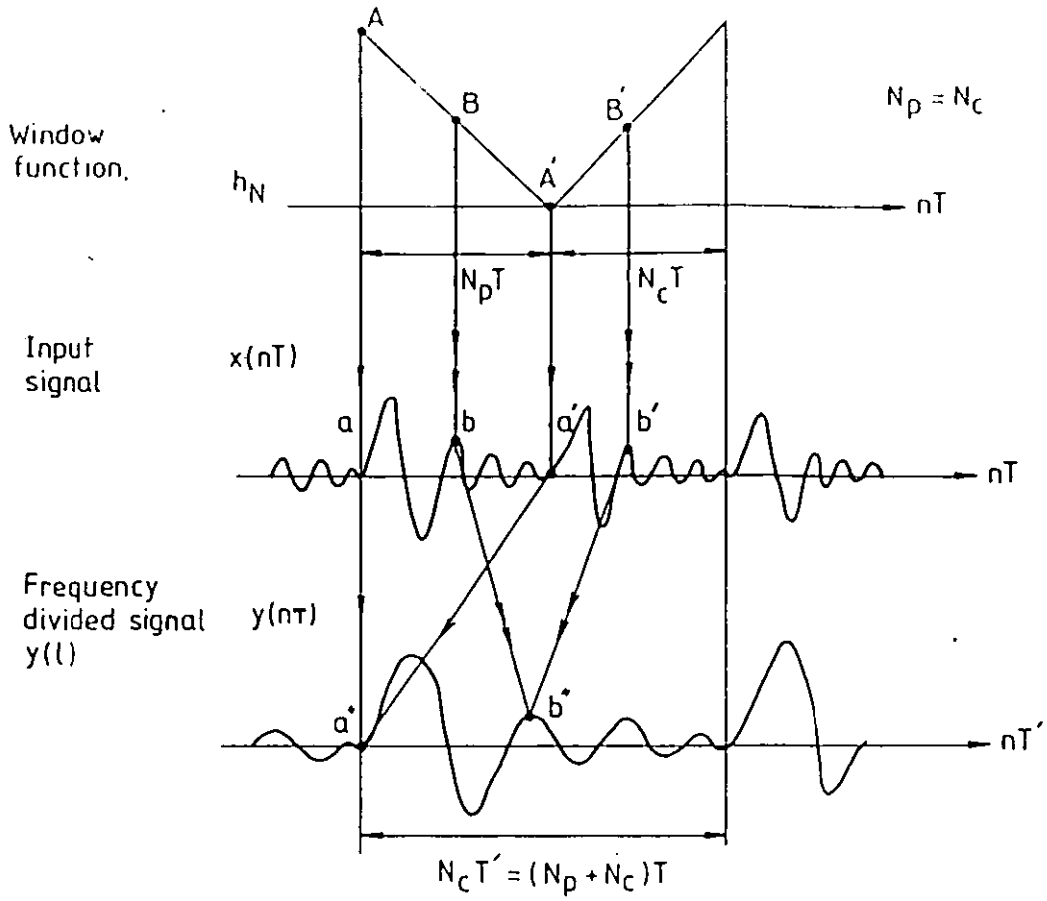


FIG. 6·39 THE TDHC PROCESS (REF 61)

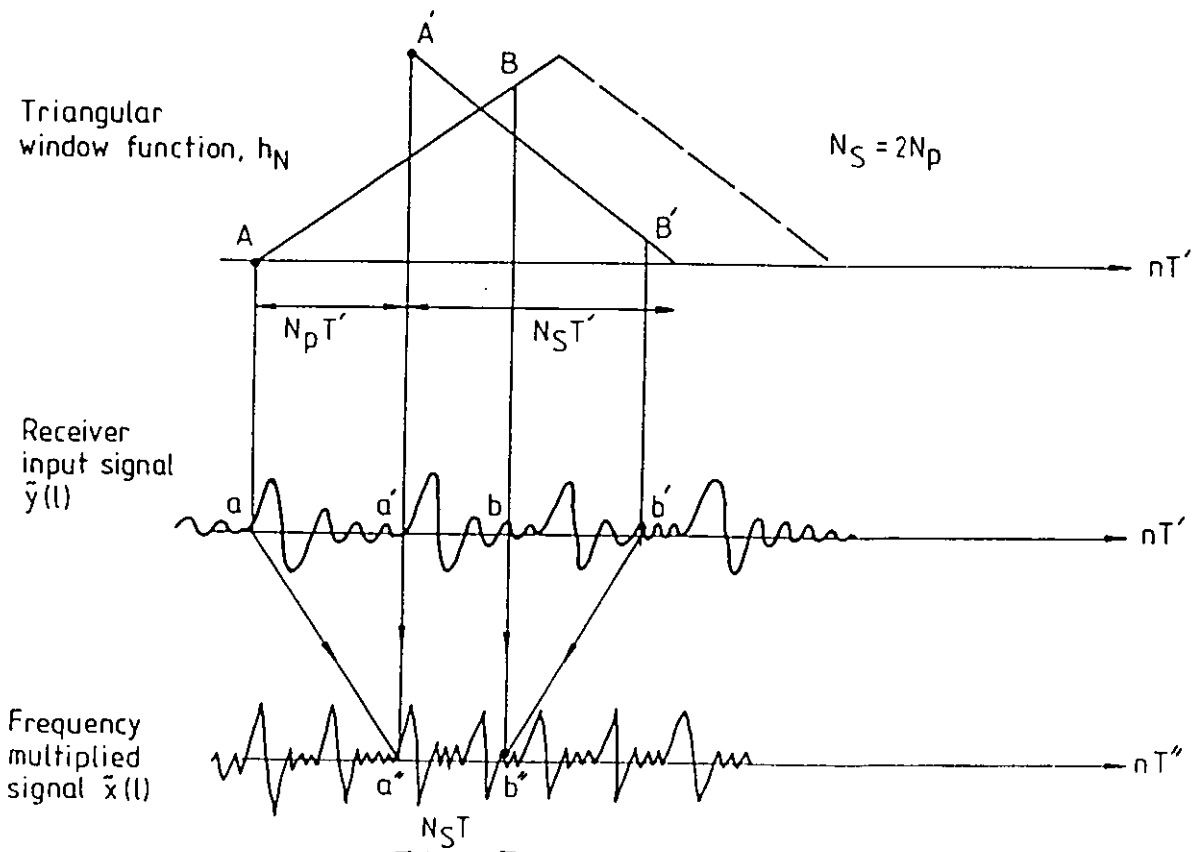


FIG. 6·40 THE TDHE PROCESS (REF 61)

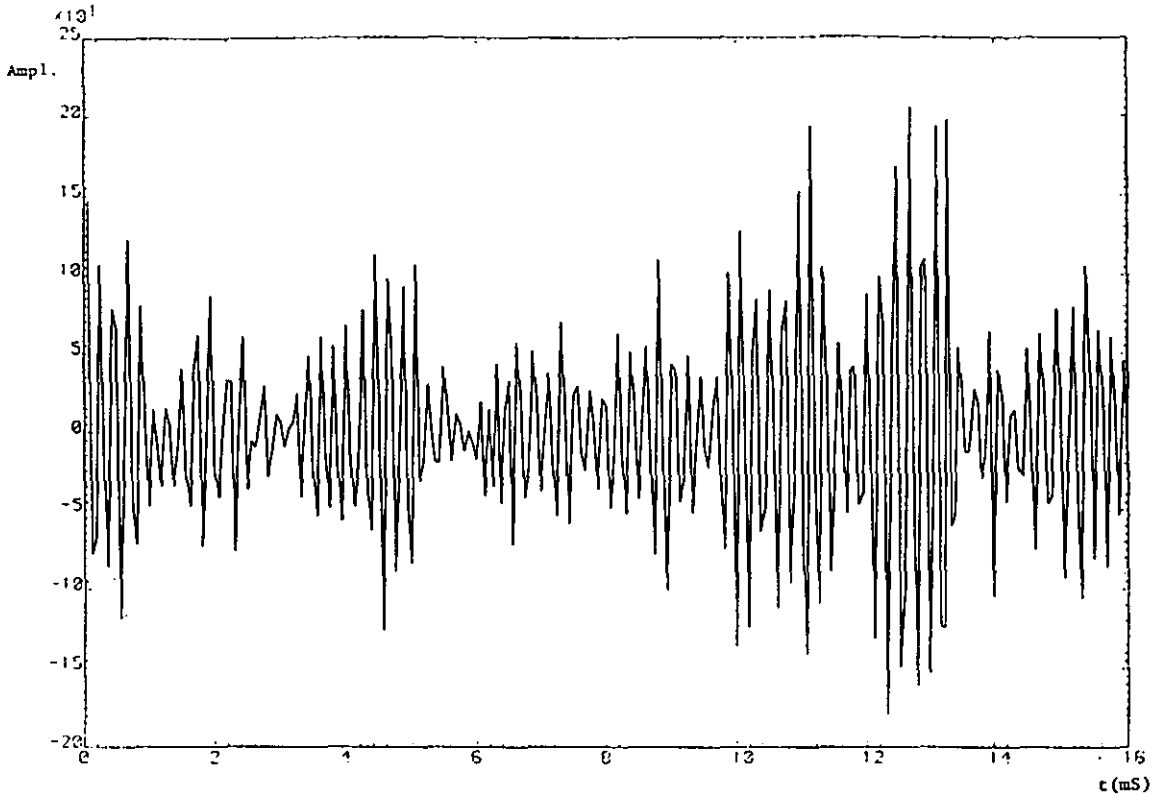


Fig 6.41(a) Input waveform of an /s/ signal from "sister".

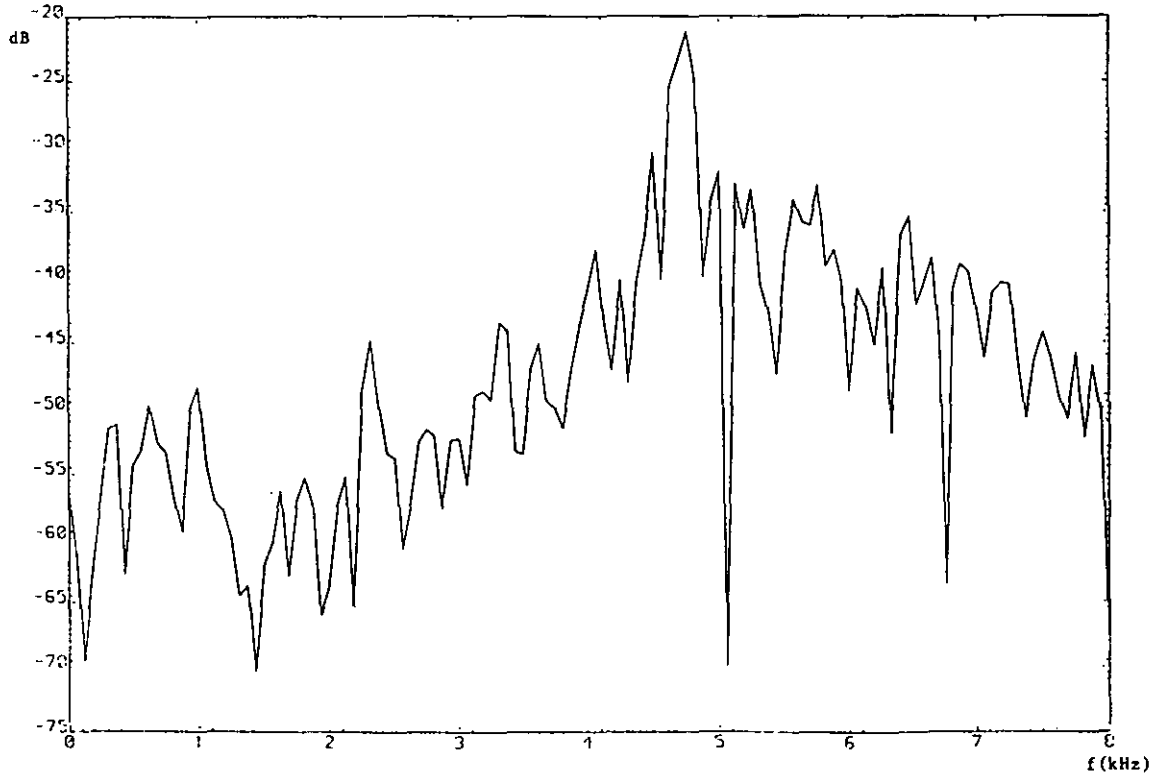


Fig 6.41(b) Spectrum of the input signal in (a).

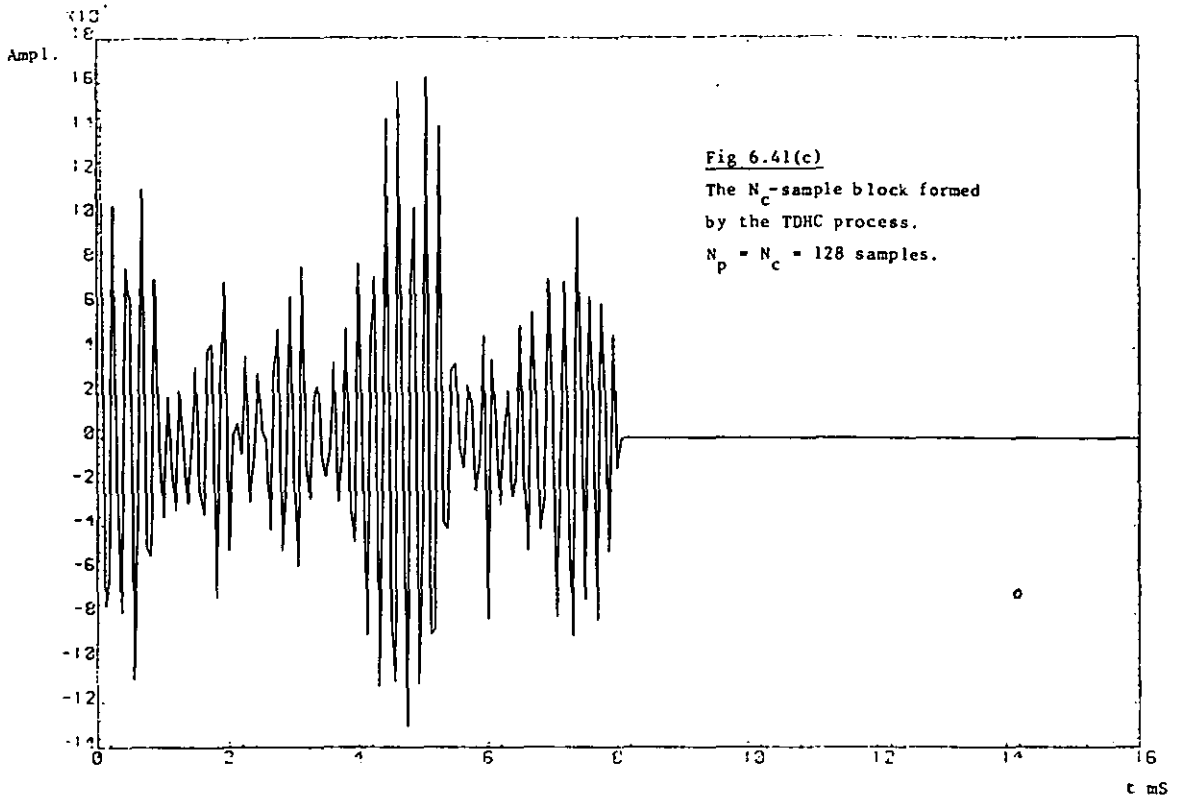


Fig 6.41(c)
The N_c-sample block formed by the TDHC process.
$N_p = N_c = 128$ samples.

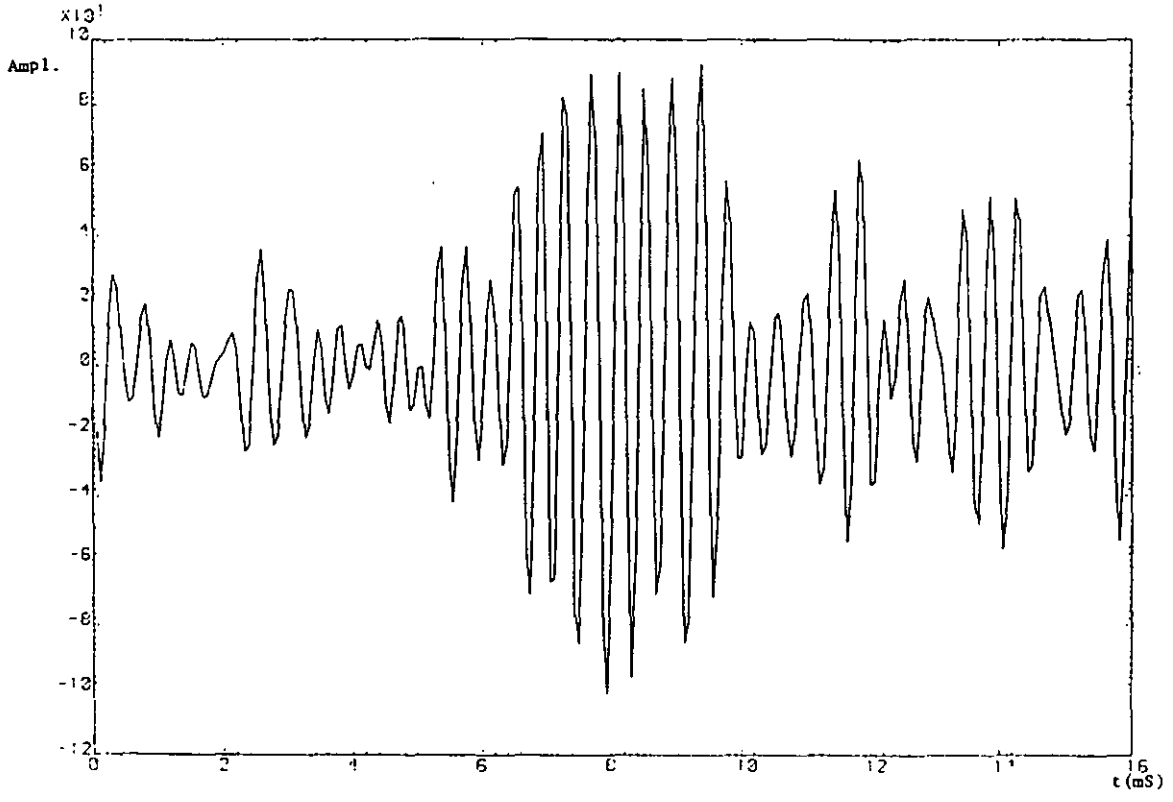


Fig 6.41(d) The 16 mS-block formed by doubling the sampling interval of the waveform of (c) and filtering from 300 Hz to 3400 Hz by an FIR filter.

