ANALYSIS AND CORRECTION OF THE HELIUM SPEECH EFFECT BY AUTOREGRESSIVE SIGNAL PROCESSING

by George Duncan

Thesis submitted for the degree of
Doctor of Philosophy,
University of Edinburgh,
1983.

Declaration of Originality

This thesis has been composed in its entirety by myself, and reports research, conducted exclusively by my own hand, within the Wolfson Microelectronics Institute and the Department of Electrical Engineering, University of Edinburgh.

ABSTRACT

The use of helium-oxygen respiratory mixtures in deep diving operations creates severe communication difficulties by distorting divers' speech, rendering it unintelligible to the listener.

This thesis proposes a new means of correction for the helium speech effect suitable for implementation as a real time electronic system. The proposed system is termed 'the residually-excited linear predictive coding (RELPC) helium speech unscrambler', and is based on autoregressive (AR) signal processing. Identification of this system is achieved in part by an analysis of acoustic events in helium speech, in order to specify those distortions affecting the perception and intelligibility of human speech. The analysis, which relies in the main on AR spectrum estimation techniques, quantifies certain known aspects of the helium speech effect, which is shown to be nonlinear in nature, in addition to exposing new acoustic phenomena which may contribute to the overall degradation of helium speech intelligibility.

The operation of certain existing in-service helium speech unscramblers is also examined in detail, in order to locate those deficiencies in their overall signal processing strategies which directly affect the performance in terms of intelligibility afforded by these systems. Helium speech unscrambling by frequency transform processing is identified as offering potentially the best fidelity of the resulting speech. Direct application of the frequency transform is, however, shown to produce undesirable effects in the resulting speech waveform.

The RELPC system conserves ease of nonlinear correction for the helium speech effect through frequency domain transformations, but improves fidelity of the unscrambled speech by complementary use of AR signal processing.

The RELPC system is therefore an improved solution to the helium speech unscrambling problem in consideration of the speech mechanism as a short-term linear time-invariant filter system.

ACKNOWLEDGEMENTS

The author wishes to extend his grateful thanks to the following, both for the kind sacrifice of their valuable time and for their extremely useful help and willingness to discuss aspects of the work presented in this thesis: Dr. J. Laver, Mr. N. Dryden, Dr. C. Shuken and Mr. S. Hiller, Dept. of Linguistics and Phonetics, University of Edinburgh; Mr. B. Thomson, Subsea Offshore Inc., Aberdeen; Mr. C. Roper, Dive Equipment Technology Ltd., Aberdeen; Mr. C. Fyffe, Institute of Geological Sciences, Edinburgh; Dr. J. Fulton, Dept. of Mathematics, University of Edinburgh; Dr. M. Jack, Dr. J. Hannah, Dr. C. Cowan, Dr. J. Dripps, Mr. J. Nash and Mr. M. Rutter, Dept. of Electrical Engineering, University of Edinburgh. Thanks are also extended to those who may not be mentioned here but whose helpful contributions are nonetheless acknowledged. Last, but by no means least, to my wife, Catherine, for typing this thesis.

This research has been supported by the Procurement Executive, Ministry of Defence.

«Il n'y a point d'effet sans cause,» repondit modestement Candide, *«tout est enchainé nécessairement, et arrangé pour le mieux dans le meilleur des mondes possibles.»*

François-Marie Arouet
de Voltaire (1694-1778)
Candide, Chp.III.

"There is no effect whatsoever without there being a reason for it," replied Candide unassumingly, *"everything is necessarily linked together, and all is for the best in the best of all possible worlds."*

Cette œuvre est dedicacée à ma femme, *Catherine*.

TABLE OF ABBREVIATIONS AND SYMBOLS

Abbreviation	Meaning
AR	autoregressive
a.r.f.	autoregressive filter
D/A	digital-to-analog
dB	decibel
DFT	discrete Fourier transform
equ.	equation
Fig.	Figure
FFT	fast Fourier transform
Hz	Hertz
kHz	kiloHertz
l.m.s.	least mean squares
LTI	linear time-invariant
mS	millisecond(s)
p.e.f.	prediction error filter
PSD	power spectral density
rec.	recording
RELPC	residually excited linear predictive coding
STFT	short-time Fourier transform
$\sum_{n=0}^{N}$	discrete summation from n=0 to n=N
$\int_{a}^{b} y(x)dx$	continuous integration over the range a to b
$a(f) \rightleftharpoons b(t)$	is the dual transform of (e.g. a(f) transforms to b(t) or a(t) transforms to b(f))
$\simeq$	is approximately equal to
$a \times b$ $a.b$	a times b (non-vector multiplication)
$\|x\|$	modulus of x

CONTENTS

IV (Continued)

CHAPTER I

A SHORT CHRONICLE OF MAN IN THE HYPERBARIC UNDERWATER ENVIRONMENT AND SYNOPSIS OF THESIS.

1.1 INTRODUCTION

The remorseless efforts of man to overcome the obdurate barrier of the seas and rivers and to exploit the natural riches of the ocean floor have endured many centuries; even today man still cannot claim to be the master of an environment which, by its very quintessence, is intractably hostile to his physiological and mental well-being.

Some of the earliest attempts to sustain the excursion of man into the ocean environment can be traced to native divers, who would take natural sponges to the surface, wring them dry, and then, after cover--ing them with oil, dive to depths of some 40-50 feet carrying the sponges in their mouths, breathing the air trapped within the capillary channels of the sponge in order to prolong their stay under water. Such a technique is, by the very nature of the respiratory system of man, very limited in terms of the duration in the underwater environ--ment it accords to the subject. Indeed, Sir Edmund Halley, in 1716, writes of this method [1]:

"....it cannot be believed that a Supply, by this means obtained, can long subsist a diver. Since by Experiment it is found that a Gallon of Air, included in a Bladder, and by a Pipe reciprocally inspired and expired by the Lungs of a Man, will become unfit for any further Res--piration, in little more than one Minute of Time; and though its Elasticity be but little altered, yet in passing the Lungs, it loses its Vivifying Spirit, and is rendered effete, not unlike the Medium in Damps, which is present Death to those that breath it; and which in an instant extinguishes the brightest Flame, or the shining of glowing Coals, or red hot Iron, if put into it."

The earliest man-made contrivance for prolonging the working period of man underwater involved the use of a helmet from a suit of armour, onto which two pipes were fitted, the one taking air supplied from bellows to the diver, the other returning the expired air to the sur--face. Such a device was impracticable, however, beyond a depth of 3 fathoms, the main problem being the sown-leather suit to which the helmet was attached, which was apt to spring leaks at such depths, much to the consternation of the diver involved.

This problem was obviated somewhat by the inception of the diving bell, so-called because of its conical shape. This device was closed at the top (apex) end, but open to the sea at the bottom end in order to allow the diver to carry out his work. Of this device, Sir Edmund writes:

"....and if the Cavity of the Vessel may contain a Tun of Water, a single Man may remain therein as least half an Hour, without much inconvenience, at 5 or 6 Fathoms Deep."

Sir Edmund, however, does indeed enumerate several inconveniences. At 6 fathoms, the bell is half full of water due to the ambient pres--sure; the diver cannot sustain the coldness of the water in the bell for very long, and lastly, the main inconvenience is due to the accu--mulation of the diver's expired, tepid air which renders the whole atmosphere rapidly unfit for respiration. Sir Edmund's contribution to the science was a bell which, firstly, had a glass panel inserted at the top to improve lighting. Most importantly, however, the air inside the bell could be replenished by a system of two barrels, each holding 36 gallons of air, which were alternately lowered by a pulley down to the diver, who then emptied the fresh air into the bell. There were also a stop-cock near the top of the bell so that tepid air could be vented before it made the atmosphere of the bell intolerable.

This technique, first used in 1691, both prolonged the useful stay of the diver under the sea and increased his working depth, and was to last almost a century until the development of pumps capable of pro--ducing compressed air in 1788. Extolling the virtues of his device, Sir Edmund writes:

"Therewas such a plentiful supply of air, that I myself have been One of the Five who have been together at the Bottom, in nine or ten Fathoms of Water, for above an Hour and Half at a time, without any sort of ill consequence: and I might have continued there as long as I pleased, for any thing that appeared to the contrary. Besides, the whole Cavity of the Bell was kept entirely free from Water, so that I sat on a Bench, which was diametrically placed near the Bottom, wholly drest with all my Cloaths on. I only observed, that it was necessary to be let down gradually at first, as about 12 foot at a time; and then to stop and drive out the Water that entred, by recei--ving three or four Barrels of fresh Air, before I descended further."

It is indeed a paradox that, some two centuries later, it was to be shown beyond question that it was the ascent to the surface, and not the descent to the ocean floor, which was to prove most incommo--dius. Further still, the "Vivifying Spirit" which Sir Edmund so painstakingly sought to restore to the respiratory mixture, and which was later identified in 1774 by J. B. Priestly to be oxygen, was to prove to be one of the most toxic elements known to man when respired in a hyperbaric atmosphere.

1.2 PHYSIOLOGICAL AND MENTAL EFFECTS OF A HYPERBARIC AIR ATMOSPHERE

While observing Irish immigrant labourers during the construction of the Eads bridge over the Mississippi at St. Louis, U.S.A, in the mid-19th century, the French doctor A. Jaminet observed on their return to the surface, having spent several hours working on the bridge substructure at a pressure of 4 atmospheres in submerged

caissons, that several of the men suffered severe convulsions and haemorrhageing from the nose, and indeed there were a series of fatalities [2].

In fact, of the 160 diving labourers employed in the project, some 30 or so became severely and permanently paralysed, whilst 12 suffered sudden death. Many others experienced violent convulsions, and pains in the muscles and joints coupled with temporary paralysis, especially in the lower limbs. It is also recorded that a certain construction company in England at that time lost 10 of its 24 divers, 3 from sudden death and 7 dying as a result of severe paralysis after several months.

A few years later, in 1876, it was another Frenchman, Professor P. Bert, who identified for the first time the root causes of decom-pression sickness, or "the bends" as it came to be colloquially known. Prof. Bert was able to show that in saturation diving, in which the diver can be considered as 'saturated' with gas for a specific depth (pressure), the crucial factor determining the diver's susceptibility to "the bends" was not, as believed, the rate of compression or descent into the water, which could in fact be carried out at any desired rate, but rather depended on the rate of recompression or ascent out of the water [3].

At depth, the diver's bloodstream can be considered to be saturated with the component gases of air for the corresponding ambient pressure. If recompression occurs too rapidly, that is, the pressure is relieved too quickly, then the gas appearing out of solution from the blood-stream does not have enough time to diffuse through the skin and muscle tissue, but instead accumulates in the veins, acting as an air-lock, thereby, at the very worst, halting circulation and causing death [4,5]. The most troublesome gas in this respect has been shown

to be nitrogen, which constitutes some 78% by volume of air. The oxygen has been shown to recombine with the blood almost as quickly as it appears.

However, Prof. Bert was able to demonstrate that it was the oxygen in the compressed air at depth which was toxic to man and was the cause of violent convulsions. Further investigation has shown that breathing a gaseous mixture whose partial pressure of oxygen exceeded 0.8 bar for long periods in hyperbaric conditions will produce violent convulsions and may even result in death.

With the advance of technology, deeper dives were now becoming possible, and with them came the discovery that, not only could oxygen promote convulsions, but nitrogen too was very troublesome, producing a state of narcosis in the diver which worsened with increasing depth or pressure [6]. Investigations have shown that the first symptoms are headaches, vertigo and vomiting, leading to the rapid onset of delirium tremens, during which phase it appears that the most common hallucination is for the diver to visualise someone once very close to him but now dead walking across the sea bed towards him, and actu-ally start a conversation with him [7].

Another distressing condition which can occur is dyspnea, described as being "like trying to breathe air through a straw". This ailment arises through respiring a very dense gas mixture under pressure.

Despite these distressing indications, it in fact takes a very experienced diver to recognise them himself, and in most cases it is the surface crew who first realise that the diver is undergoing nitrogen narcosis.

1.3 THE HYPERBARIC HELIUM-OXYGEN ATMOSPHERE

In order to obviate these dangerous and all-to-easily accomplished physiological conditions, and in addition increase the maximum possible diving depth, a lightweight gas was sought to replace nitro--gen and the other heavy gases present in air. The most convenient replacement so far has proved to be helium. It has various advantages. It is inert; non-toxic; non-inflammable; colourless, odourless and tasteless, and of course is very light. Note, however, that the use of a lighter respiratory mixture does not affect the diver's suscept--ibility to decompression sickness, nor does it significantly reduce the waiting periods during ascent.

The Americans were the pioneers in the use of a helium-oxygen (heliox) respiratory mixture in the early 1930's, and the Royal Navy, in 1948, succeeded in lowering a diver to 540ft in the open sea - a record which was to last until the late 1960's. Today, working depths of 500ft are commonly attained using a helium-oxygen mixture.

The percentage volume of helium in the respiratory mixture - compr--ising typically of around 96% helium and 4% oxygen - may at first appear excessively high. However, recalling that it is the partial pressure of oxygen that is critical, then although the volume of oxy--gen is low, its partial pressure is sufficient to sustain life.

One of the difficulties in using a lighter-than-air respiratory mixture is that it tends to find escape routes from equipment in a way that heavier gases would not, and therefore causes problems in main--taining hermetic seals in breathing equipment.

However, the principal disadvantage of using a heliox mixture relates to its disastrous effect upon voice communications. The heliox mixture, with its increased velocity of sound and different acoustic impedance with respect to air, alters the speech uttered by the diver to such an extent that what reaches the surface is not an intelligible acoustic waveform, but an "incoherent jumble", which none but the extremely experienced can hope to comprehend: even then, serious and occasionally fatal mistakes in comprehension can occur [8,9].

With the realisation that oil from the ocean floors was a viable source of national wealth, there has been a dramatic increase in deep-diving operations since the early sixties, with the attendant obligatory use of oxy-helium respiratory mixtures.

1.4 SYNOPSIS

Contemporaneously with the increased frequency of diving operations which involve long term endurance of a heliox environment, a range of electronic systems has been developed to enhance the intelligibility, under such environmental circumstances, of the divers' squeaky voice emissions, known as "helium speech", an example of which can be heard in recording A1.

There has, however, generally been cause to criticise the performance of these systems in terms of the intelligibility of the unscrambled speech which they produce. It is the purpose of the work presented within this thesis to determine the specifications for a new and improved helium speech unscrambler system in order to enhance speech intelligigility in respect of the listener.

The philosophy of approach to the research in this thesis has been

to investigate the problem of improving on existing unscrambler systems from two considerations. Firstly, that there may be as yet undiscovered acoustic phenomena in helium speech and/or phenomena which are known to occur, but have not been directly linked to the degradation in intelligibility of raw helium speech. Secondly, there may be aspects of the signal processing strategies, designed into existing unscrambler systems, which are detrimental in themselves to the intelligibility of the unscrambled helium speech.

As with all system problems in signal processing, it is imperative to possess as much information as possible about the entity to which processing is to be applied, and to this end, Chapter II of this thesis furnishes relevant details, in respect of the human speech signal, which are known to have a direct bearing on speech intelligibility under normal everyday atmospheric conditions. These are then related to the known acoustic events of helium speech in terms of how changes in various acoustic phenomena affect speech intelligibility vis-à-vis the listener. Here, it is also shown that the diver himself can adapt his own speech while breathing heliox, presumably in an attempt to render it more intelligible as he himself judges. Chapter II concludes with a brief discussion of the existing unscrambler systems currently in service and their relationships to the presently accepted acoustic phenomena of helium speech.

As part of the research undertaken and reported in this thesis, the opportunity arose to make new recordings of a diver in situ in a (simulated) pressured heliox environment. As will be appreciated from the summary of the presently-known characteristics of helium speech in Chapter II, there is still some controversy in respect of certain acoustic events, and so the procurement of these recordings presented

the possibility of a fresh examination and assessment of helium speech taken in comparison to normal air speech.

Chapter III presents the experimental scenario, material used for the recordings and the mathematical techniques based in the main on autoregressive signal modelling, used to facilitate the acoustic analysis of several facets of the recorded speech. The results from the various analyses are presented in detail. From these results, it has been possible to verify several of the existing premises regarding helium speech. The treatment and consideration here, however, have also produced new phenomena which have not before been reported, and which represent an original contribution to the science of the subject. The explanations offered here for these phenomena are, to a certain extent, conjectural in the sense that supportive measurements of the diver's physiological condition proved impossible to arrange. They are, nonetheless, based on a knowledge of events known to occur in normal air speech.

Whilst the research presented here relates primarily to the acoustic events of helium speech, the working environment of the diver can, however, contribute to the overall problem by acting as a noise source which may compound the problem of rendering the helium speech once more intelligible. Chapter III therefore concludes with an analysis of one of the major ambient noise sources encountered in diving operations, relating to filtering equipment for cleansing excess carbon dioxide from the respiratory mixture. The acoustic characteristics of this noise are investigated in detail and their likely effects upon the foregoing analyses are explored. Once again, this investigation is thought to provide an original contribution to the knowledge of the subject.

In Chapter IV, attention is turned to the unscrambler systems themselves in order to identify deficiences in their processing mechanics which may have an adverse effect on the quality of the unscrambled speech. Two systems are simulated and considered here in detail. The first concerns a system based on time-domain processing, this being the most popular real time device to be found in on-line service today. As is shown in Chapter III and as is indeed widely accepted, the translation from air to helium speech is nonlinear in nature. The time domain system is constrained to linear correction of the helium speech, however. Additionally, certain aspects of its functioning are identified as contributing to a degradation in the intelligibility of the unscrambled speech. As one of the simplest systems to implement, however, its overall performance can be considered here to act as a yardstick by which to measure the performance of more complex unscrambler systems.

The second system investigated in Chapter IV represents a helium speech unscrambler based on signal processing via the short-time Fourier transform (STFT), and permits nonlinear operations to be performed on the speech signal, thereby in theory providing improved intelligibility of the unscrambled helium speech. However, the use of the STFT in this application relies heavily on the apparent phase-insensitivity of the human ear as compared to the relative importance of the short-time magnitude spectrum in perception of the speech signal. This supposition is shown to be questionable, as discussed in the conclusions of Chapter IV, and moreover, it is demonstrated that the ear is indeed sensitive to the spectral phase of the speech signal.