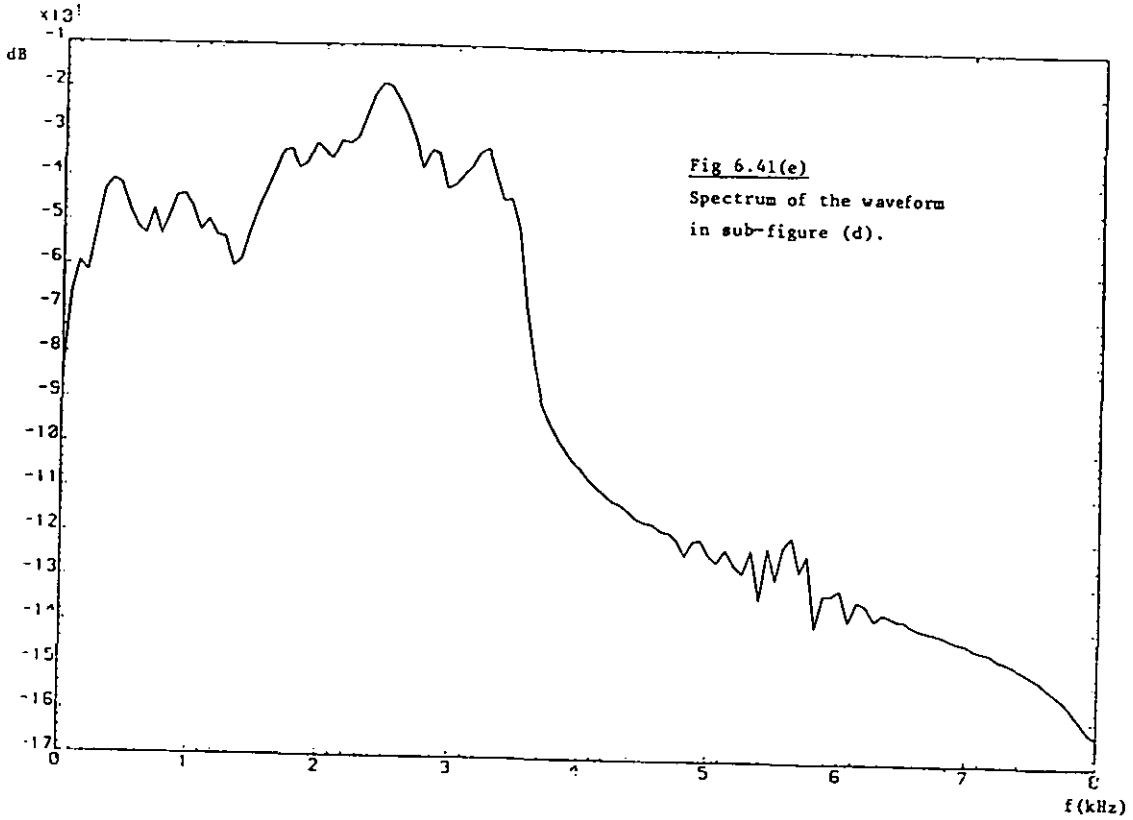


Fig 6.41(e)
Spectrum of the waveform in sub-figure (d).

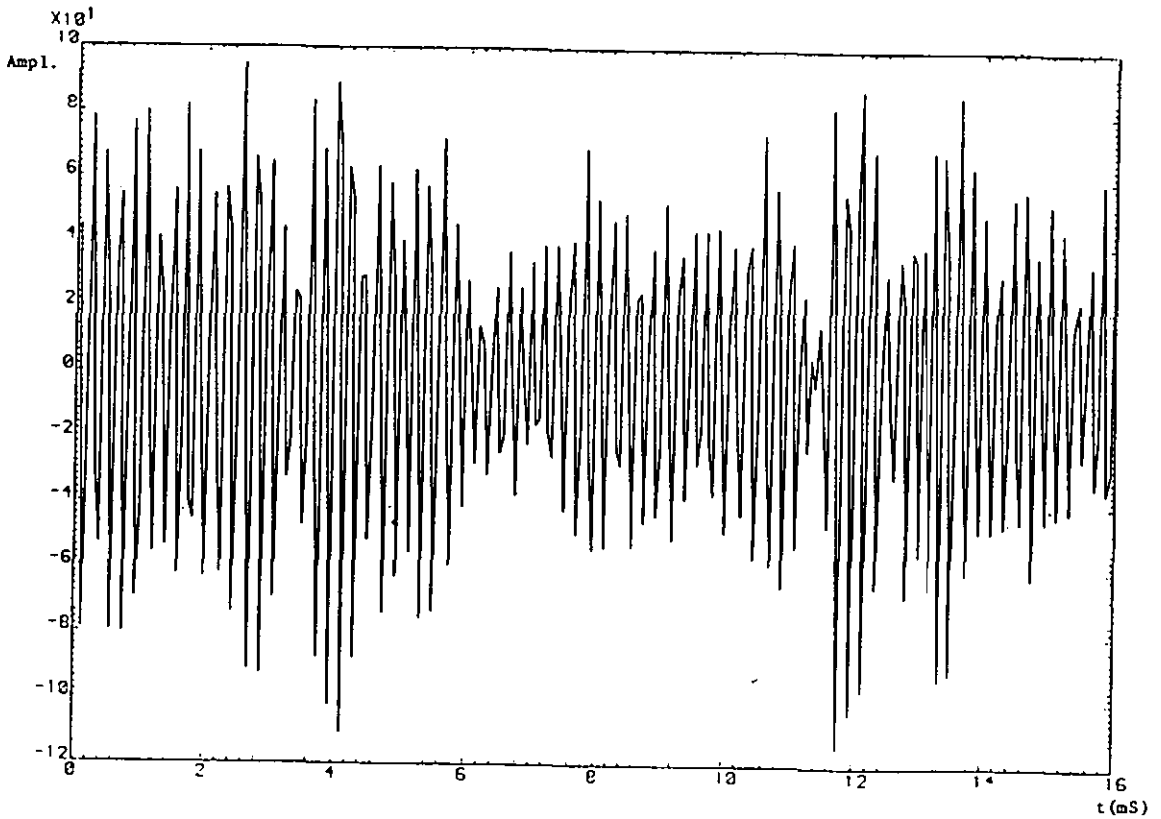


Fig 6.41(f) Output waveform from the TDHE process, N_p = 128 and N_s = 256 samples.

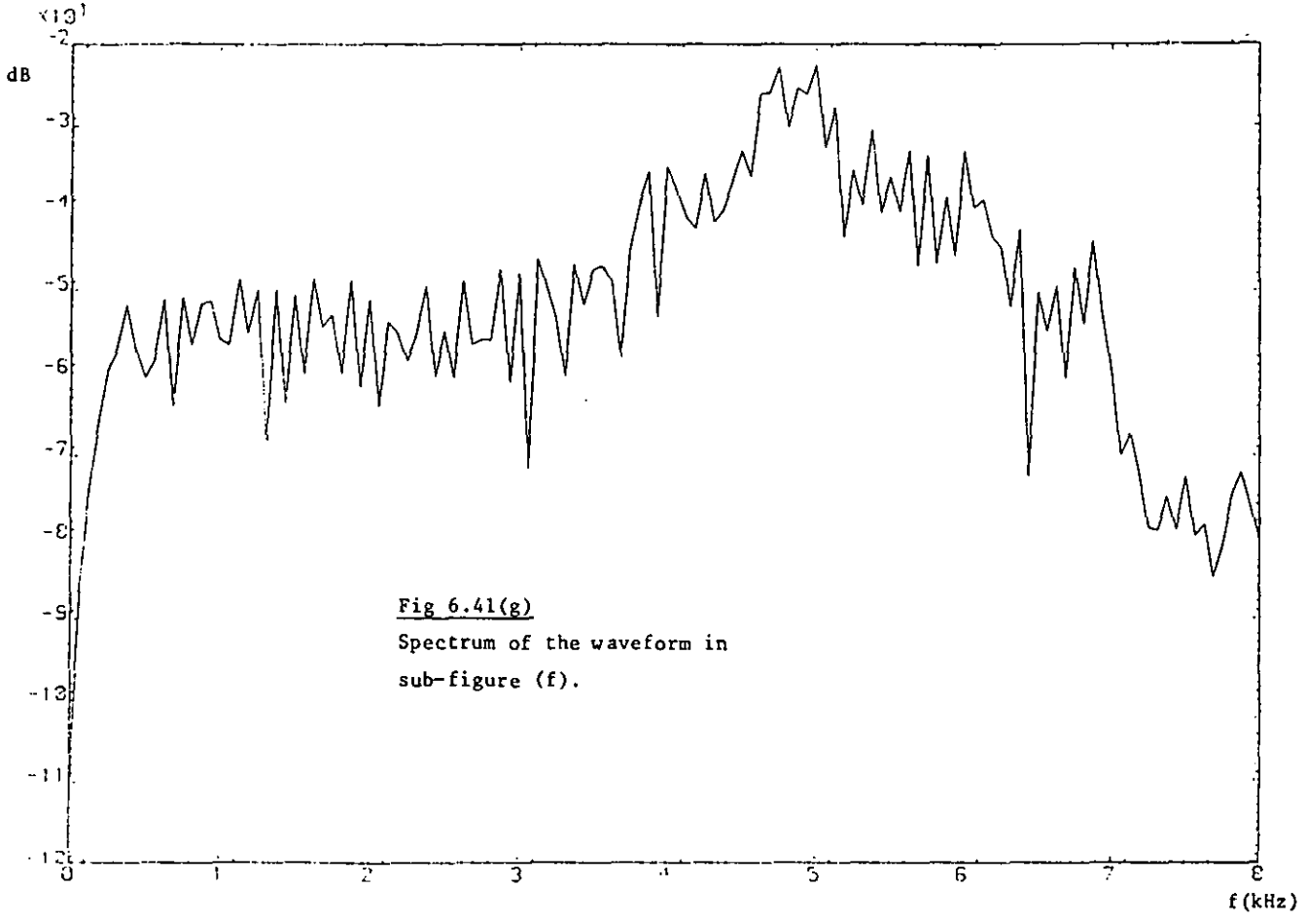


Fig 6.41(g)
Spectrum of the waveform in sub-figure (f).

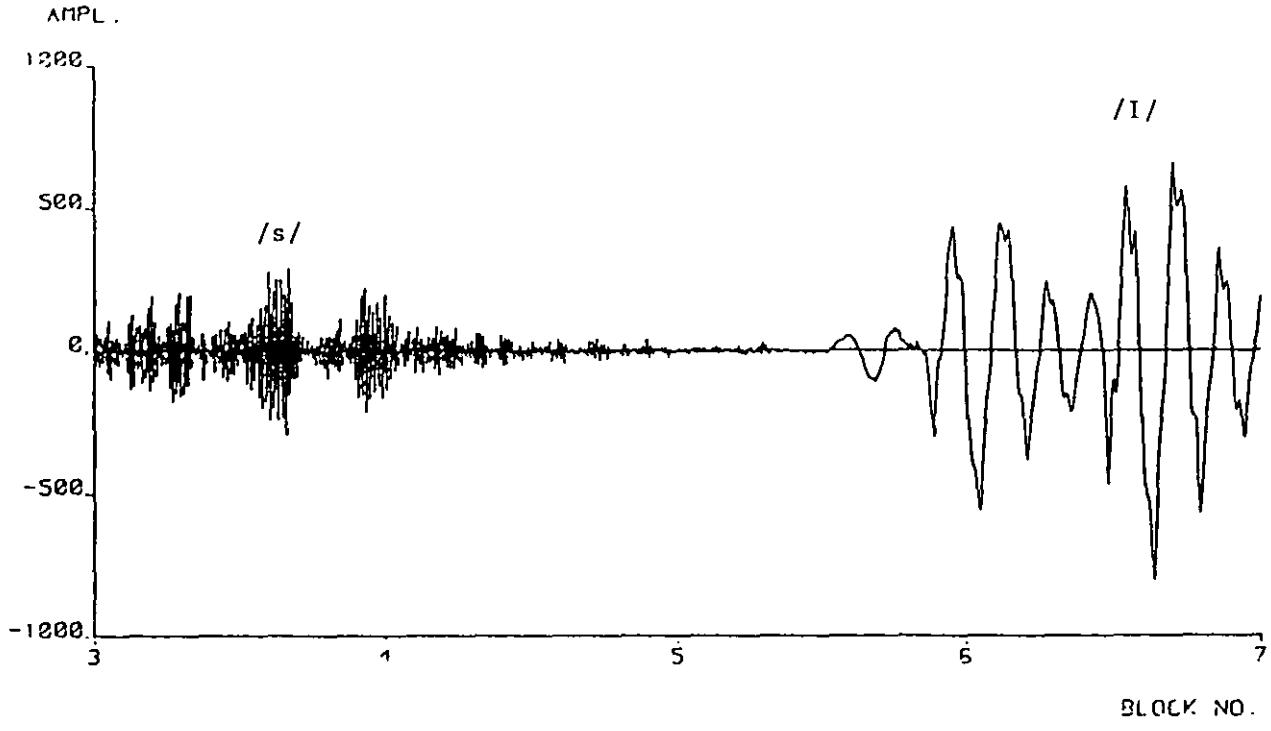


Fig 6.42(a) Original signal, 0.3-7.6 kHz.

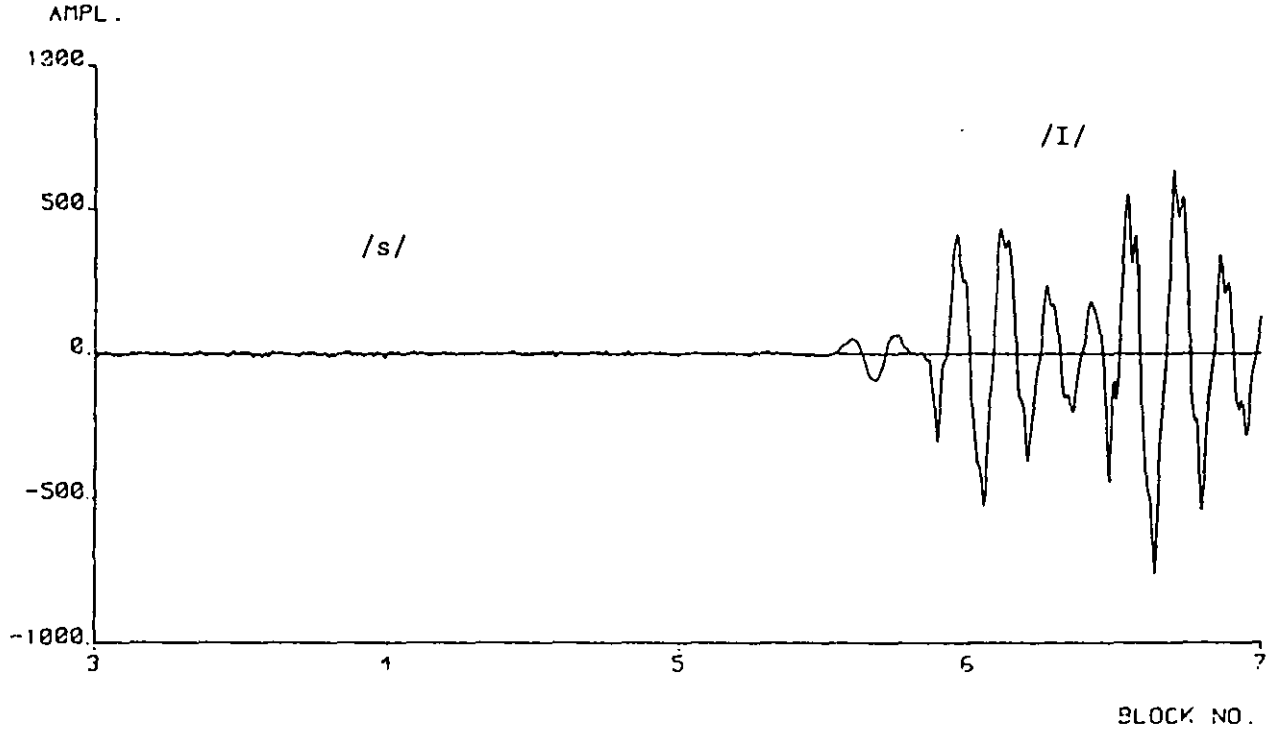


Fig 6.42(b) Bandpass filtered signal, 0.3-3.4 kHz.

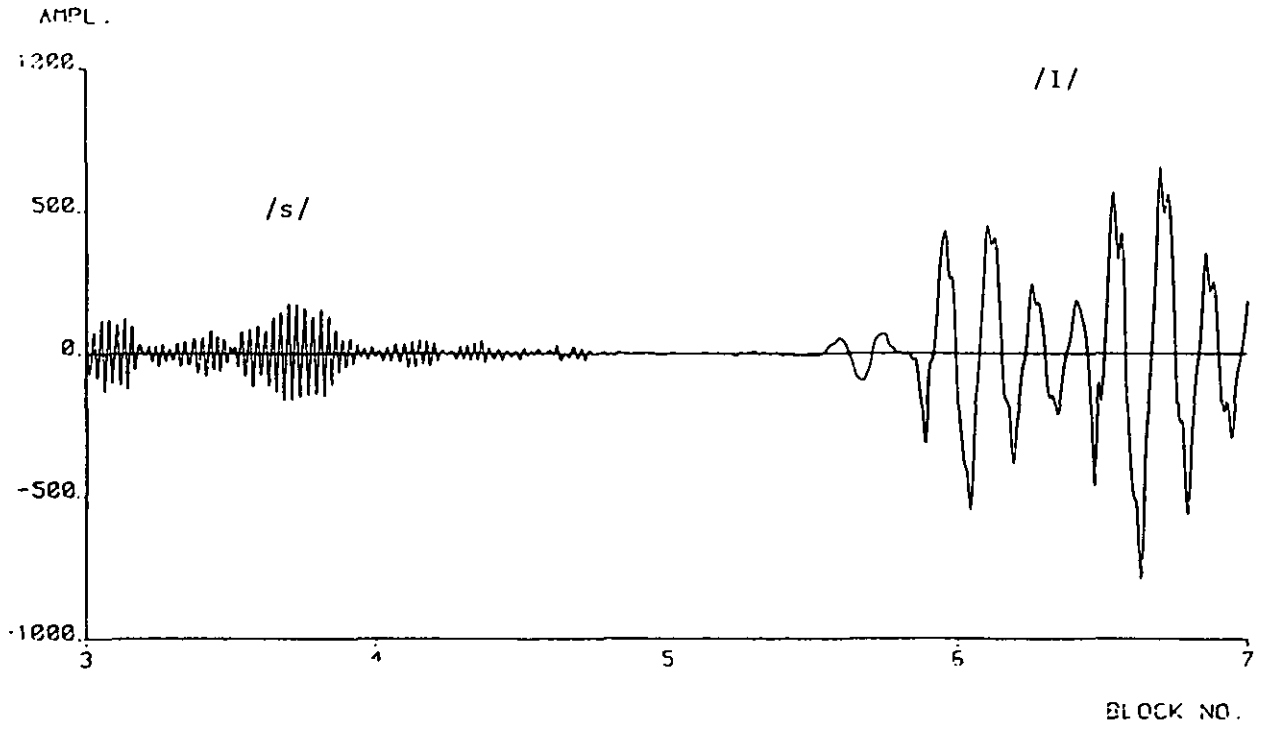


Fig 6.42(c) TDHC processed signal, 0.3-3.4 kHz.