The results of the research presented in Chapters III and IV channel directly to the novel residually-excited linear predictive coding (RELPC) unscrambler system, based on autoregressive signal processing techniques,

which is presented and simulated in Chapter V, and which represents the culmination of the original work presented in this thesis. The bases for the design of this system are given in detail, with consideration given to the requirements of an eventual processing system permitting unscrambling of the helium speech signal in real time. Furthermore, it is argued that, under the assumption that the human speech mechanism can be considered to be a linear time-invariant system in the short term, then the RELPC unscrambler is an improved unscrambler system in terms of ease of nonlinear correction and compatibility with the required criteria concerning those aspects of the speech signal which are important to human speech perception.

Chapter VI reiterates the main conclusions arising from this research into both helium speech and unscrambler systems for enhancing intelligibility in a helium-oxygen environment, and suggestions for future work in this field are proposed.

CHAPTER II

THE SPEECH MECHANISM IN RELATION TO HELIUM SPEECH AND ELECTRONIC UNSCRAMBLER SYSTEMS.

Signal processing demands an a priori knowledge of the desirable characteristics of the end-product waveform which will result from the processing applied to the input signal. Here, the resulting acoustic waveform must conform to the needs of the perceptual mechanism of the human brain such that the information which the talker wishes to convey is in a form which is intelligible to the listener. In normal speech in air, the human interpretation of speech depends upon a complex interplay of both temporal and acoustic events, the most salient of which are described in the first part of this chapter in terms of method of production by the talker and perception by the listener. In the second part of this chapter, the effects of a pressured helium-oxygen atmosphere on the acoustic waveform are expounded, and related to those perceptual factors directly affecting the intelligibility of helium speech. Lastly, the known properties of helium speech are related in brief to the processing strategies of presently available on-line helium speech unscrambler devices.

2.1 THE SPEECH MECHANISM

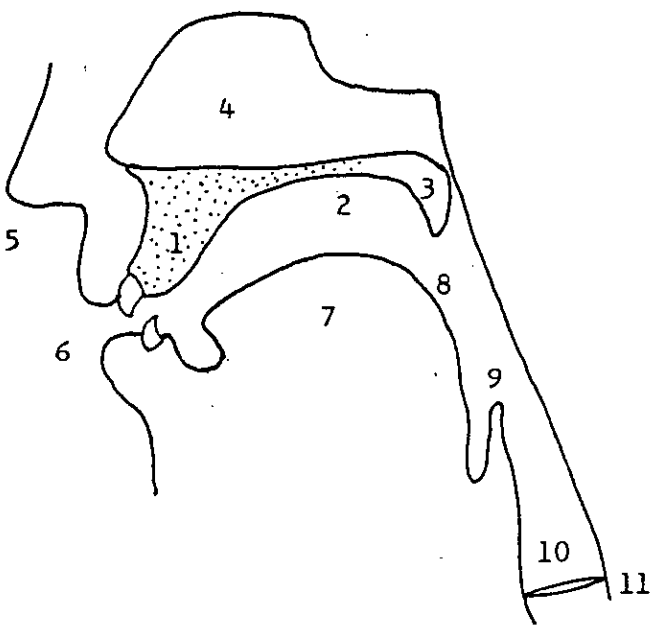
The acoustic waveform of speech is the result of a complex interaction of events within the vocal tract, which comprises the non-uniform tube from the vocal cords to the lips and includes the nasal cavity

(see Fig. 2.1). There are a variety of different temporal and spectral events, each of which, together with its associated method of physical production, has some influence upon the intelligibility of the speech waveform. For the purposes of the present treatise, however, the three principal ways only in which the vocal tract can be excited to produce an acoustic signal will be discussed.

2.1.1. THE VOICED ACOUSTIC WAVEFORM

Here, the vocal tract can be likened to a pulse-excited causal, minimum-phase filter. The non-uniform tube is approximated by a series of short concatenated sections each having uniform cross-sectional area, being terminated at one end by the vocal cords, through which the pulse excitation enters the system, and at the other extremity by the lips [10,12], from which the pressure waveform carrying the acoustic information travels out into the surrounding medium (see Fig. 2.2(a)). As a pressure wave enters the system through the vocal cords, it undergoes various transmissions and reflections at the boundaries of each acoustic tube section (see Fig. 2.2(b)), such that the system as a whole can be regarded as a series of resonating cavities, with each cavity contributing to the overall waveform shape and the characteristic distribution of the power spectral density, which, for voiced sounds, exhibits distinct areas of resonant energy, or formants, centered on certain frequencies.

In air at normal temperatures and pressures, the tissue of the vocal tract walls and articulators, or moveable elements, such as the palate, tongue, teeth and lips, can be essentially considered to be perfect acoustic reflectors, thereby forming lossless boundaries. The



1. Hard palate.
2. Soft palate.
3. Velum.
4. Nasal cavity.
5. Nostrils.
6. Lips.
7. Tongue.
8. Pharynx.
9. Epiglottis.
10. Glottis.
11. Vocal cords.

Fig. 2.1 Section through the human vocal apparatus.

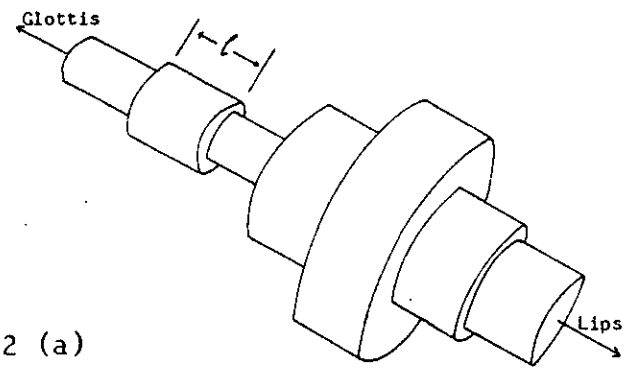


Fig. 2.2 (a)

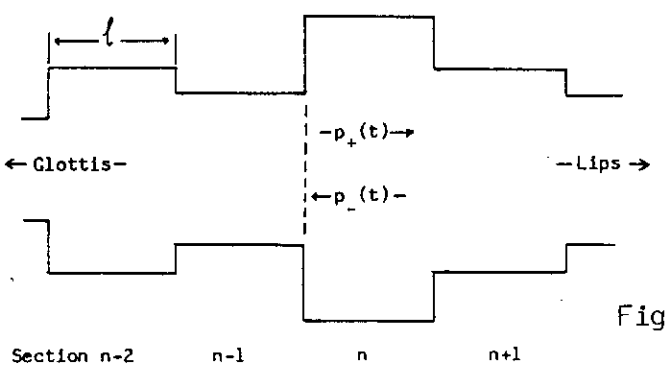


Fig.2.2(b)

Fig. 2.2 (a) Concatenated acoustic tube model of the vocal tract and (b) tube section showing resonating cavities with forward ($p_+(t)$) and backward ($p_-(t)$) pressure waves.

vocal tract filter is not, however, time-invariant in the long term, but can in most cases of voiced speech be so considered over periods of approximately 10-30mS [12].

Vocal tract shape can be altered by repositioning the articulatory members, the most mobile of which is the tongue [13]. Changing the shape of the various cavities of the acoustic tube in this way will alter their respective contributions to the overall waveform shape, thereby modifying the filtering characteristic of the vocal tract on the excitation source to produce changes in the power spectral density distribution of the resulting speech signal. The excitation source for the vocal tract filter cannot, however, be considered to be a perfect Dirac impulse. In order to produce the vocal tract excitation function, firstly, the vocal cords are drawn together by muscular action, and the adduction is completed by the reduction in pressure between the edges of the vocal folds due to high-velocity air which is blown from the lungs through the glottis. Very rapidly, sub-glottal pressure builds up as the lungs continue to expel air, until the pressure necessary to blow the vocal cords apart is reached. The cords separate under this pressure, allowing a pulse of air to be injected from the over-pressured sub-glottis through the gap between the vocal cords. This gap effectively takes the form of a Venturi tube due to the fashion in which the cords separate and as a result of the local drop in air pressure in the constricted passage between the cords, coupled with elastic tensions in and on the vocal cords and the relieved sub-glottal pressure, the vocal cords are forced back towards each other [14,15]. The instant of glottal closure is usually the point at which excitation of the vocal tract is most powerful [16,17]. This process is repeated at the fundamental frequency rate, being between 50 and 250 times per second

in the voiced speech uttered during ordinary conversation for male voices [18].

Figure 2.3(a) shows a typical pulse train as produced by the vocal cords. Due to its periodicity, the power spectral density of the vocal tract excitation function exhibits an harmonic structure in which the strongest component is normally the fundamental repetition rate and the fall-off of spectral power with increasing frequency is of the order of -12dB/octave (see Fig. 2.3(b)) [19,20].

The combined effect, then, of the vocal tract filter function and the vocal cord excitation source is to produce a pressure waveform emanating from the lips whose power spectral density will reflect the periodicity of the excitation source (Fig. 2.4(a)) and, in addition, will contain concentrations of energy at the formant frequencies F1, F2, F3, etc. corresponding to the filtering action of the vocal tract. The lips themselves typically act as an acoustic horn radiator, thereby imparting an emphasis of +6dB/octave to the overall power spectral density [21]. An example of the power spectral density for the waveform corresponding to the vowel "ee" is shown in Fig. 2.4(b).

In terms of a mathematical model approximating to the vocal tract filter function, V(s), this can best be provided by an all-pole transfer function, the number of poles N and their locations s_n corresponding to the positions of N formant frequencies Fn in the power spectral density, in addition to a low frequency zero s_ℓ to account for spectral emphasis due to the lip radiation characteristic, that is:

$$V(s) = (s-s_\ell)\prod_{n=1}^{N}\left(\frac{1}{(s-s_n)(s-s_n^*)}\right) \qquad (2\text{-}1)$$

where s is the complex frequency variable in the continuous Laplace transform domain [22,23,24].

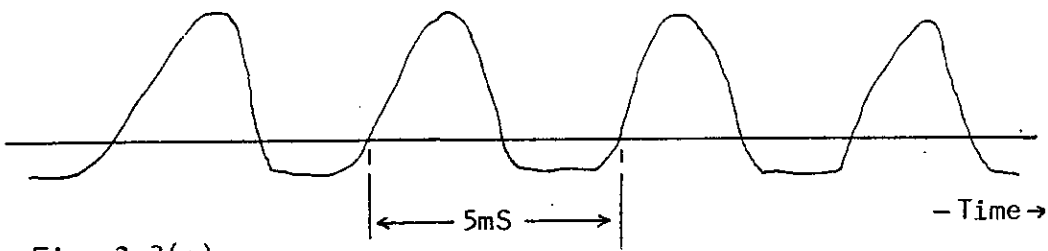


Fig. 2.3(a)

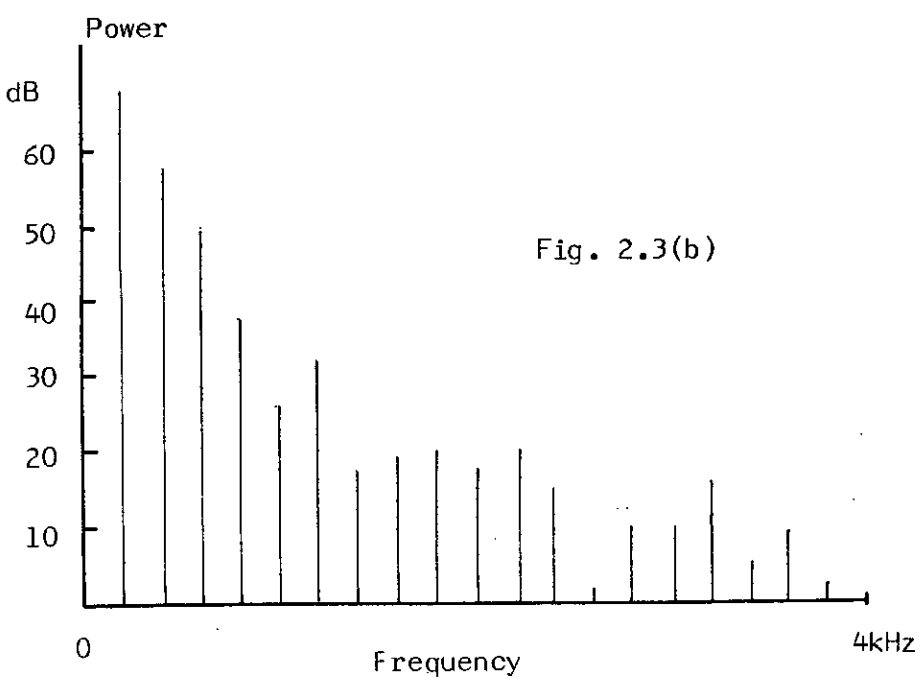


Fig. 2.3(b)

Fig. 2.3 (a) Vocal tract excitation waveform produced by the vocal cords and (b) corresponding power spectral density over a 25mS period.

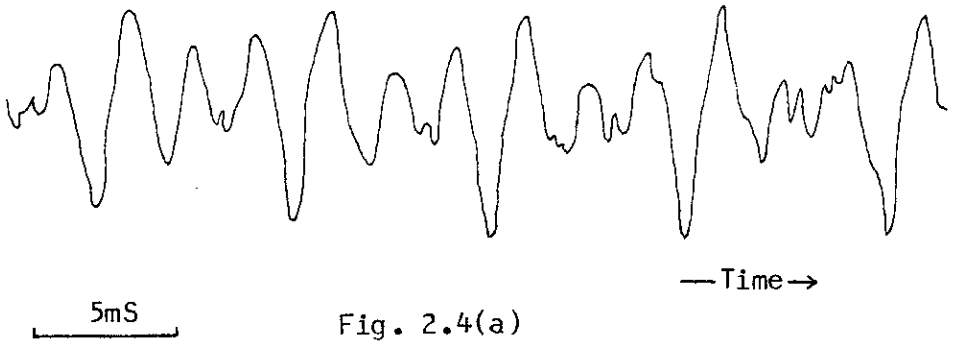


Fig. 2.4(a)

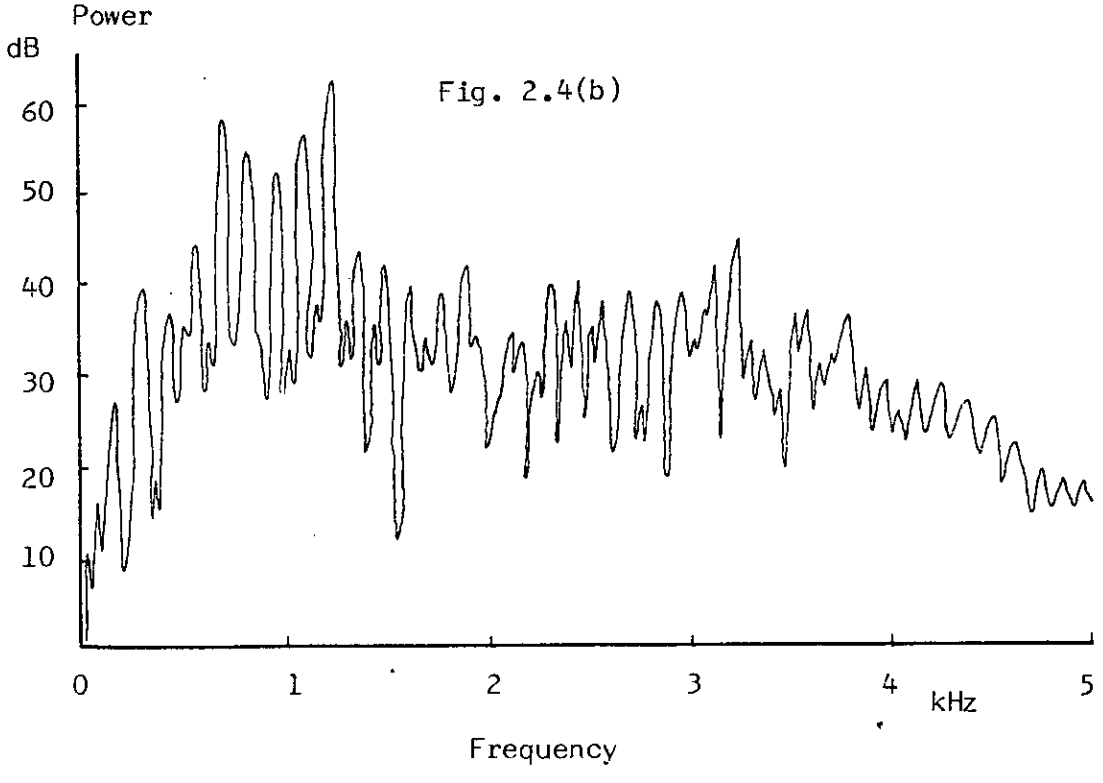


Fig. 2.4(b)

Fig. 2.4 (a) Acoustic waveform for the voiced vowel /a/ and (b) corresponding power spectral density over a 25mS period.

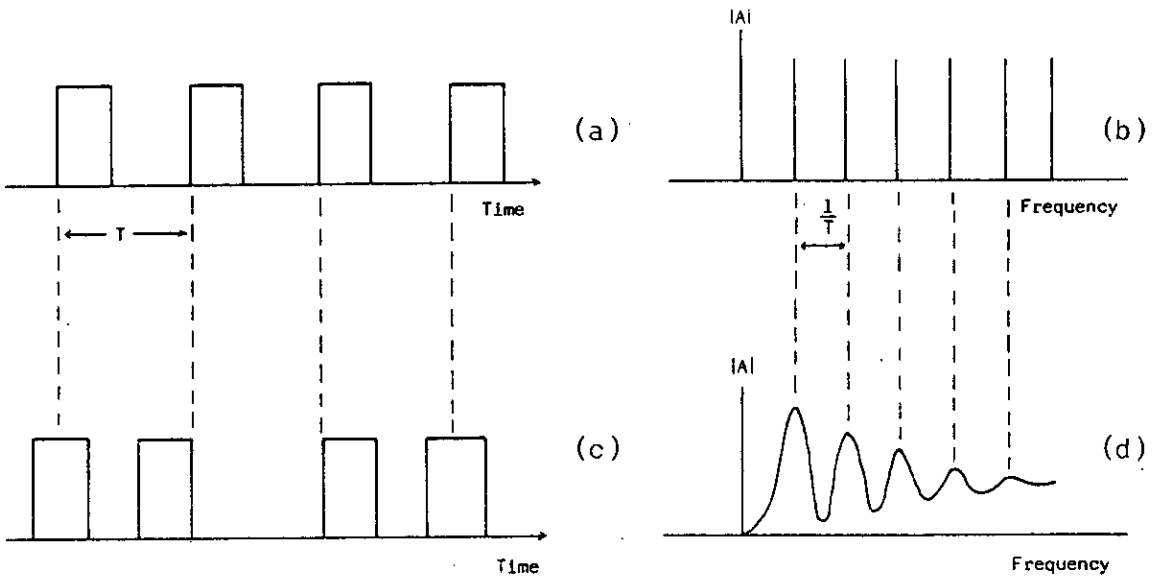


Fig. 2.5 (a) Periodic pulse train and (b) corresponding line spectrum. (c) Pulse train with jitter and (d) power spectrum.