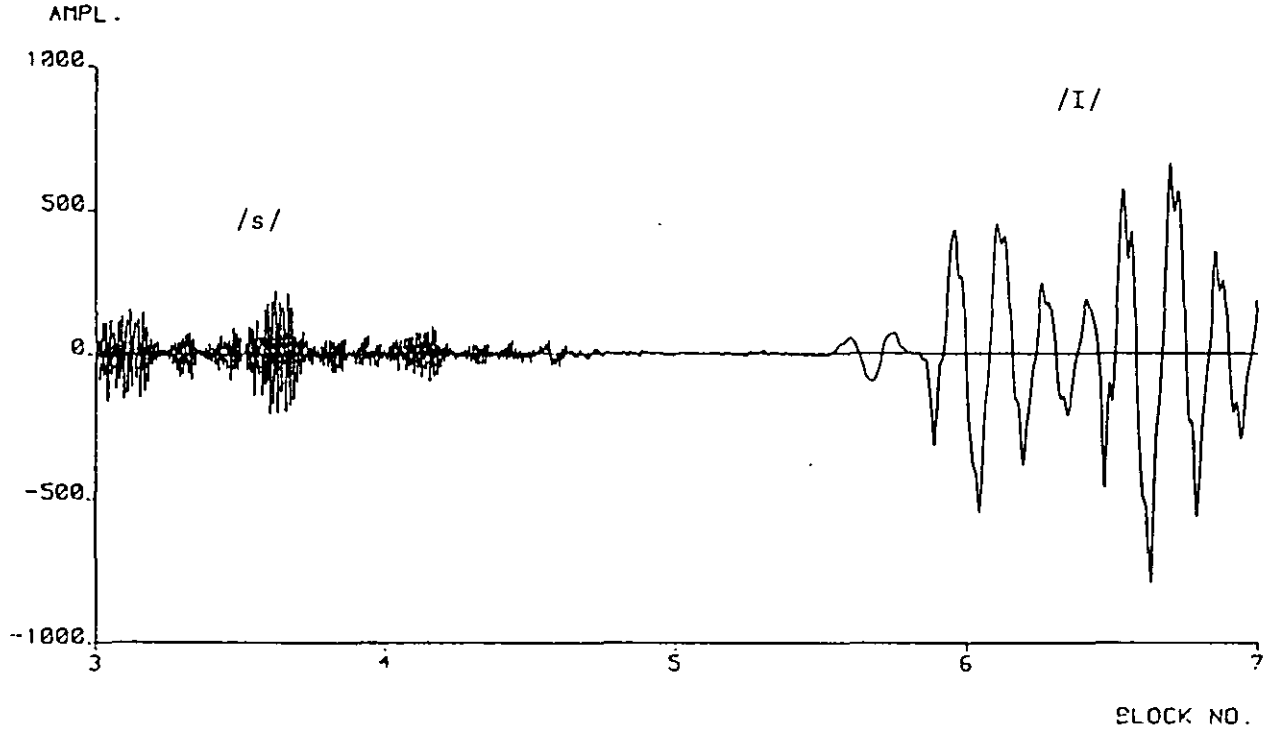
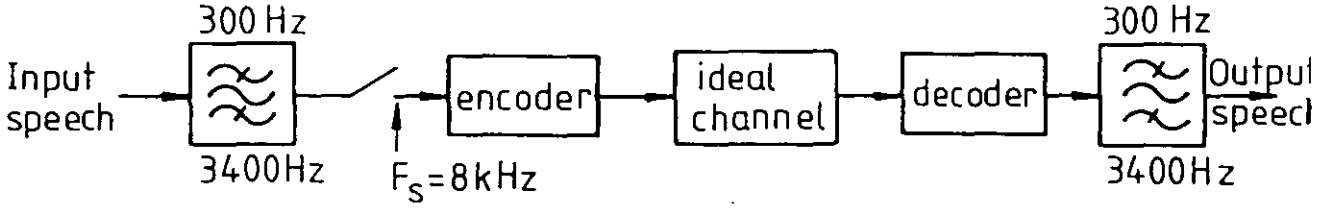
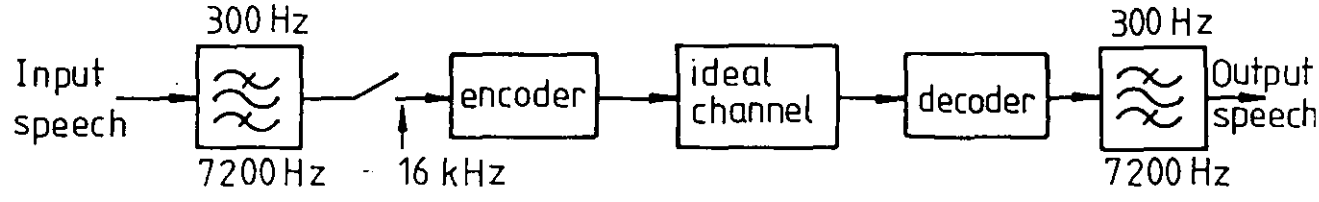


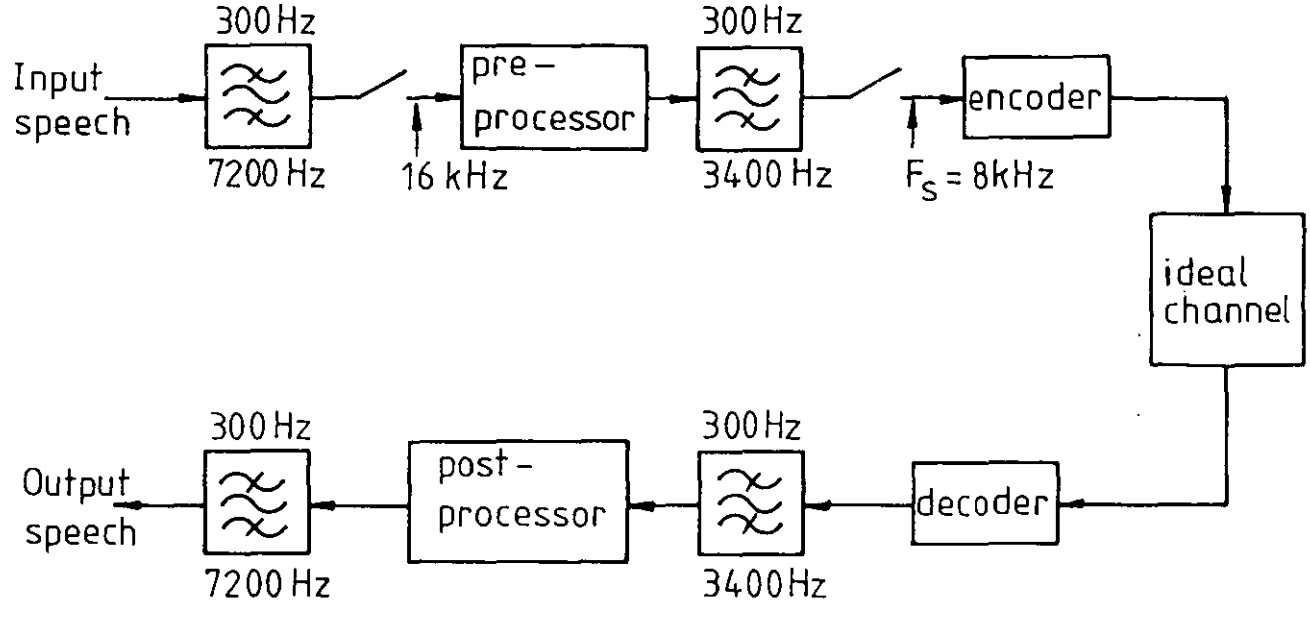
Fig 6.42(d) TDHE processed output signal, 0.3-6.8 kHz.



(a) INPUT SPEECH BANDLIMITED TO 3400 Hz



(b) INPUT SPEECH BANDLIMITED TO 7200 Hz



(c) BANDWIDTH COMPRESSED INPUT SPEECH
(FROM 0·3 – 7·2 kHz INTO 0·3 – 3·4 kHz)

FIG. 6·43 (REF 89)

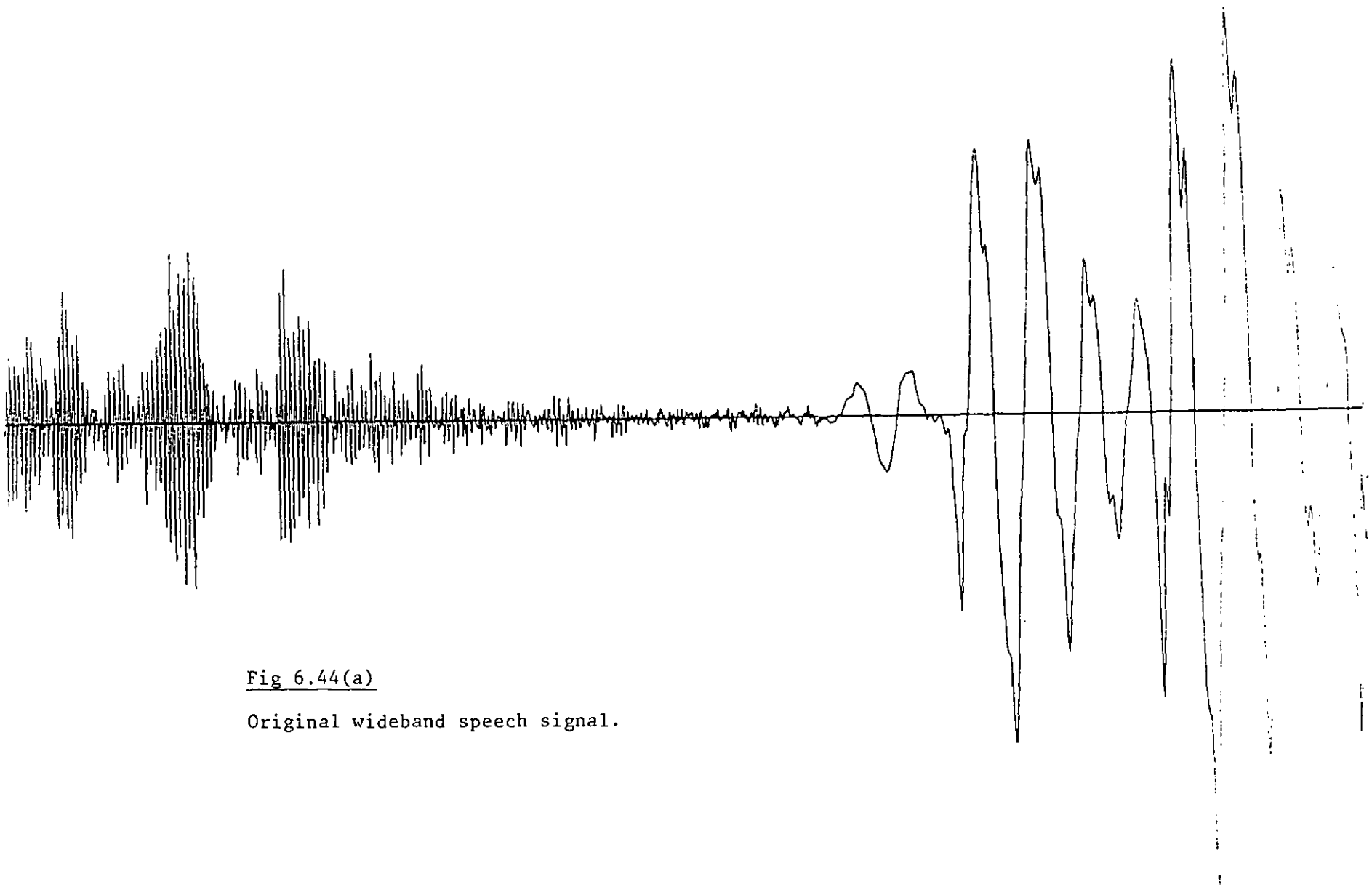


Fig 6.44(a)

Original wideband speech signal.

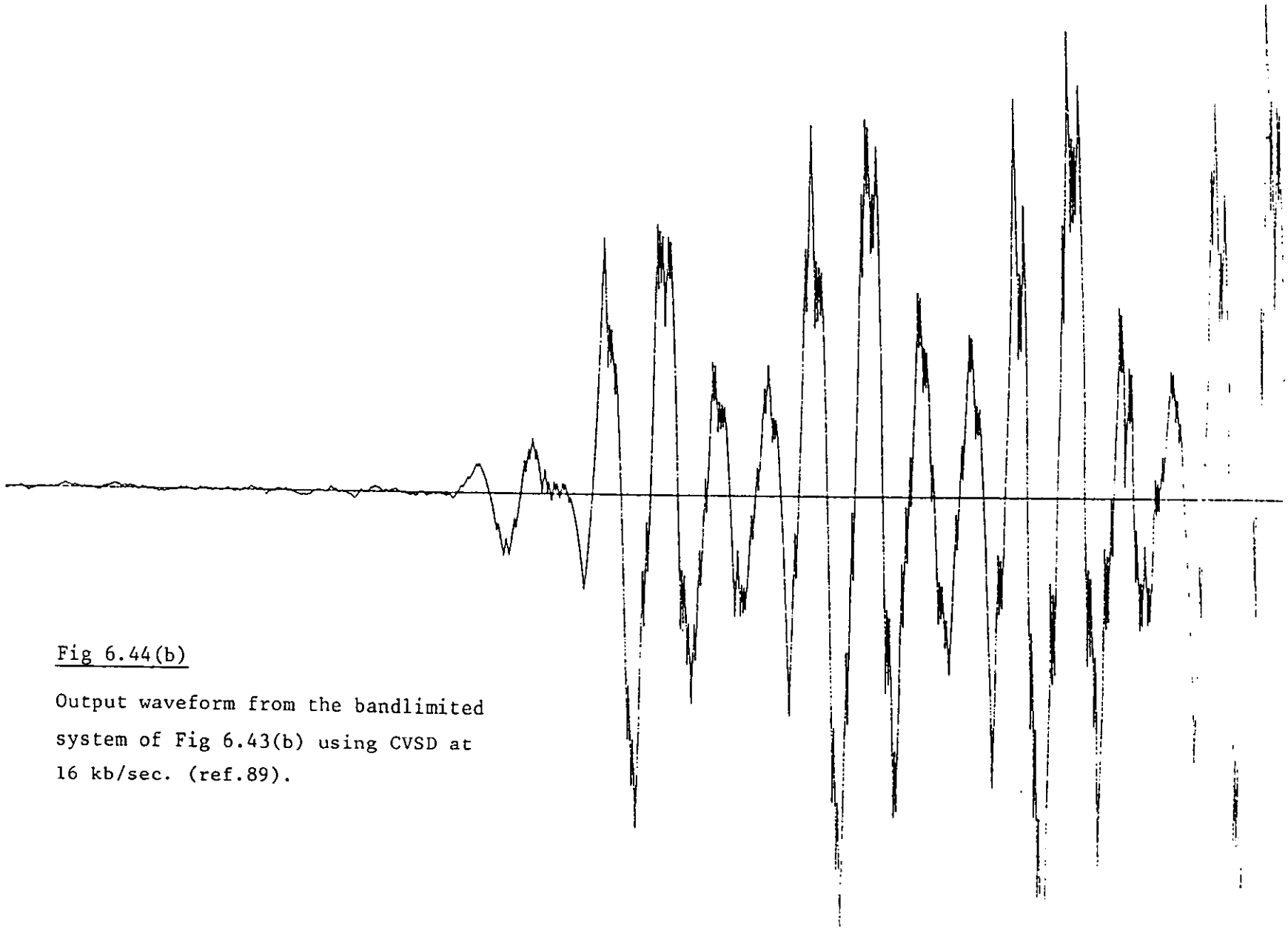


Fig 6.44(b)

Output waveform from the bandlimited system of Fig 6.43(b) using CVSD at 16 kb/sec. (ref.89).

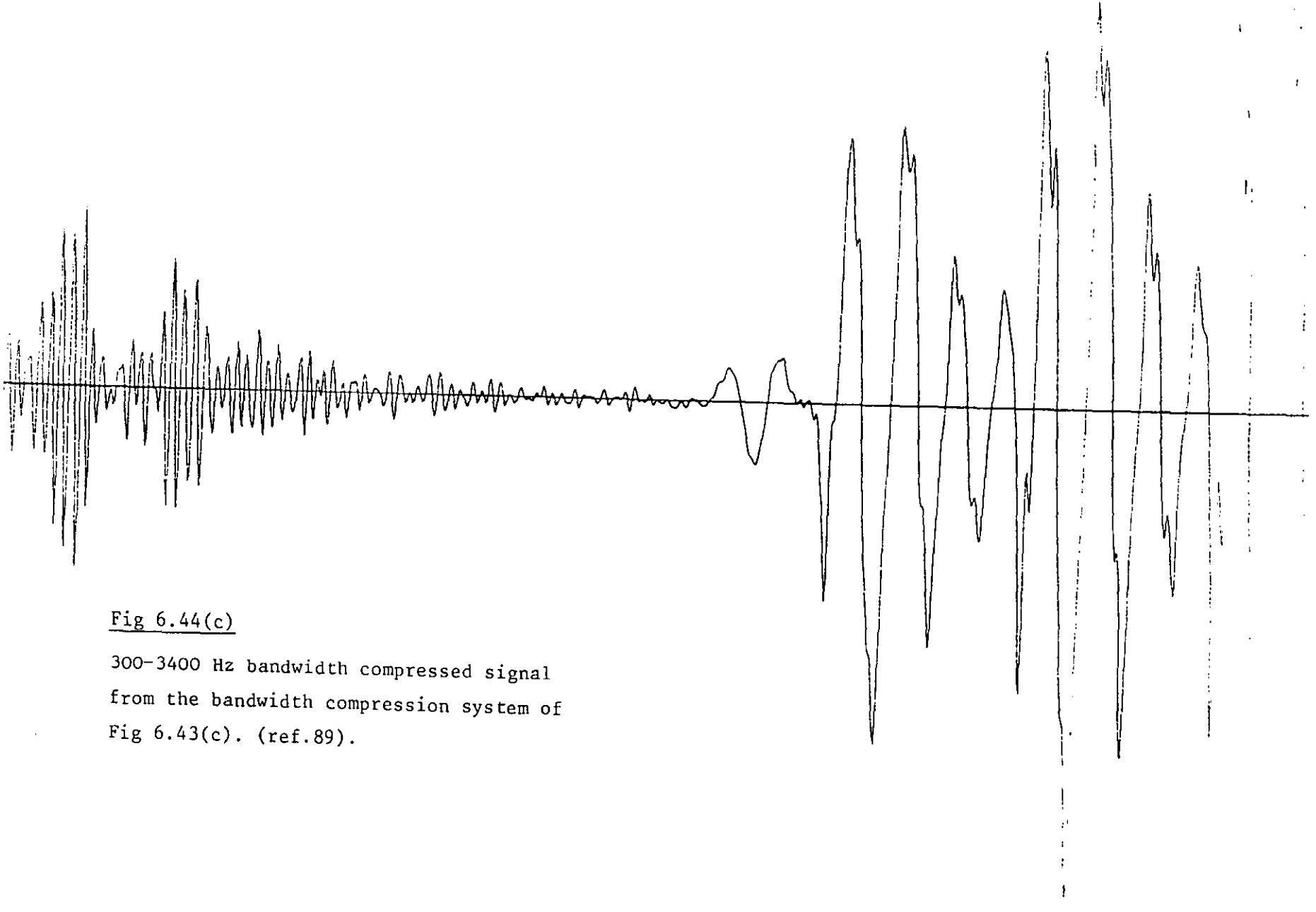


Fig 6.44(c)

300–3400 Hz bandwidth compressed signal from the bandwidth compression system of Fig 6.43(c). (ref.89).