Finally, note that the vocal tract excitation function is not, in fact, absolutely periodic, but rather is quasiperiodic in the short term over 20-50mS. The effect of this is to smear the energy of each spectral line in Fig. 2.3(b) according to the amount of jitter around the mean fundamental frequency [25]. This is illustrated in Fig. 2.5, which shows an exactly periodic pulse train (Fig. 2.5(a)) and its corresponding spectrum (Fig. 2.5(b)). Notice that when jitter is introduced into the waveform (Fig.2.5(c)), the spectral peak spacing remains constant, but the corresponding energy for each line is smeared (Fig. 2.5(d)).

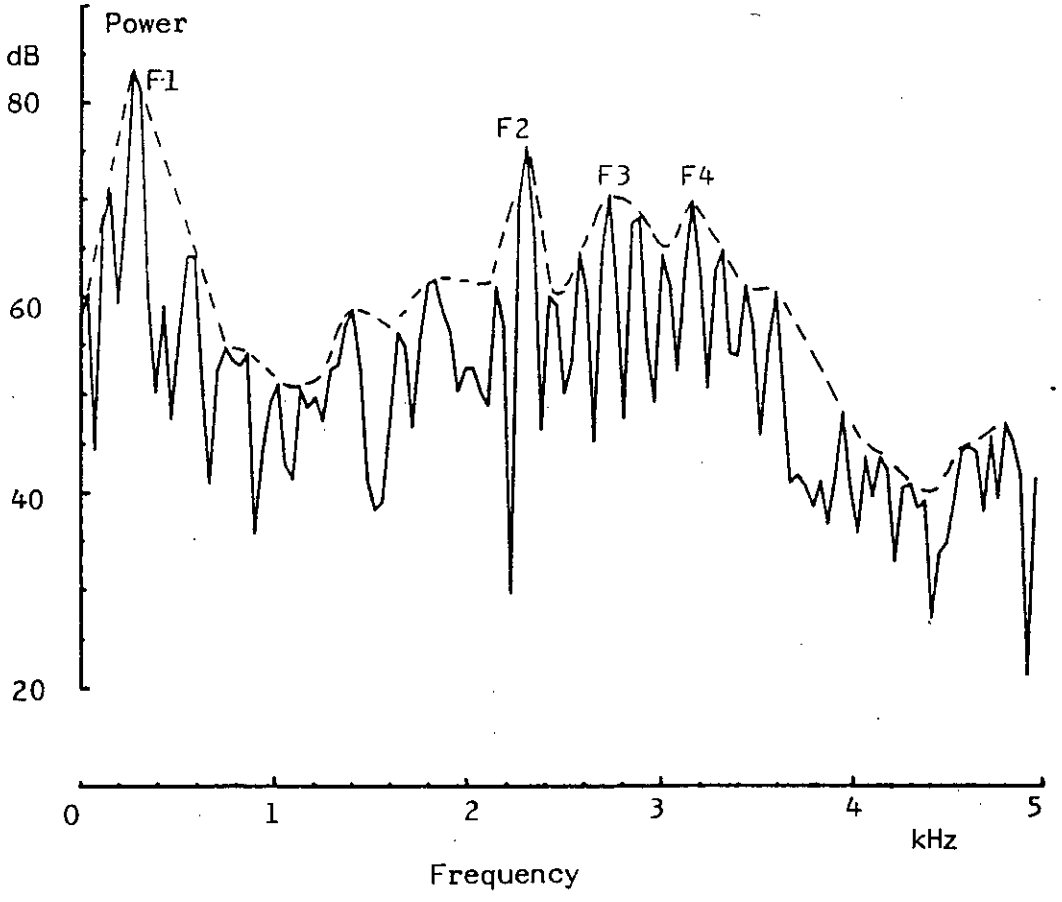


Fig. 2.6 (a) Power spectral density (PSD) for a 25mS segment through the vowel "oo" .

2.1.2 PERCEPTUAL CORRELATES OF THE VOICED ACOUSTIC WAVEFORM

The quality of the voiced speech sounds as perceived by the listener, or phonemes, depends on several acoustic attributes of the received waveform. Speech intonation, for example, is a function of the variation in time of the fundamental frequency, whereas the pitch of the perceived speech signal is, strictly speaking, the psychological impression of the tone of voice, and is a function of both the fundamental frequency and overall intensity of the voiced waveform [26]. However, the term 'pitch' is often used to signify only fundamental frequency, and indeed the two terms will be used interchangeably throughout this thesis.

Individual voiced phonemes, such as the vowels /a/ and /e/, are

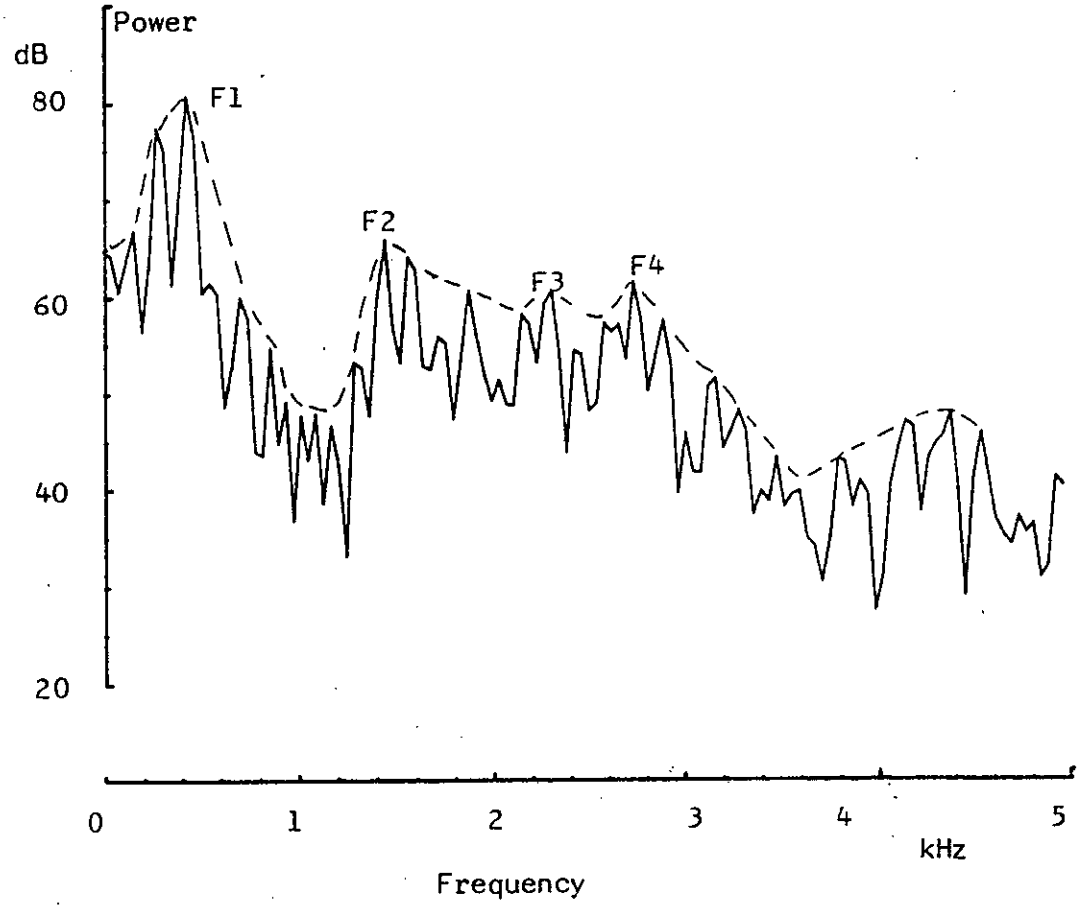


Fig. 2.6 (b) PSD for the vowel "ee".

recognised by the listener according to the centre frequencies of the resonance peaks, or formants, present in the short-time power spectrum of the signal, see Figs. 2.6(a)&(b). The most important factor in the recognition of individual vowel phonemes is the relative ratios of these formant centre frequencies to each other, as opposed to their absolute frequency values [10,11,19,27,28,]. It is this property which helps explain why the listener can perceive exactly the same vowel phoneme spoken by a small boy, whose absolute values of formant frequency will be high, and an adult male speaker, whose absolute formant frequencies will be much lower [29]. Although the corresponding power spectrum for any particular phoneme will in general contain many formants

it is generally accepted that the first 3 or 4 only play the most crucial role in speech perception [19,27,30,31]. As an approximate guide, the range of variation of the voice fundamental frequency (F0) and first 4 formant centre frequencies (F1-F4) for vowel sounds uttered by average male subjects is as follows [32]:

F0 - 60-240Hz

F1 - 150-850Hz

F2 - 500-2500Hz

F3 - 1500-3500Hz

F4 - 2500-4500Hz

Female subjects have an average one octave higher fundamental pitch, but only 17-20% higher formant frequencies.