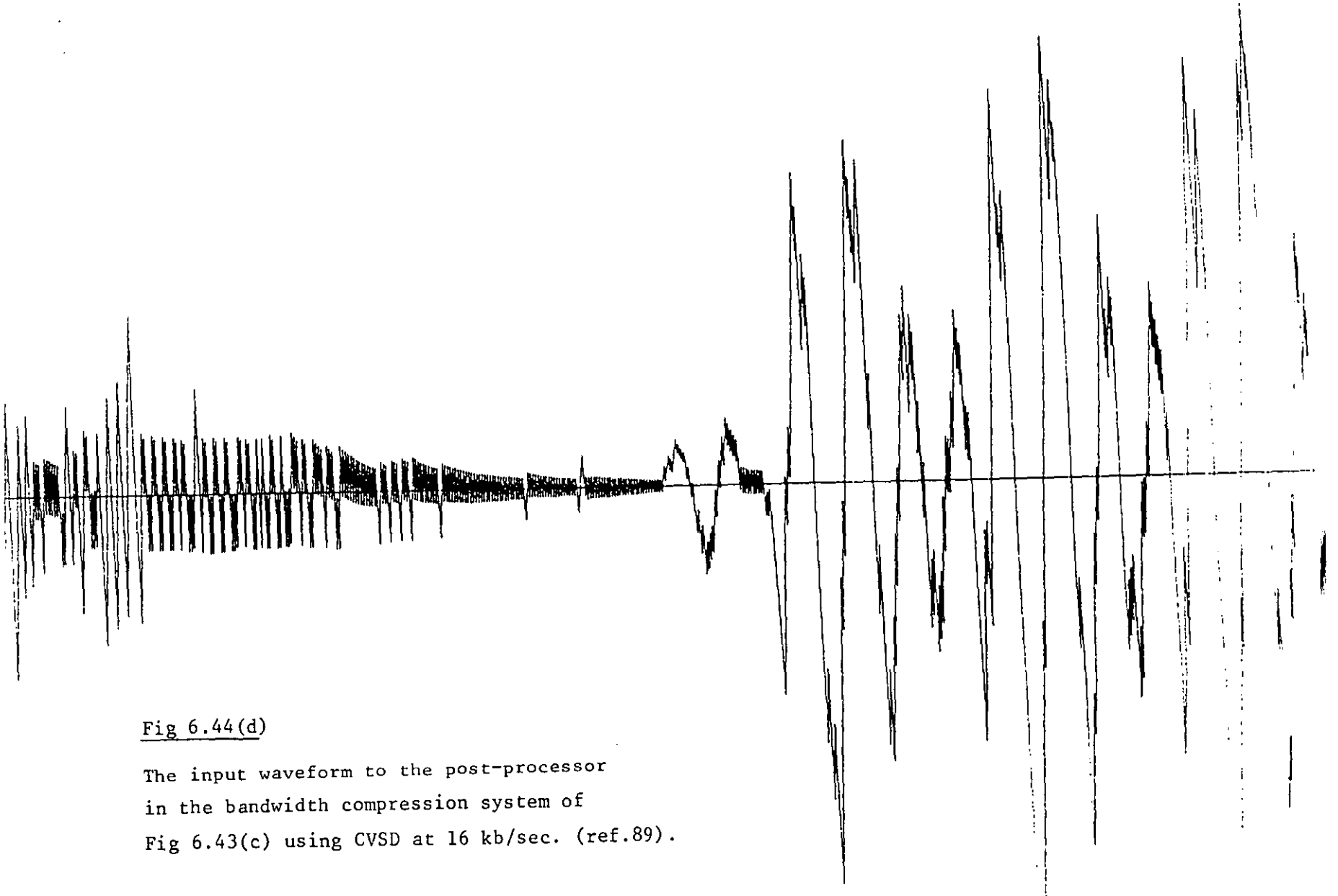


Fig 6.44(d)

The input waveform to the post-processor in the bandwidth compression system of Fig 6.43(c) using CVSD at 16 kb/sec. (ref.89).

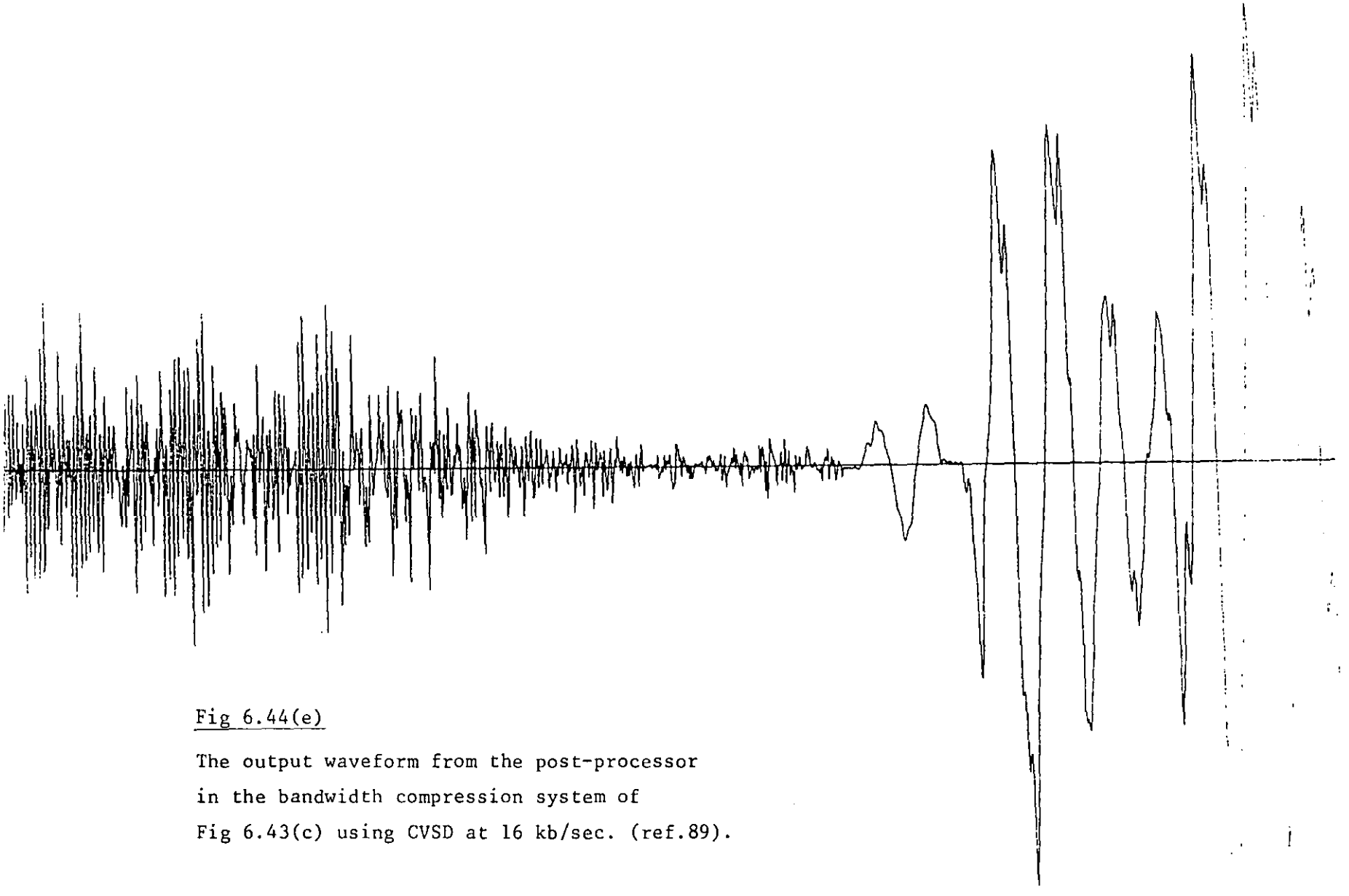


Fig 6.44(e)

The output waveform from the post-processor in the bandwidth compression system of Fig 6.43(c) using CVSD at 16 kb/sec. (ref.89).

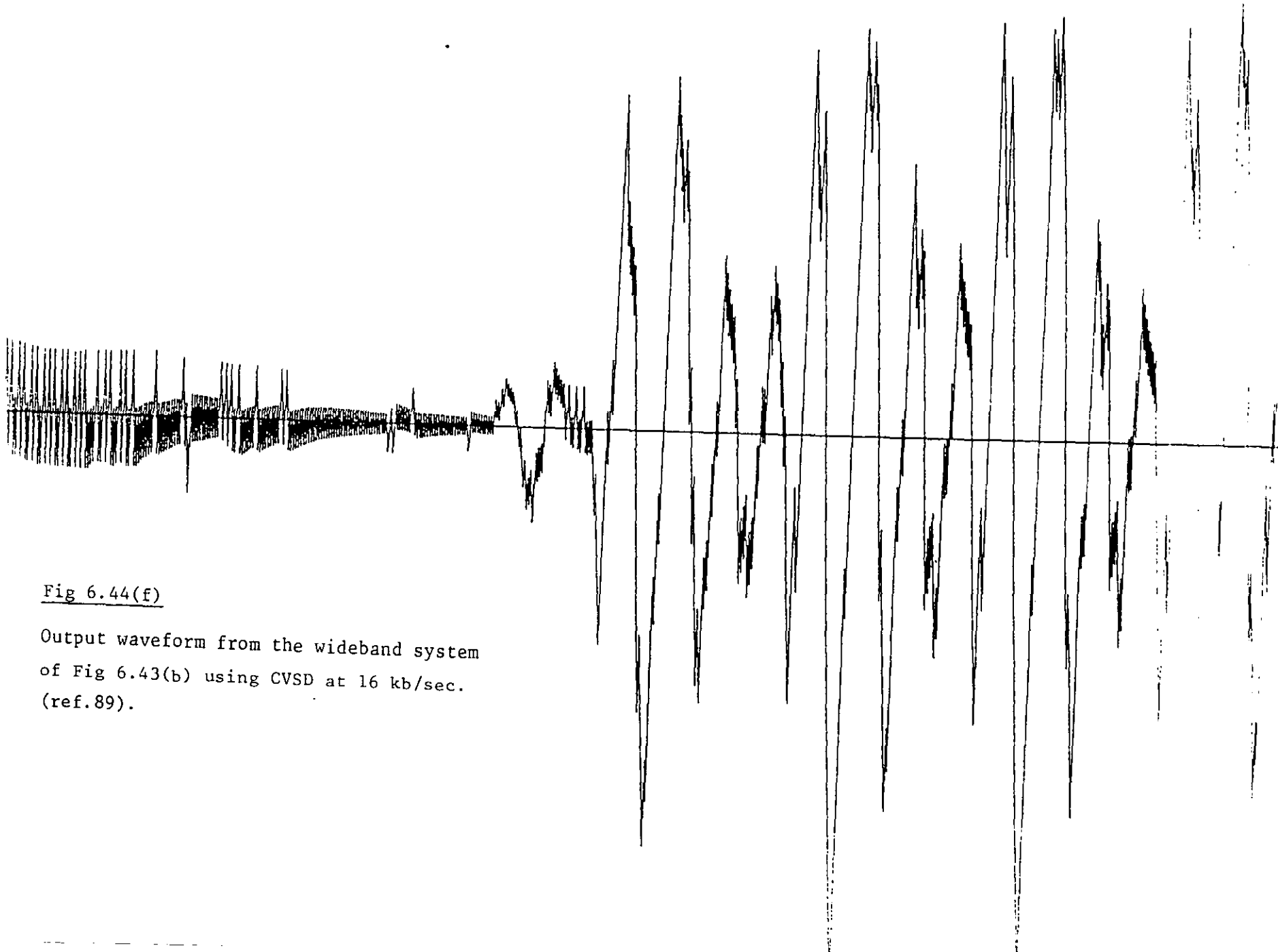


Fig 6.44(f)

Output waveform from the wideband system of Fig 6.43(b) using CVSD at 16 kb/sec. (ref.89).

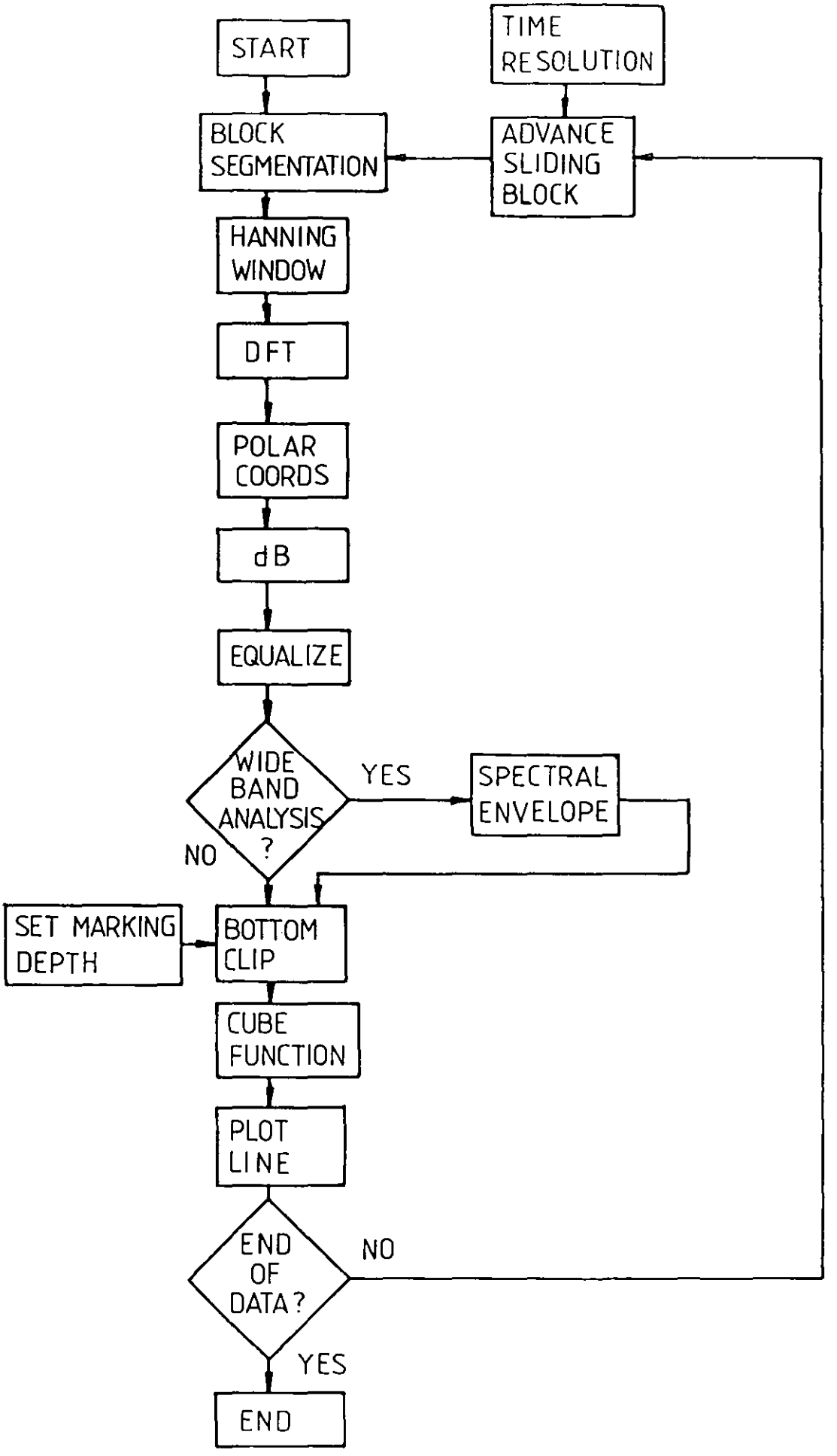


FIG. 7·1 STRATEGY FOR OBTAINING AN APPROXIMATE SPECTROGRAPHIC DISPLAY ON THE LINE PRINTER

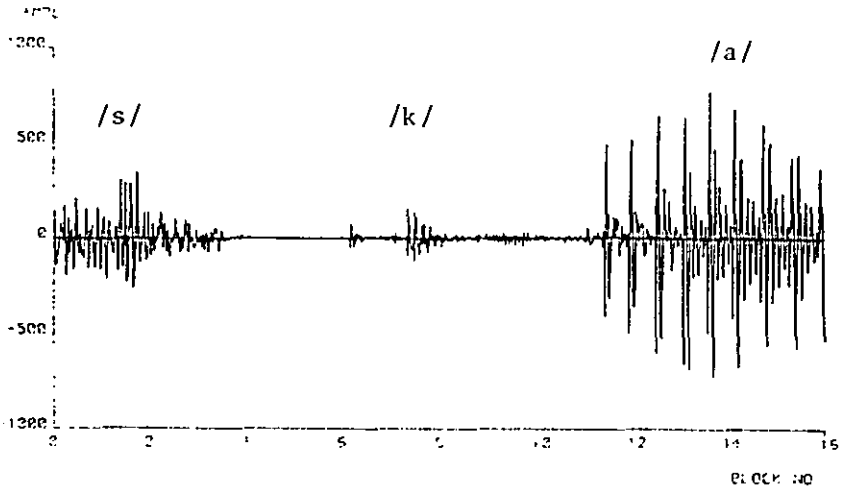
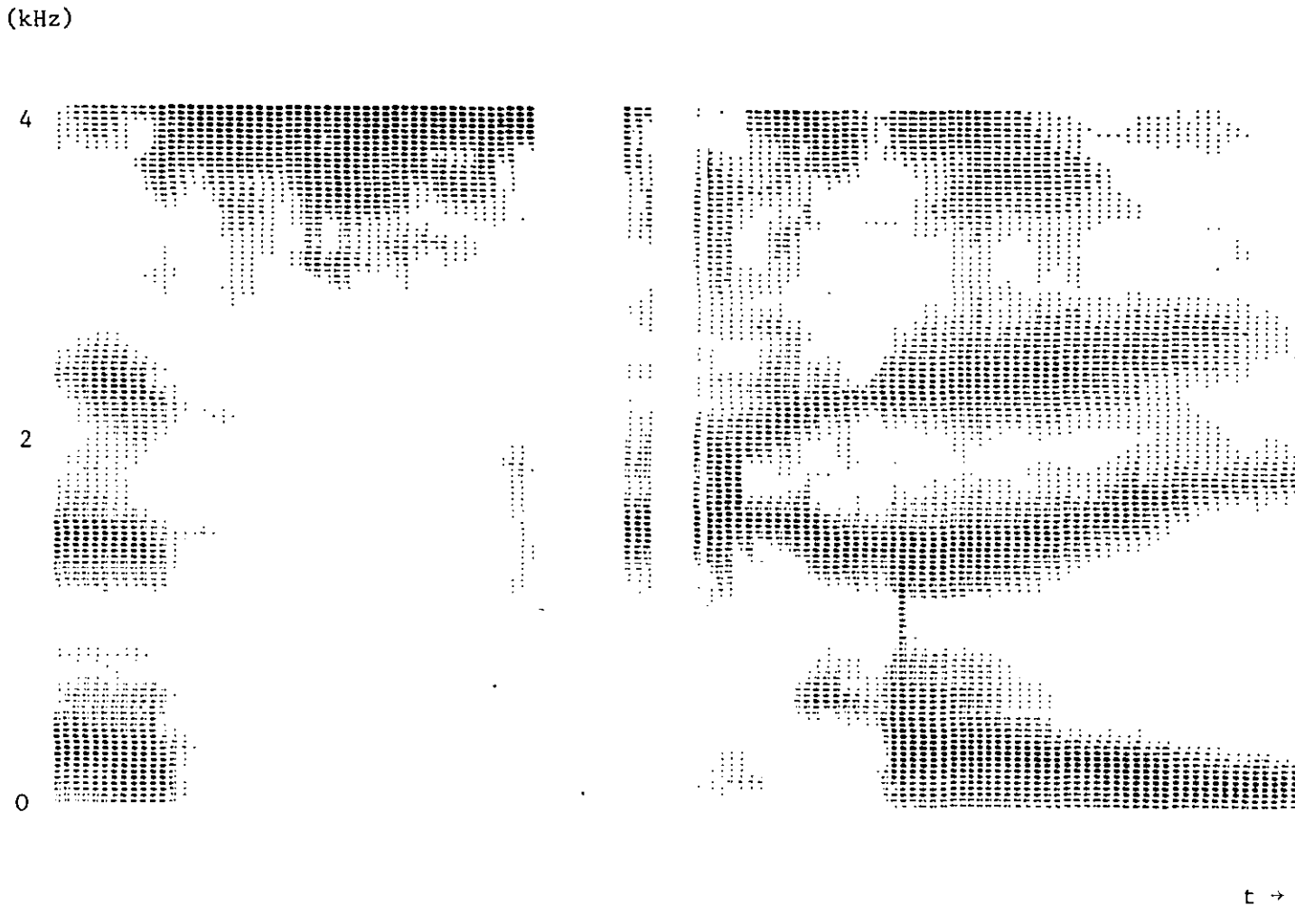


Fig 7.2 Line-printer spectrogram
(a) Wideband analysis.



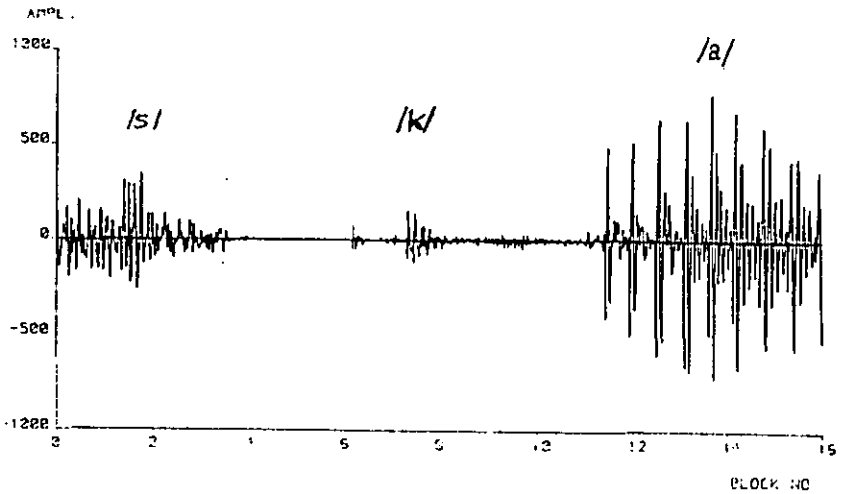
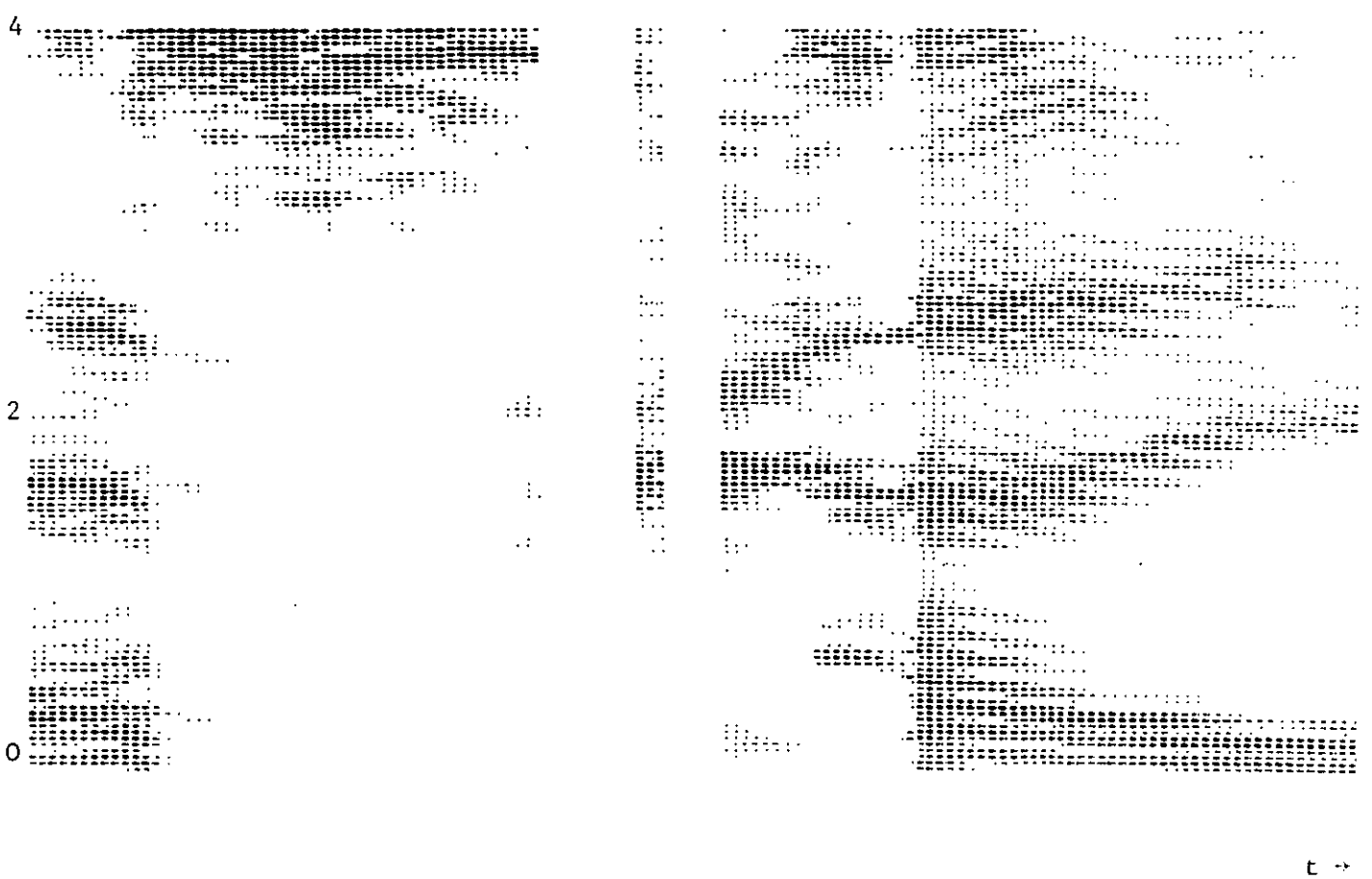
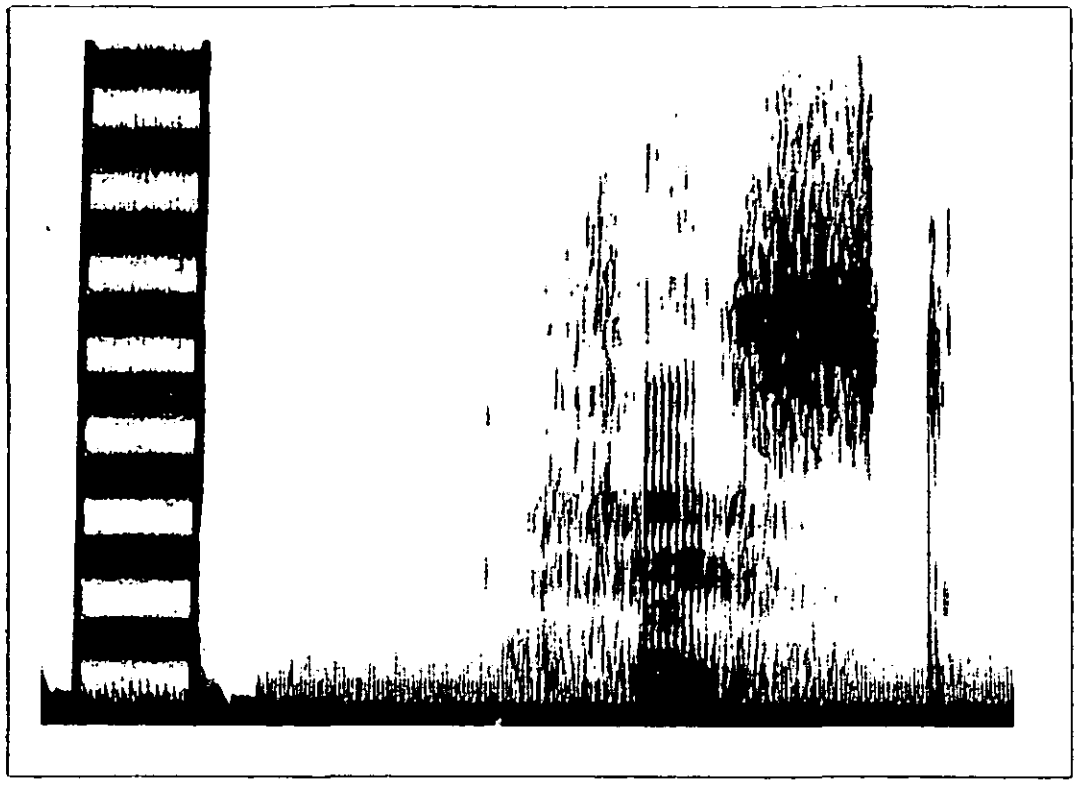
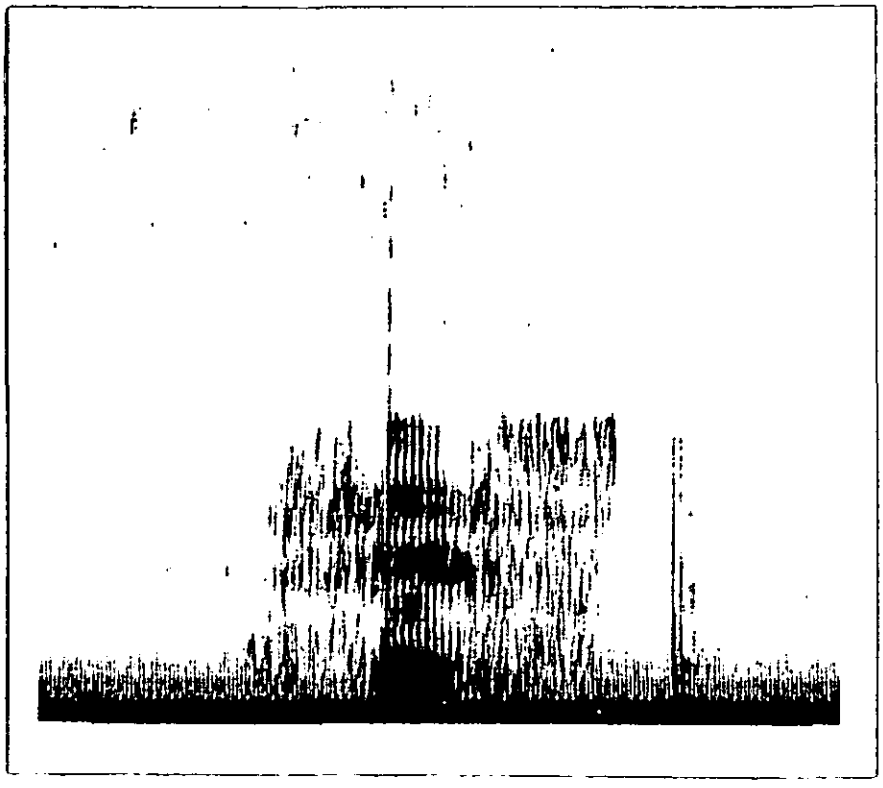


Fig 7.2 Line-printer spectrogram
(b) Narrowband analysis.



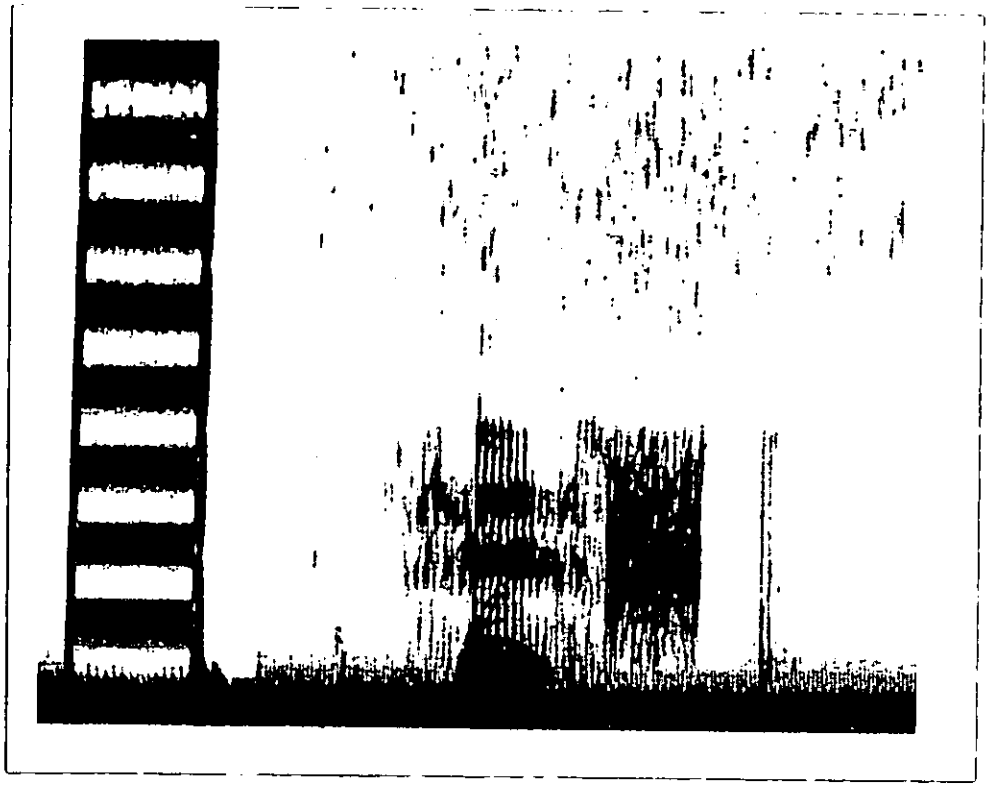


(a) Original signal "fist" bandlimited from 0.3 to 7.6 kHz.