The relative amplitudes of formants have some bearing upon the perceived quality of the particular phoneme [33,34]. For example, increasing the energy in the acoustic waveform to make the phoneme appear louder is reported to affect the intensities of the formants from the second formant upwards, whereas the first formant remains relatively unaffected [17]. However, increased vocal effort in this way is not expected to alter the formant centre frequencies to any great extent; at best perhaps by some 10-20Hz.

Another most important perceptual quality is that of nasality. Theoretically, a nasal quality may result from the coupling of any resonating side-branch, that is, an acoustic shunt element, to the main vocal tract. However, the main side-branch resonator contributing to the nasal quality of voiced sounds is, in air at normal temperatures and pressures, the nasal cavity itself. Refering to Fig. 2.1, it may at first appear that rotating the velum down and away from the back pharynx wall is in itself sufficient to allow air-flow into the nasal

cavity and hence induce nasal resonance. In reality, however, nasal resonance is a complex function related to nasal chamber dimensions, and in particular to the ratio of the nasal port openings and the area of the velic opening in the vocal tract [14,35,36].

The acoustic correlates of resonance in the nasal cavity are firstly, that nasal resonances and associated anti-resonances are introduced into the vocal tract transfer function, leading to a modified power spectral density distribution. The most commonly reported nasal formant frequencies are between 200-300Hz and around 1kHz [37], with another around 2kHz, their associated bandwidths being of the order of 300Hz for the lowest formant increasing to 1kHz for those near to 2.5kHz [38]. Nasal anti-resonances (anti-formants) are each paired with a nasal formant, and values of between 500-700Hz have been reported for the lowest [39], with another between 900Hz and 1.8kHz [37].

Secondly, as regards the acoustic waveform overall, the most general effect is an overall loss of power, which can be directly attributed to the introduction of the nasal anti-resonances, which absorb acoustic energy, especially in the higher frequencies. A further effect of nasality is the detuning of existing vocal tract formants such that their bandwidths increase, with an attendant decrease in formant peak amplitude [40]. This general attenuation is considered to be responsible for the drop in intelligibility of nasal voices [41,42].

2.1.3 UNVOICED FRICATIVE SPEECH PRODUCTION

In the case of frication, the vocal tract can still be approximated by a non-uniform acoustic tube; in this instance, however, the excitation source, rather than being a train of quasiperiodic pulses, is best

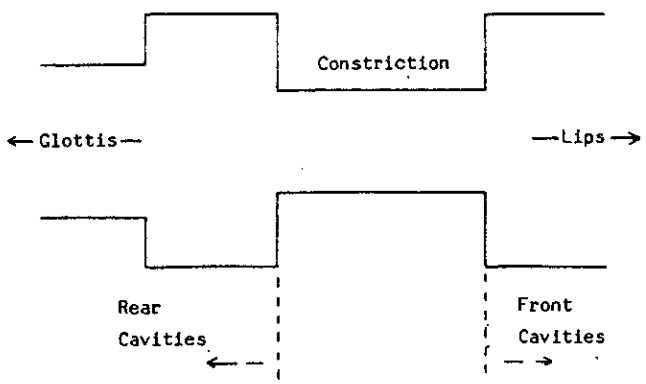


Fig. 2.7 (a) Acoustic tube model for frication.

approximated by a random train of impulses or a white noise source. In consequence, for any particular vocal tract configuration corresponding to a particular vowel, the power spectrum would in this case, exhibit no line structure, and formant amplitude would be expected to decrease whilst formant bandwidth would increase [43]. However, the production of the excitation source itself produces other changes in the power spectrum. Specifically, the excitation source is no longer situated at the vocal cords, which are held apart by muscular tension to permit the free passage of air, but rather the source is located at a point of constriction within the vocal tract, such as by the bringing together of the teeth and lower lip, as in /f/, or by the action of the tongue, as in "sh" [18]. The forcing of air at high velocity through the constriction produces turbulent flow in the vicinity of the constricting passage and possibly also at the teeth. Noise is generated as a result of the turbulent flow, and it is this noise

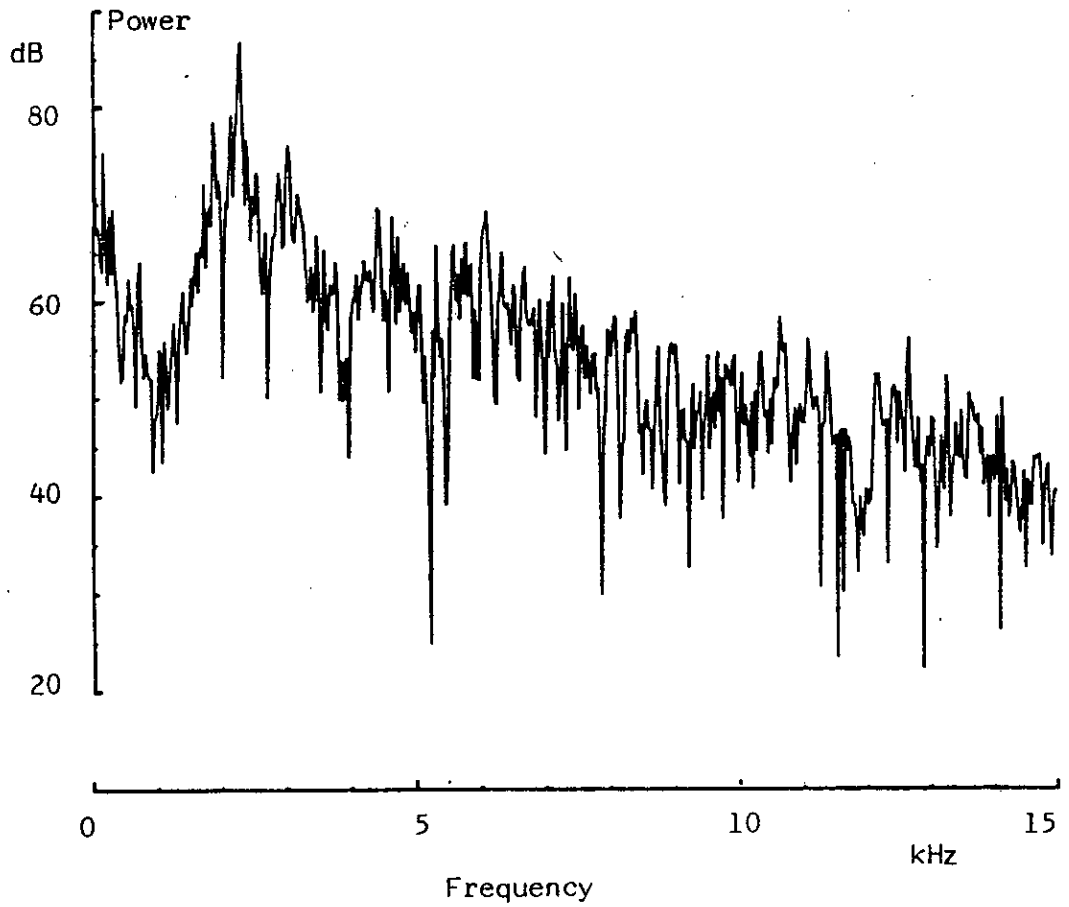


Fig. 2.7 (b) Power spectrum for the fricative "sh".

which acts as the excitation source for the frontal cavities between the constriction and the external atmosphere [19]. An example of the acoustic tube model for frication is shown in Fig. 2.7(a). In the case of the fricative "sh", the constriction shown is formed by the tongue and the hard palate. The power spectral density corresponding to "sh" is shown in Fig. 2.7(b). Notice that the spectrum exhibits a formant-like structure.

From the equation for the characteristic acoustic impedance, Zo, for a section of tube [44],

$$Zo = \frac{\rho c}{A} \qquad (2\text{-}2)$$

where ρ = gas density, c = velocity of sound, A = cross-sectional area,

it can be seen that there may be acoustic coupling between the frontal cavities and the rear cavities from the glottis to the constriction for those rear cavities having natural frequencies close to the frequencies of minimum impedance of the constriction. The classical interpretation of this effect on the vocal tract transfer function is to place zeroes in the vicinity of the rear cavity vocal tract poles in the low and high frequencies, corresponding to large values of constriction acoustic impedance, such that the effect of these poles on overall power spectrum is minimised. However, the spectral characteristic will still exhibit a definite formant structure [10,19]. Thus the vocal tract transfer function for the case of frication is now of the form:

$$V(s) = (s-s_\ell)\prod_{n=1}^{N}\frac{1}{(s-s_n)(s-s_n^*)}(s-s_g) \qquad (2\text{-}3)$$

Note that the acoustic horn radiation characteristic of the lips is still indicated by the initial zero term, and the additional real-axis zero at s_g accounts approximately for losses of energy due to the turbulent air-flow over the constriction.

2.1.4 PERCEPTION AND FRICATIVE PHONEMES

Since the power spectral density exhibits a formant structure for fricatives, then as might be expected, perception of individual fricative phonemes is based to some extent on the relative centre frequencies and bandwidths of formants, particularly in the case of fricatives whose relative power is high, such as /s/ and /ʃ/ ("sh" as in 'show')[26].