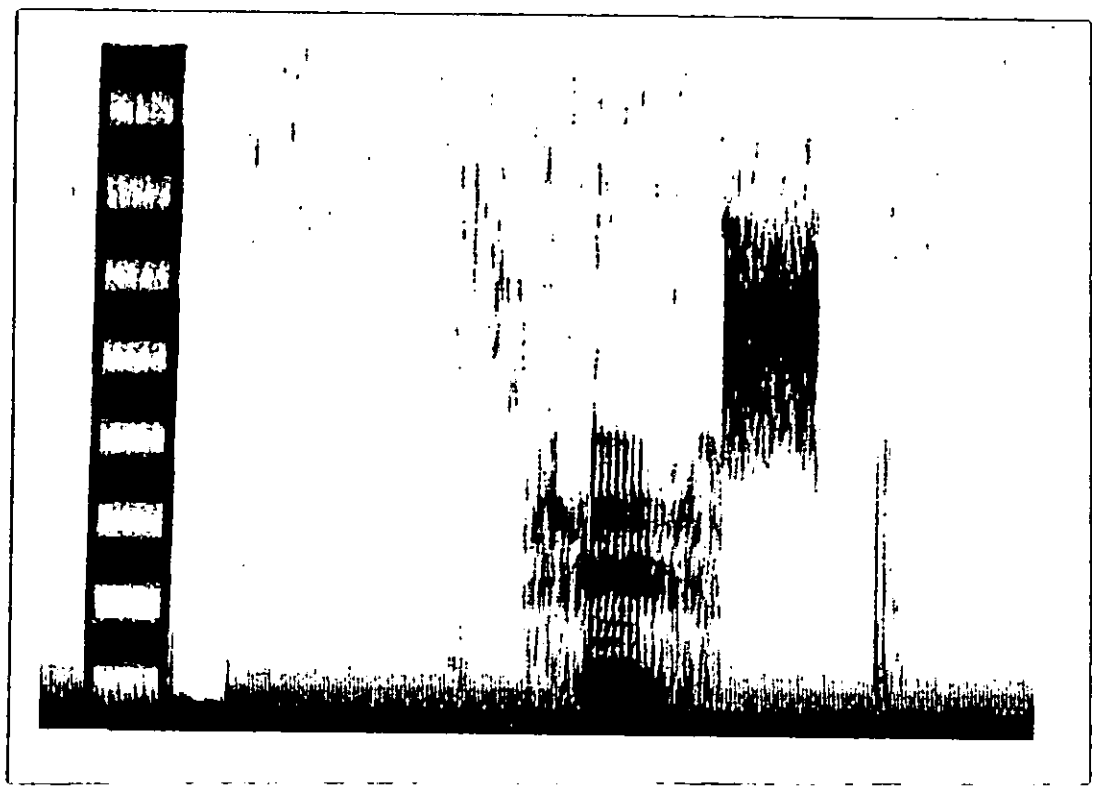


(b) Bandlimited signal from 0.3 to 3.5 kHz

Fig. 7.3 Spectrograms of the signals from some of the processes developed in this thesis

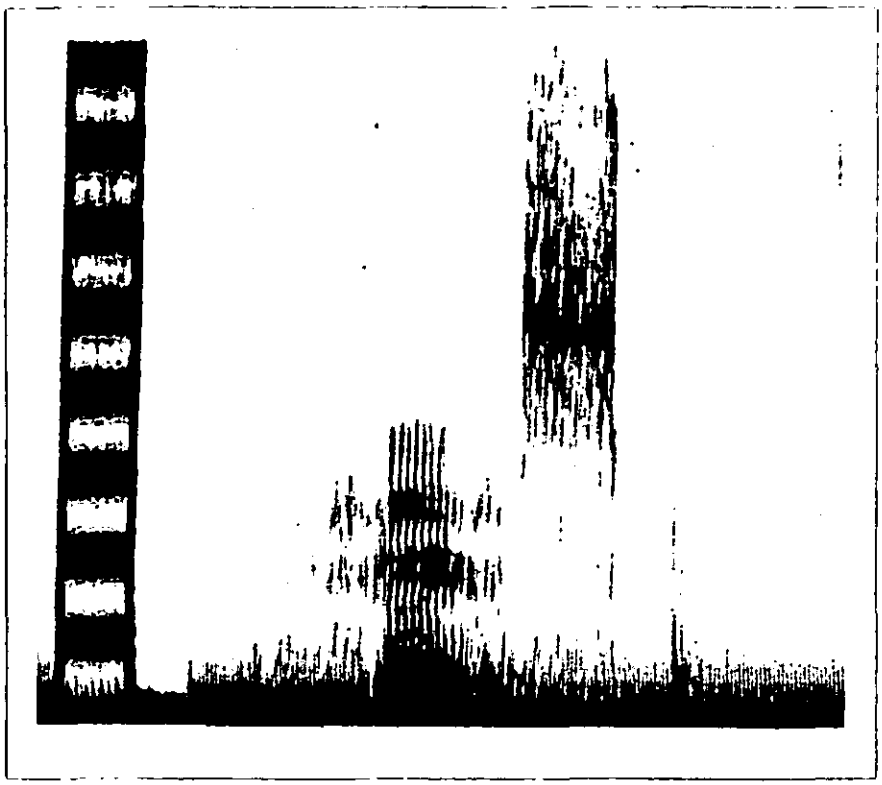


(c) Transmitted signal from the VUBS system

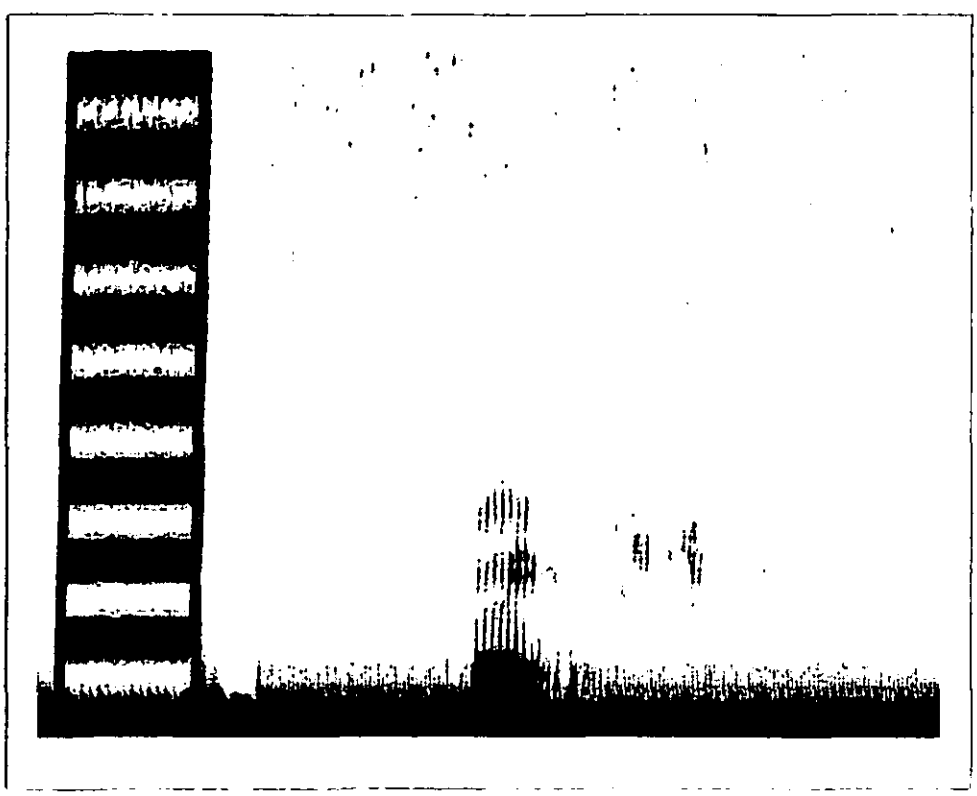


(d) Output signal from the VUBS system

Fig. 7.3 (Continued)

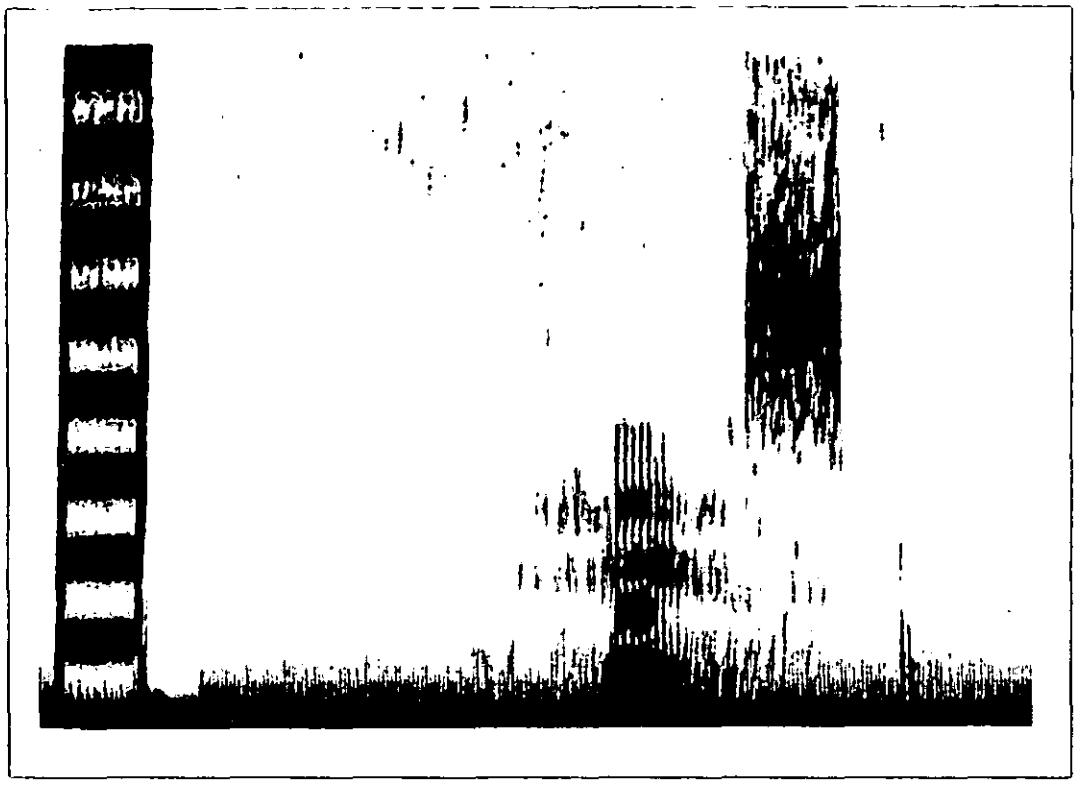


(e) Output signal from the random phase processor

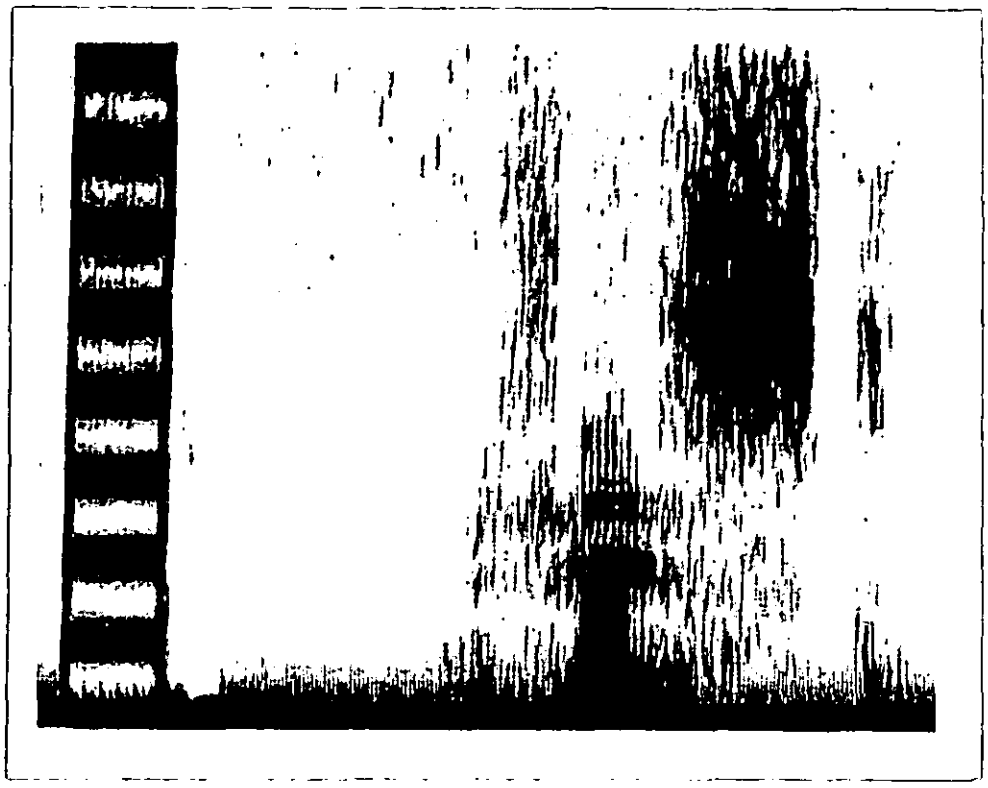


(f) Transmitted signal by the fixed frequency mapping system

Fig. 7.3 (Continued)

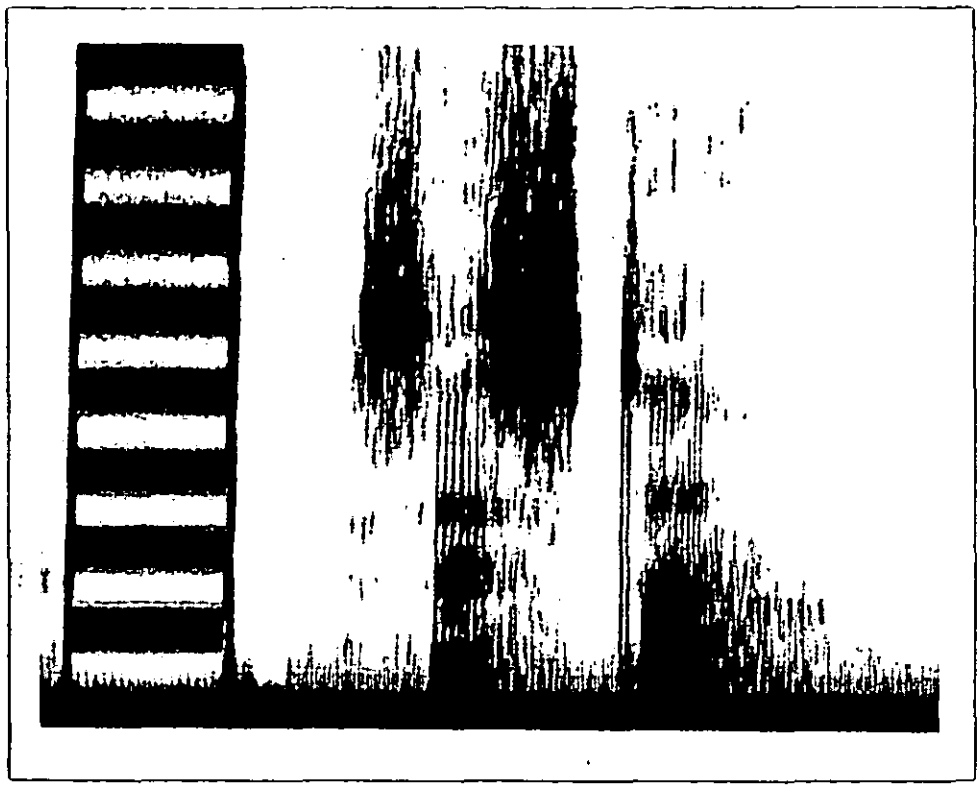


(g) Output signal from the fixed frequency mapping system

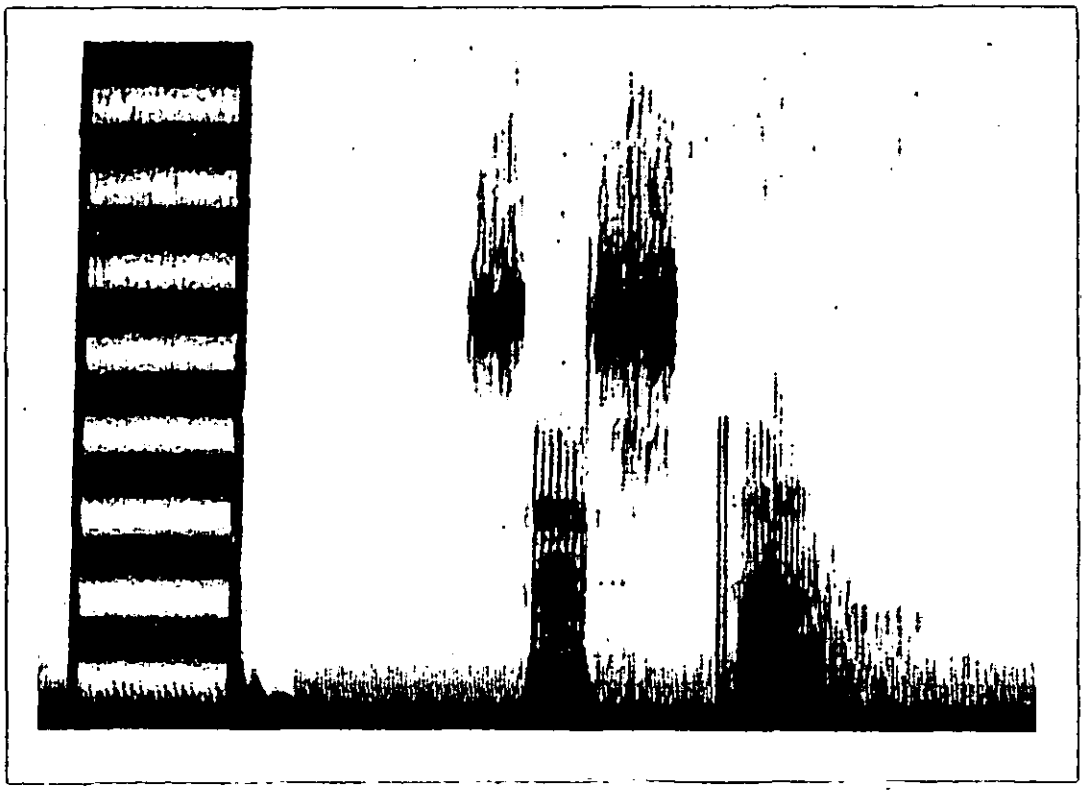


(h) Output signal from the adaptive frequency mapping system

Fig. 7.3 (Continued)

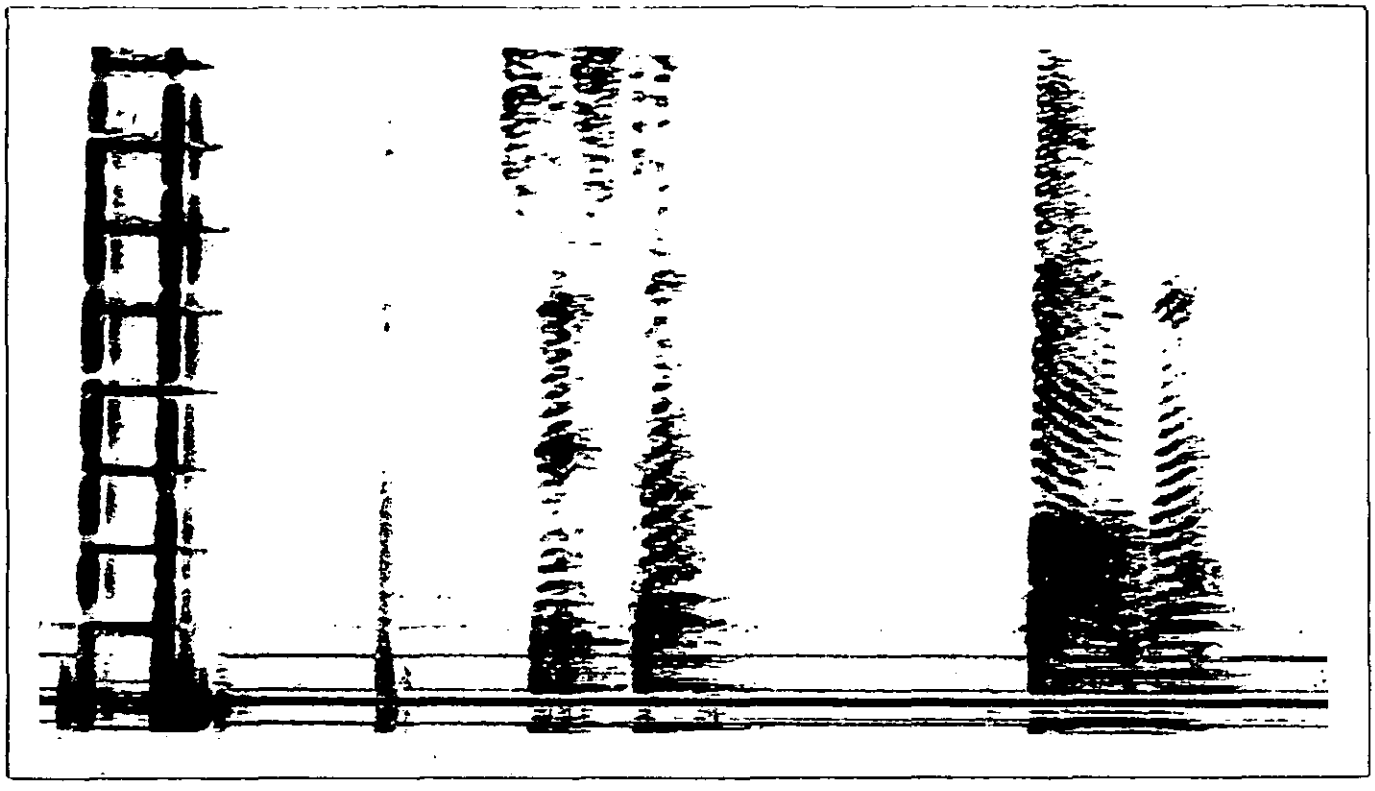


(i) Original signal "sister" bandlimited from 0.3 to 7.6 kHz

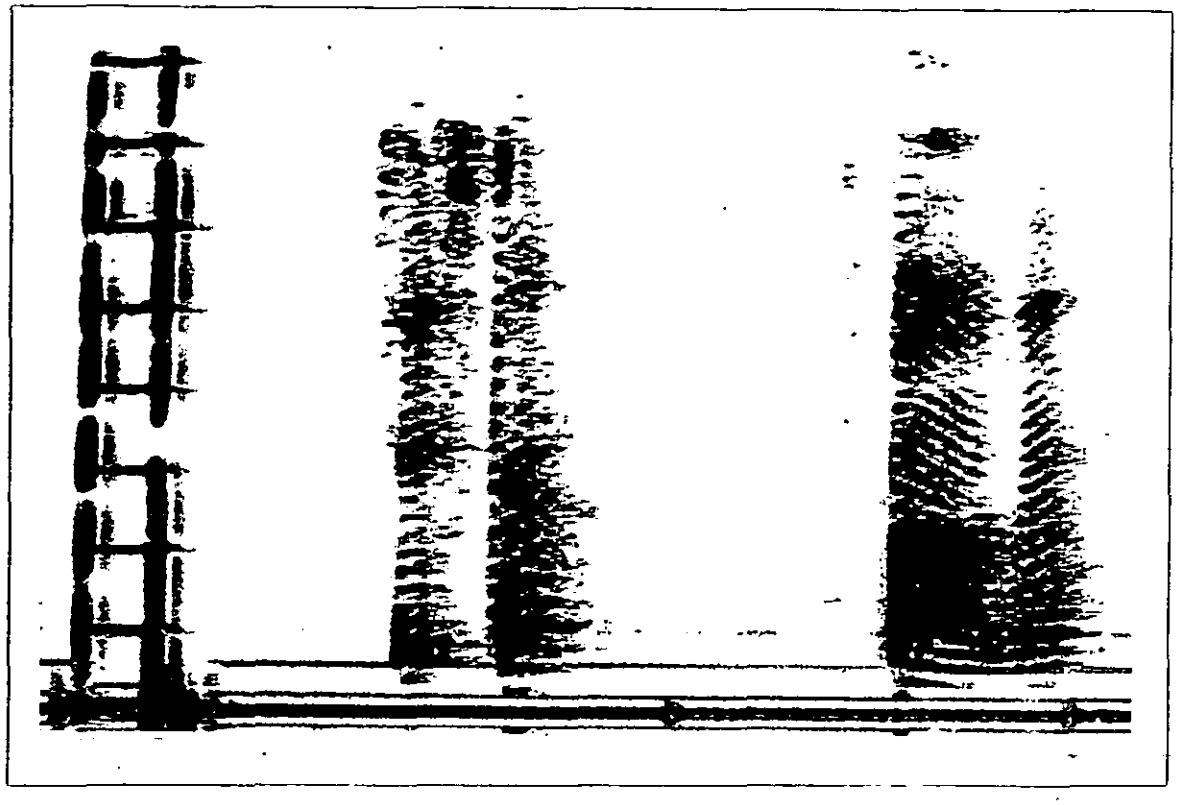


(j) Processed signal by the high frequency regenerator system

Fig. 7.3 (Continued)



(k) Original signal "sister" and "father" lowpass filtered from 0 to 7.6 kHz



(l) Bandlimited signal from 0.3 to 3.4 kHz

Fig. 7.3 (Continued)

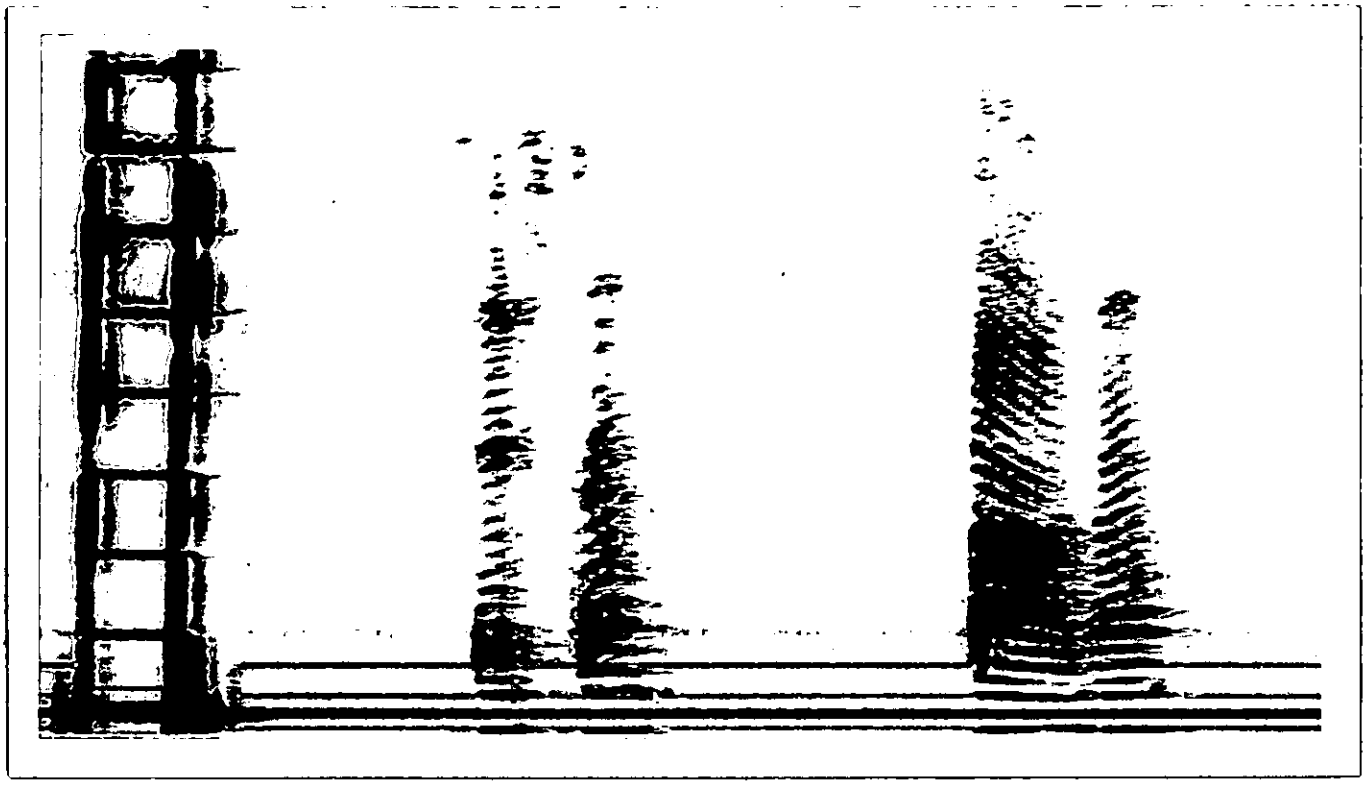


Fig. 7.3(m) Processed signal from the baseband synthesis system using a single oscillator driven at the pitch frequency

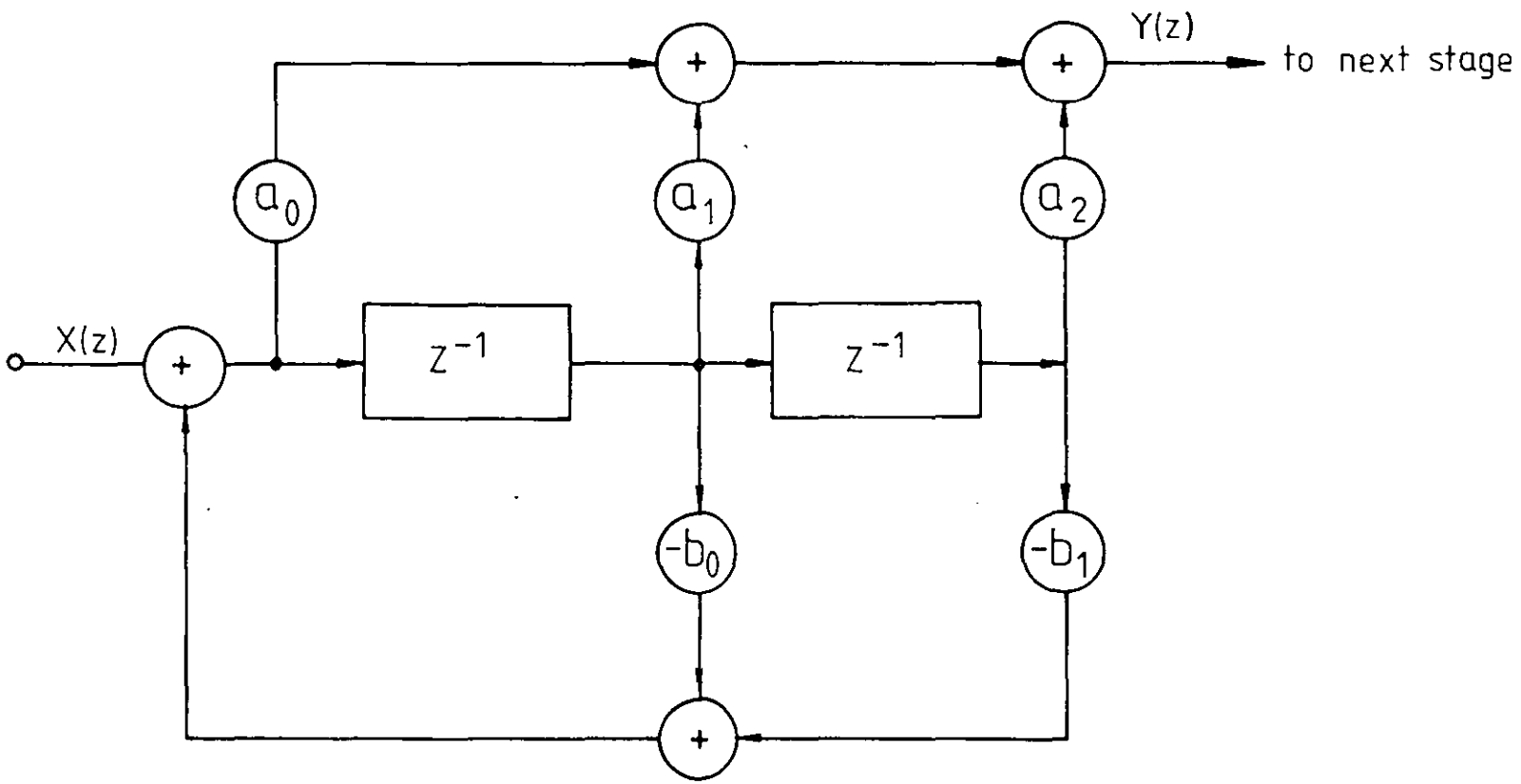


FIG. A1. SECOND ORDER STAGE FROM A SERIAL FORM OF A DIGITAL FILTER

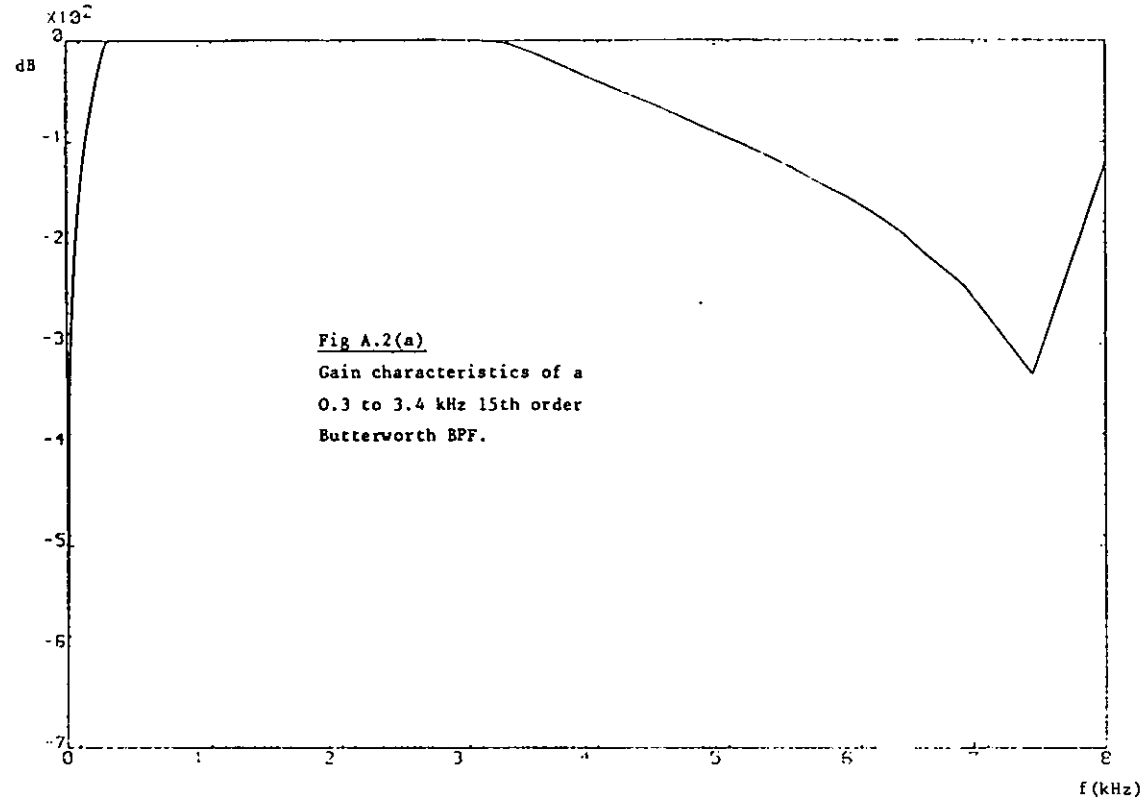


Fig A.2(a)
Gain characteristics of a
0.3 to 3.4 kHz 15th order
Butterworth BPF.

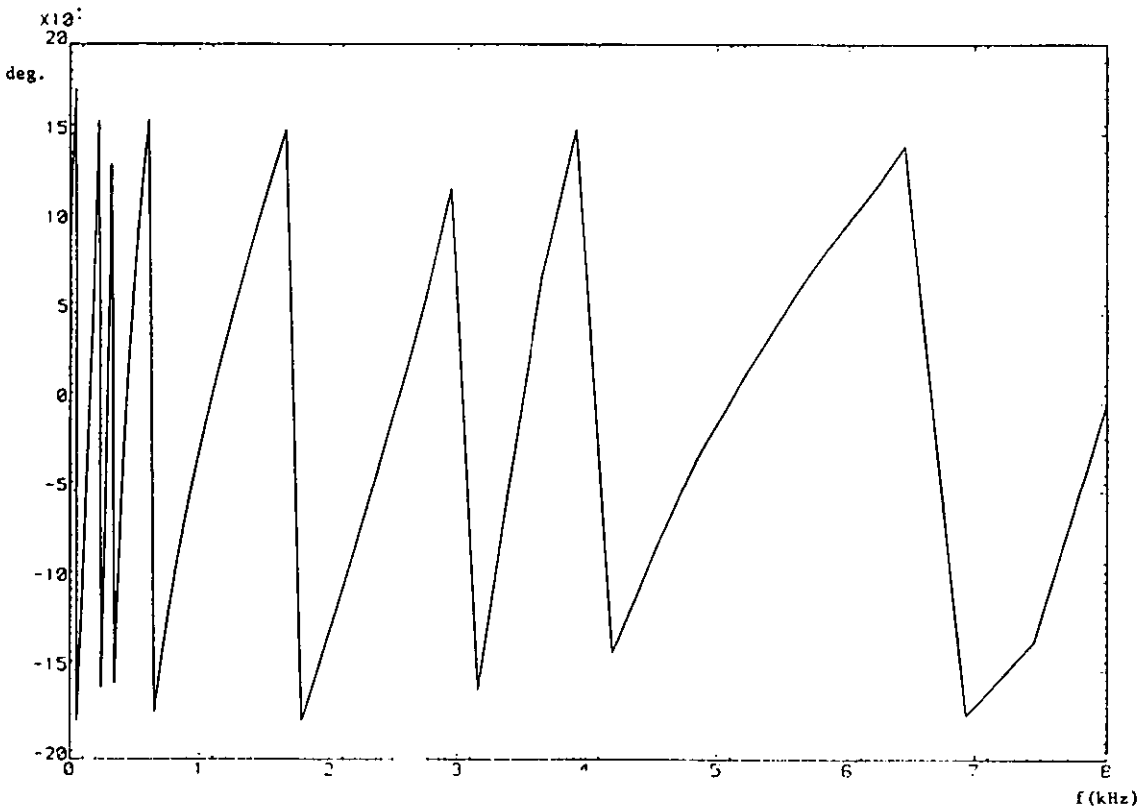


Fig A.2(b) Phase characteristics of the filter in sub-figure (a).