However, for certain of the lower intensity fricatives, such as /f/ and /θ/("th" as in 'think'), there is an important contribution to perception from the segmental environment of the consonant; that is to

say that, in the case of a consonant preceded or followed by a vowel, then perception of the fricative phoneme is based not only on spectral attributes of the consonant itself, but also on formant transitions into the vowel, on the directions of these transitions, and on other secondary acoustic characteristics such as vowel duration and relative ratio of power from vowel to consonant [45,46]. This perceptual mechanism is perhaps best explained by a consideration of the production and perception of the next class of unvoiced speech sounds, plosive or stop consonants.

2.1.5 PLOSIVE CONSONANT SPEECH SOUNDS

In plosive or stop consonant production, there is a build-up of pressure at some location due to closure of the vocal tract by, for example, the teeth and lips. The sudden release of pressure causes a transient excitation of the vocal tract which results in the sudden onset of sound. The characteristic power spectral density for any particular plosive consonant, such as /p/ and /t/, is considered to be concentrated within some particular region of the power spectrum [47,48,49] For the plosive /t/, for example, the main concentration of spectral power is in the region above 4kHz.

2.1.6 PERCEPTION OF PLOSIVE PHONEMES

It is generally considered that the spectral concentration of energy associated with the plosive burst is not, in itself, sufficient to allow identification of a particular plosive phoneme. Experiments have shown, for example, that a plosive burst centered at one frequency may

be heard as /k/ when associated with one vowel, but as a /p/ when associated with another vowel, implying that identification of a particular plosive is dependent to some extent on the nature of the following vowel [50,51,52]. Further experimentation has since shown that the most important attribute of the following vowel in this respect relates to the direction in frequency and duration in time of the second formant transition. Specifically, in the absence of a plosive burst itself, all second formant transitions perceived as one particular plosive phoneme have a virtual point of origin within the same frequency region [53,54]. This principle is illustrated in Fig. 2.8, which demonstrates that a silent interval, containing no plosive burst energy whatsoever, followed by the 2nd formant frequency transitions as indicated and combined with the first formant as shown, will produce the perceptual impression of having heard the stop consonant /t/. Note that all second formant transitions have a virtual point of origin in frequency, as shown by the dotted arcs, in the region of 1.8kHz. Similar results may be obtained for a /p/ plosive (virtual origin=700Hz) and a /k/ plosive (virtual origin=3kHz) [50].

2.1.7 PERCEPTION OF THE COMPOSITE ACOUSTIC WAVEFORM

Just as the perception of certain fricative consonants and plosives is dependent on formant transitions following the speech sound itself, so the perception of the composite acoustic waveform of normal speech reflects a perceptual system based on a complex interplay of acoustic events and perceptual cues involving not only subsequent formant transitions from consonant-to-vowel, but also preceding formant transitions from vowel-to-consonant, and other factors such as temporally

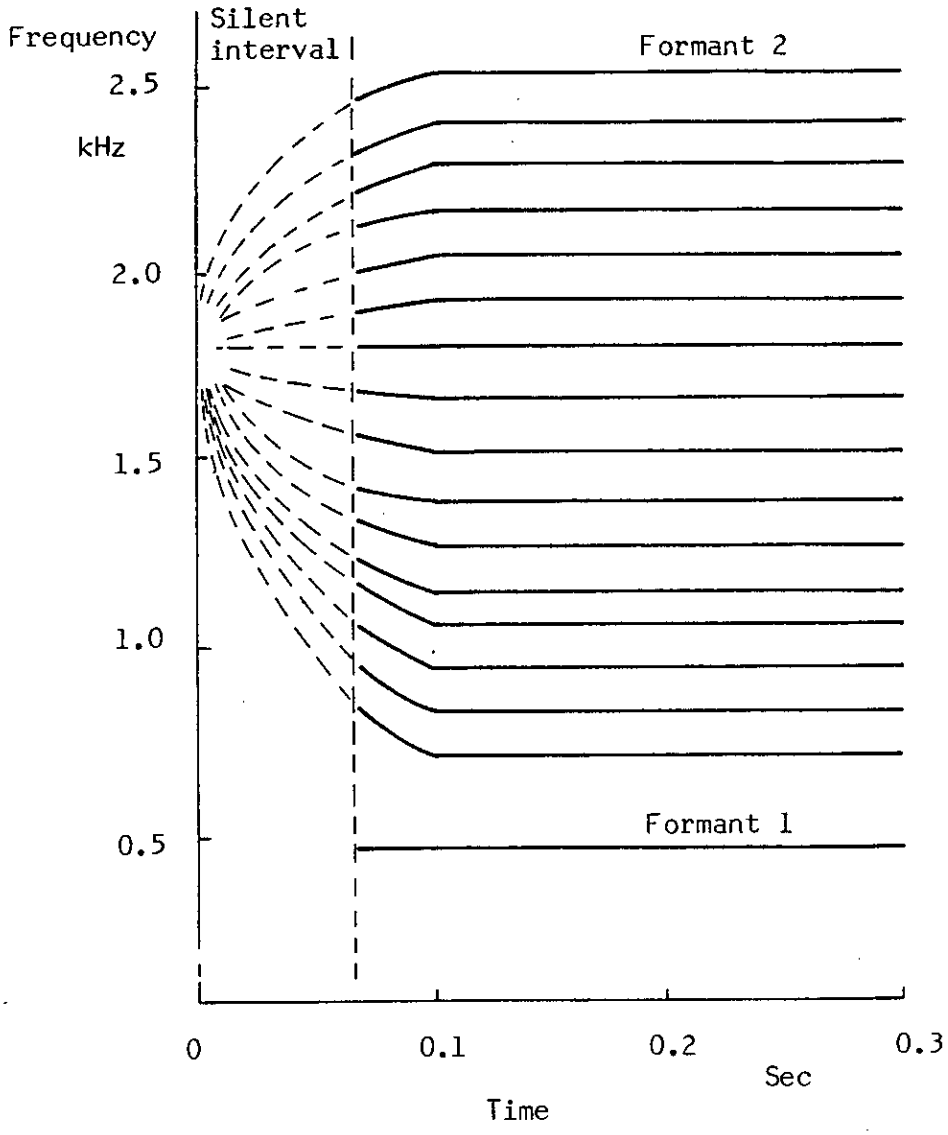


Fig. 2.8 Second formant transitions all perceived as the stop consonant /t/ (from (26)).

fluctuating intensity ratios from sound to sound, changes of speech production from voiced to unvoiced and indeed, combinations of both of the latter classes of speech production as in, for example, the voiced fricative /z/ [55,56,57,58]. Differentiation between the consonants /b/ and /p/, for example, depends not only on vowel formant transitions but also on the low intensity voicing present up to the instant of plosive release [28,52]. In particular, whilst voiced nasal consonants may be identified from changes in formant bandwidth and relative amplitude (see section 2.1.2), their perception depends also on transitions of the formants of the following vowel [59]. Refering to Fig. 2.8, if the silent interval before the commencement of second formant transition is replaced instead by a low intensity buzz, then the same formant trajectories now give rise to the perception of a nasal consonant; that is, those transitions with a virtual origin at 700Hz are perceived as /m/, whereas those with a virtual origin of 1.8kHz are perceived as /n/ [60].

In conclusion, although the human perception of speech is seen to be characterised by a highly developed and complex architecture, its structure nonetheless leads to a robust mechanism which is highly tolerant of diurnal variations and talker-specific idiosyncrasies in speech production [61].

The importance of perceptual cues in the identification of specific phonemes is highlighted in experiments to find difference limen for vowel formant centre frequencies. It has been found that, for example, the shift in second formant frequency needed to produce a just-noticeable difference in vowel identity for vowels spoken within a consonant-vowel-consonant framework is much higher than that for vowels uttered in isolation [62,63]. This property is exploited by speakers in everyday

conversation, since it is rare in any case that a speaker achieves exactly the same formant frequencies for a given vowel from instance to instance; rather the formant frequencies are placed into approximate frequency areas with phonemic identification aided by one or several of the factors outlined above [64].

A further parameter affecting perception and intelligibility, which has particular reference to helium speech, relates to the effects of variations of the relative power ratios of consonant to vowel. It has been shown that a small consonant-to-vowel power ratio inherently favours intelligibility, whereas an increased power ratio has an adverse effect upon intelligibility [65].

Experiments have demonstrated, for example, that in the presence of speech signal distortion, decreasing the consonant-to-vowel intensity ratio will aid intelligibility, and that in such cases, the effects of peak-clipping and amplitude compression on the signal are in fact advantageous [66], whereas increasing the speaker's overall vocal effort tends to increase the vowel-consonant intensity ratio and therefore has a detrimental effect upon intelligibility. Modulation of the consonant-to-vowel intensity ratio is also an important device employed from time to time by talkers in order to reduce articulatory effort, supplementing articulation by the variation of the intensity ratio in order to produce a structured signal conveying message information [56,67].