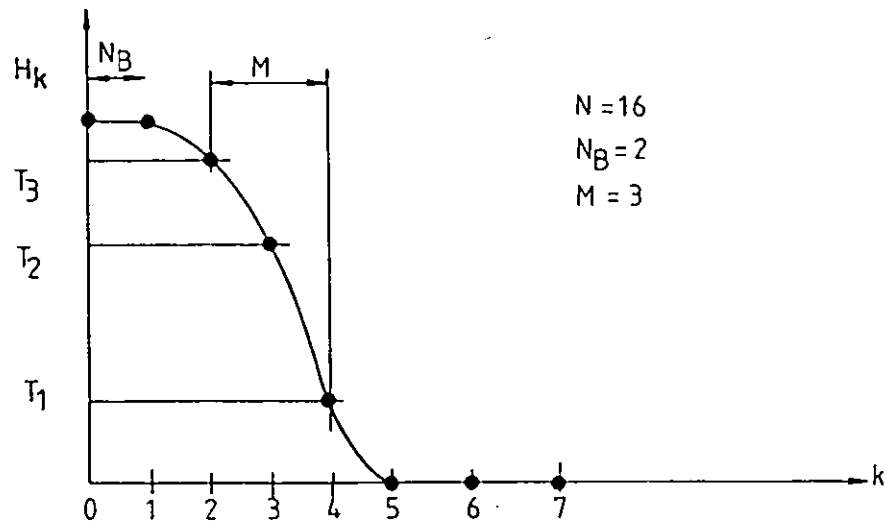


FIG. A3 TYPICAL VALUES FOR A LOW PASS FILTER

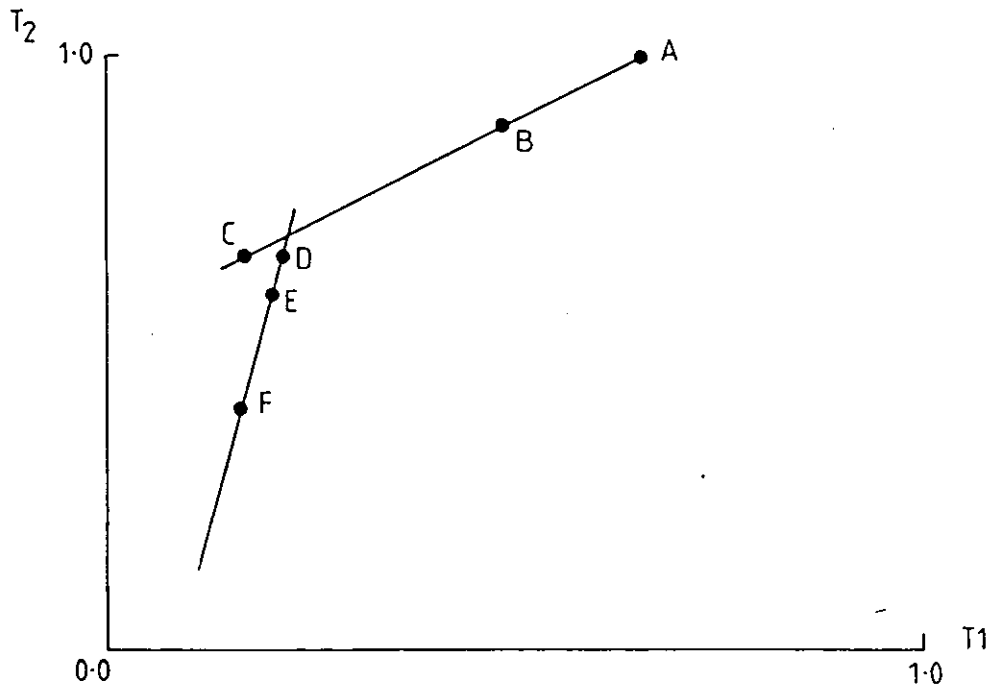


FIG. A4 ILLUSTRATION OF THE PATH FOLLOWED IN A TYPICAL SEARCH FOR TWO OPTIMUM TRANSTION COEFFICIENTS (REF 100)

N = 256

N_B	L(dB)	T_1	T_2	T_3
1	−92.07104015	0.10647949	0.19387524	0.67664281
2	−93.63089275	0.01963501	0.22197911	0.70144920
3	−87.43575478	0.01908569	0.22085960	0.69990539
5	−89.86921692	0.02305298	0.24117076	0.71635813
8	−89.21122360	0.02479248	0.24843111	0.72164702
9	−88.34475231	0.02329712	0.24253562	0.71679420
16	−88.42712784	0.02444458	0.24629538	0.71900030
32	−87.89452744	0.02577896	0.25163493	0.72307099
48	−87.84068012	0.02421875	0.24359358	0.71550480
56	−86.96756554	0.02345581	0.23957232	0.71177494
64	−87.60656548	0.02396851	0.24199281	0.71380179
80	−87.11819744	0.02351685	0.23926844	0.71102085
96	−86.78892708	0.02435913	0.24219392	0.71293931
104	−85.55295181	0.02552490	0.24590992	0.71524908
112	−85.86081982	0.02607422	0.24926456	0.71857490
120	−88.45293331	0.02683105	0.25909273	0.73130690
121	−90.09288883	0.02561035	0.25523207	0.72916388
122	−93.23881817	0.02344360	0.24778644	0.72456966
123	−99.37811375	0.01946106	0.23070314	0.71099759
124	−113.12398720	0.01351929	0.20394843	0.69037794

Table A.1

The optimum transition values T_1, T_2 and T_3 obtained for different pass-band samples N_B. L(dB) shows the level of the highest sidelobe (minimax) with respect to the main lobe height (ref.100).

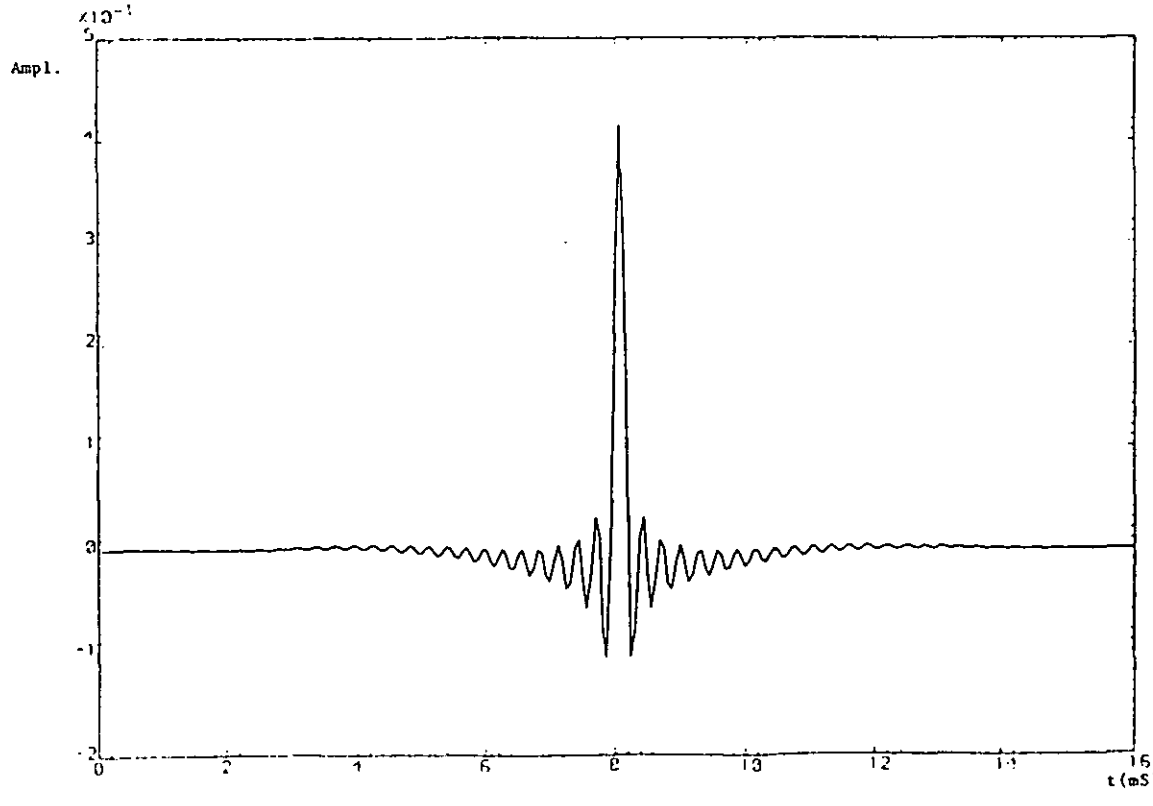


Fig A.5 Impulse response of a 0.3-3.4 kHz FIR bandpass filter.

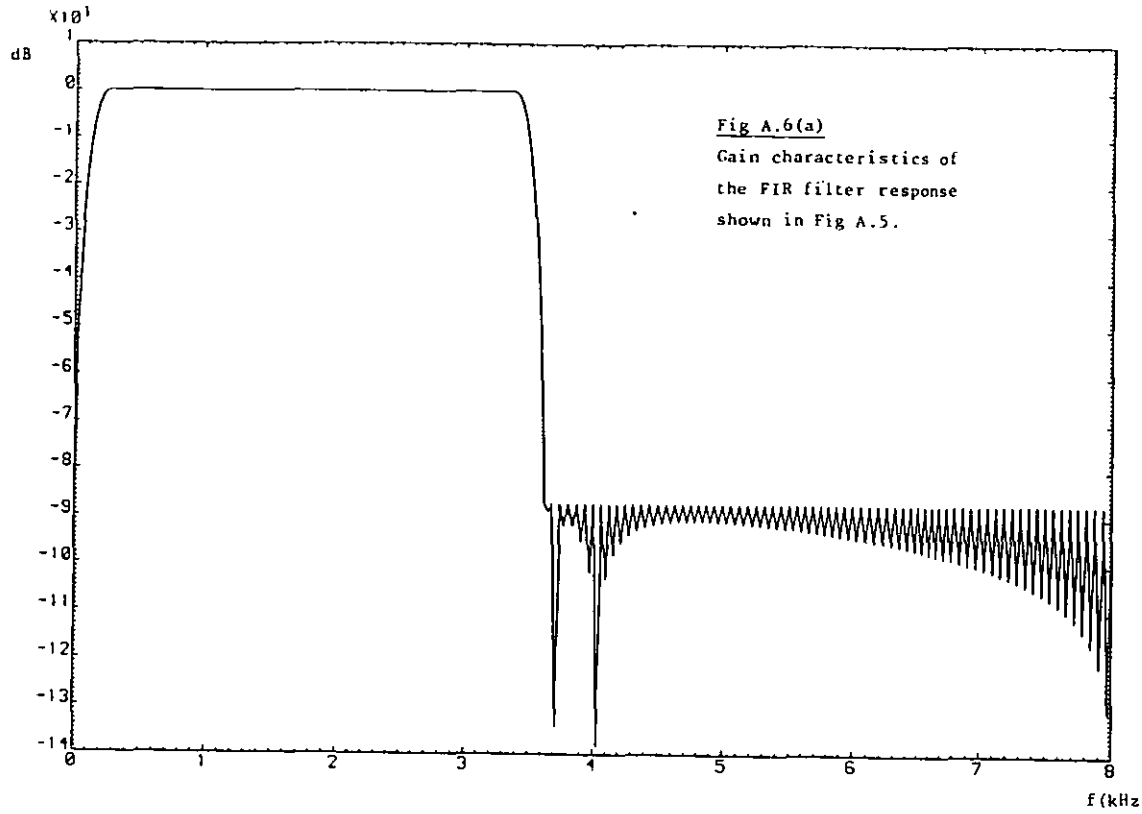


Fig A.6(a)
Gain characteristics of the FIR filter response shown in Fig A.5.

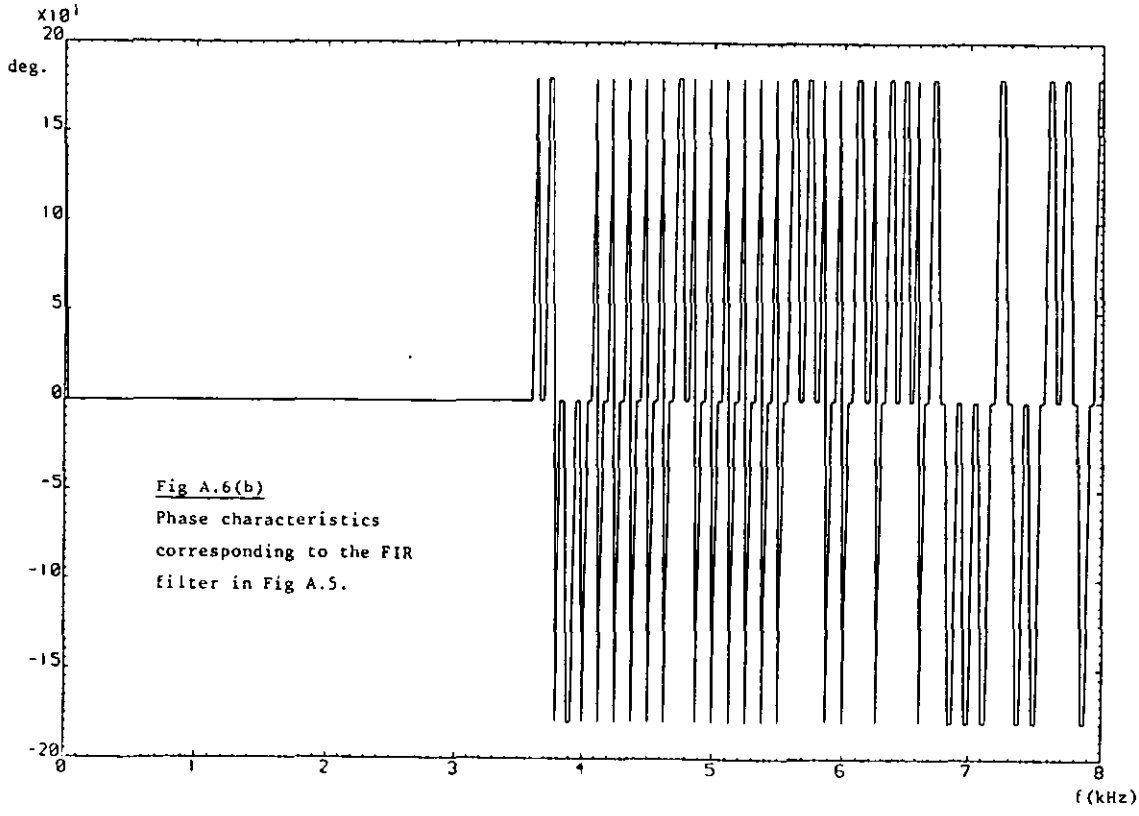


Fig A.6(b)
Phase characteristics corresponding to the FIR filter in Fig A.5.

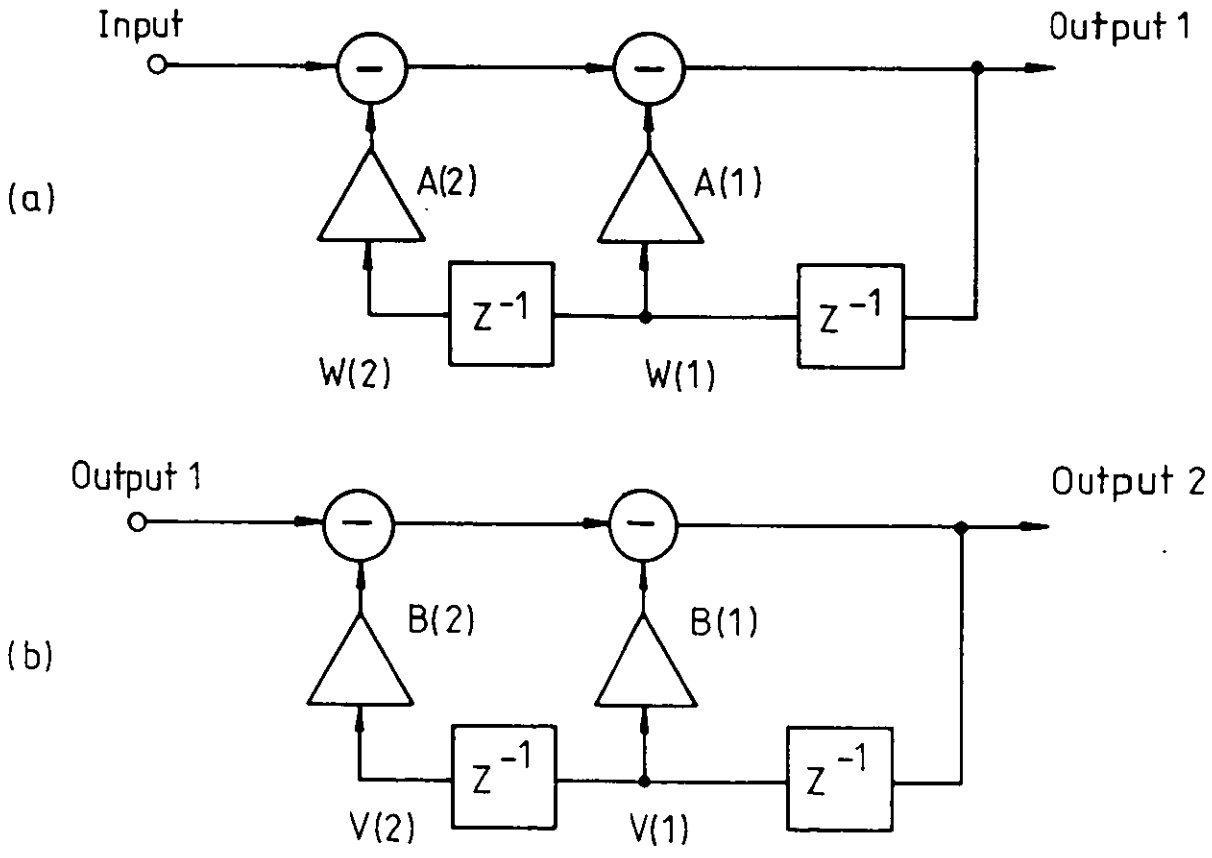

Input
Output 1
(a)
A(2)
A(1)
Z -1
Z -1
W(2)
W(1)
Output 1
Output 2
(b)
B(2)
B(1)
Z -1
Z -1
V(2)
V(1)

FIG. A7