2.2 THE EFECTS OF PRESSURE AND GAS MIXTURE ON SPEECH INTELLIGIBILITY

2.2.1 SPEECH IN A HIGH PRESSURE AIR ATMOSPHERE

Since the transmission medium for the speech signal is the atmosphere surrounding the talker, and since the signal itself is a pressure wave propagating through the medium, then the pressure waveform is dependent upon the physical properties of the medium. From acoustic theory, the speed of sound, c, in a gas is given by:

$$c = \left(\frac{\gamma P}{\rho}\right)^{\frac{1}{2}} \tag{2-4}$$

where γ is the adiabatic constant (ratio of specific heats), P is the gas pressure and ρ the density of the gas. It can be shown, however, that the speed of sound for a given gas is relatively independent of ambient pressure, since γ is, to a first approximation, independent of pressure and temperature over the ranges of temperatures and pressures survivable in the diving environment. Furthermore,

$$\rho = \frac{\rho_o P}{P_o} \tag{2-5a}$$

$$\Rightarrow \quad \frac{\rho}{\rho_o} = \frac{P}{P_o} \tag{2-5b}$$

where ρ_o is the density measured at some pressure P_o.

Substitution of equs. (2-5) into equ. (2-4) illustrates that the speed of sound is independent of pressure for a given gas. Thus, it might be expected that the mechanism and perception of human speech is relatively unaffected at high ambient air pressure: experiments have shown, however, that there is a degradation of the intelligibility of speech uttered in such an environment [68,69,70]. This drop in intelligibility has been attributed in the main to (a) the increased nasality of high-pressure air speech and (b) to a nonlinear shift of the lower

formant frequencies [71].

It has also been demonstrated that unvoiced consonant sounds suffer an attenuation of the order of -10dB with respect to voiced vowel sounds at high pressure [69]; therefore, in consideration of the discussion in section 2.1.3 relating to the effect of the vowel-consonant intensity ratio upon intelligibility, an increase in this ratio will inherently disfavour intelligibility. Attributes (a) and (b) have been related theoretically to the acoustic impedance properties of the respiratory gas and the tissue of the face and vocal tract [72]. Specifically, at normal temperatures and pressures the vocal tract walls can be essentially regarded as perfect reflectors of acoustic energy due to the large mismatch of impedance between the transmission medium and the skin tissue.

However, the characteristic acoustic impedance, Zo, of each air-filled vocal tract cavity, from equ.(2-2), increases with density, i.e. ambient pressure. At high pressures, therefore, the vocal tract walls may in fact absorb acoustic energy as the impedance mismatch is reduced, thereby producing resonatory vibrations of the wall tissue which will, in turn, affect the pressure wave produced by the vocal mechanism. This effect has been quantised by a theoretical consideration of the total volume of air enclosed by the vocal tract and acoustic properties of the vocal tract walls [69]. The result is an equation relating the resonance frequency of the closed vocal tract, Fw, to the ambient pressure P:

$$Fw = Fwo \left(\frac{P}{Po}\right)^{\frac{1}{2}} \qquad (2\text{-}6a)$$

and, substituting equ.(2-5b) gives:

$$Fw = Fwo \left(\frac{\rho}{\rho_o}\right)^{\frac{1}{2}} \qquad (2\text{-}6b)$$

where Fwo is the closed tract resonance frequency at pressure Po. At normal atmospheric pressure, Fwo is considered to be of the order of 150-200Hz [73].

The resonating vocal tract walls, under conditions of high ambient pressure, can be considered to act as a side chamber shunting the main vocal tract, and would therefore be expected to add a low frequency pole and zero to the vocal tract transfer function in a similar manner to the effect of nasal cavity resonance [39].

The resulting deterministic relationship between formant centre frequency in high pressure air, Fpn, and original formant frequency, Fn, at some pressure Po is given by:

$$Fpn = (F_n^2 + F_w^2)^{\frac{1}{2}} \qquad (2\text{-}7)$$

with the proviso that equ.(2-7) is valid only on an average basis [74], since the shunting effect of the vocal tract walls is not likely to be uniformly distributed. Published results [75] have been roughly in agreement with equ.(2-7), and an illustrative example is shown in Fig. 2.9. Notice that the nonlinearity of the formant frequency translation characteristic is most pronounced in the low frequency region of the speech spectrum. In addition, it has been argued that the bandwidth of the vocal tract resonances is not radically changed with increasing pressure. Coupling of a shunt element across the vocal tract would, of course, be expected to produce an increased damping and hence increased bandwidth of those vocal tract poles in the vicinity of the shunt element poles and zeros. Here, however, the effects of the high pressure atmosphere are such that the low frequency region most affected in this respect is translated upwards in frequency with the lowest frequencies shifted by the greatest amount, the effect of which is to nullify any low frequency formant damping.

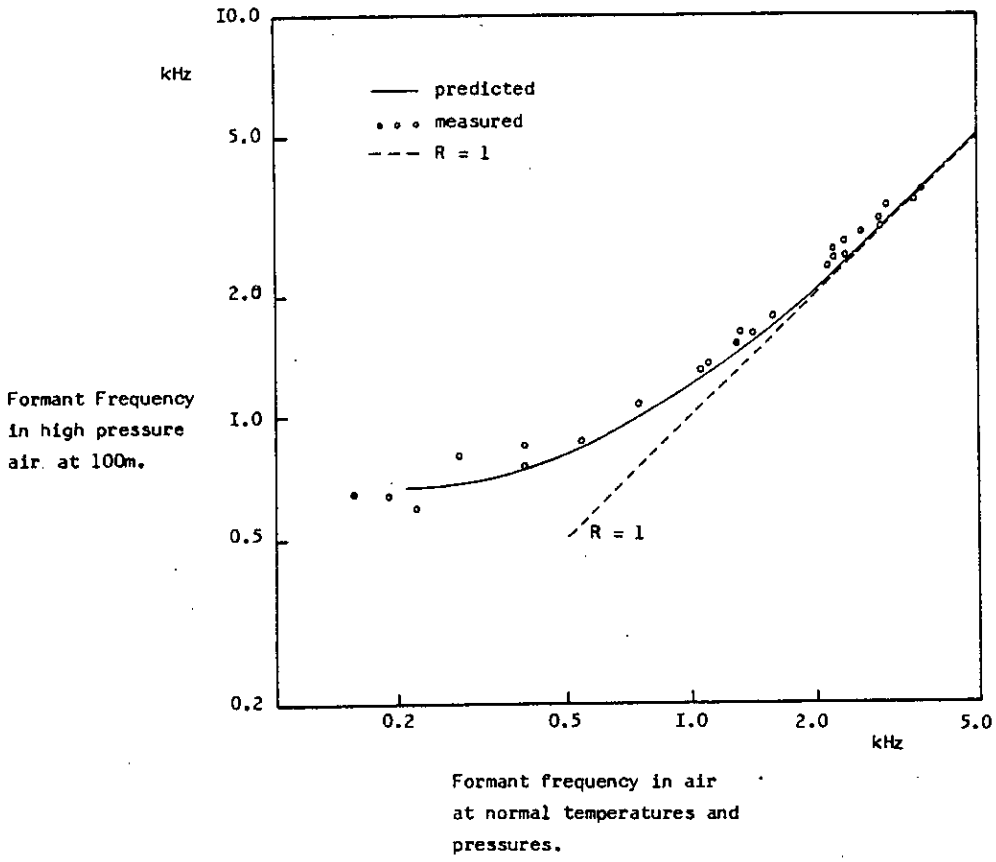


Fig. 2.9 Formant transposition for one subject in high pressure air (from (74)).

A consequence of this trend is that certain voiced sounds having an otherwise very low first formant will increase in intensity at high pressure because the low first formant Q factor is effectively increased by the upwards frequency shift. The combined effect of this, in addition to acoustic radiation idiosyncrasies related to the mechanism of production of voiced and voiceless sounds, is to promote an increase in the sound pressure level of voiced to voiceless sounds by a factor roughly proportional to $P^{\frac{1}{2}}$ (69,74).

Thus, in terms of the mathematical correlates of the effect of high ambient pressure, (a) for voiced sounds, equ.(2-1) is still valid in

that the vocal tract filter function can still be described by N poles, although the frequency locations will be moved upwards, with low frequency poles being most affected; (b) for unvoiced sounds, again the general form of equ.(2-3) is valid, except that, in this case, whilst the pole frequencies are not shifted, there is however either increased damping of the poles or some general attenuation factor in order to account for the increased vowel-to-consonant intensity ratio.

As regards the increased nasality of high pressure air speech, this has been directly attributed to the increase in frequency of the low frequency formants [69,74]. The degradation of the intelligibility of speech at high ambient pressure has been correlated to this increase in the nasal quality of the speech, and also due to the rise in the first formant (F1) frequency towards that of the second formant (F2), in that auditive distinction between F1 and F2 is impaired thereby affecting the perceived phoneme [70]; the increase in the vowel-consonant intensity ratio also deteriorates the intelligibility of high pressure speech, and plosive and fricative consonants are found to be worst affected in this respect [76].

2.2.2 EFFECTS OF PRESSURE AND GAS MIXTURE ON FUNDAMENTAL FREQUENCY

Early results, shown in Fig. 2.10, demonstrated a slight decrease in pitch period when breathing helium gas under laboratory conditions at normal ambient temperatures and pressures [77]. However, this effect has been attributed to a physical contraction of the larynx muscles, since the helium gas was colder than room temperature.

Acoustic theory would predict no increase in pitch period as a result of increased helium concentration and little change, if any,

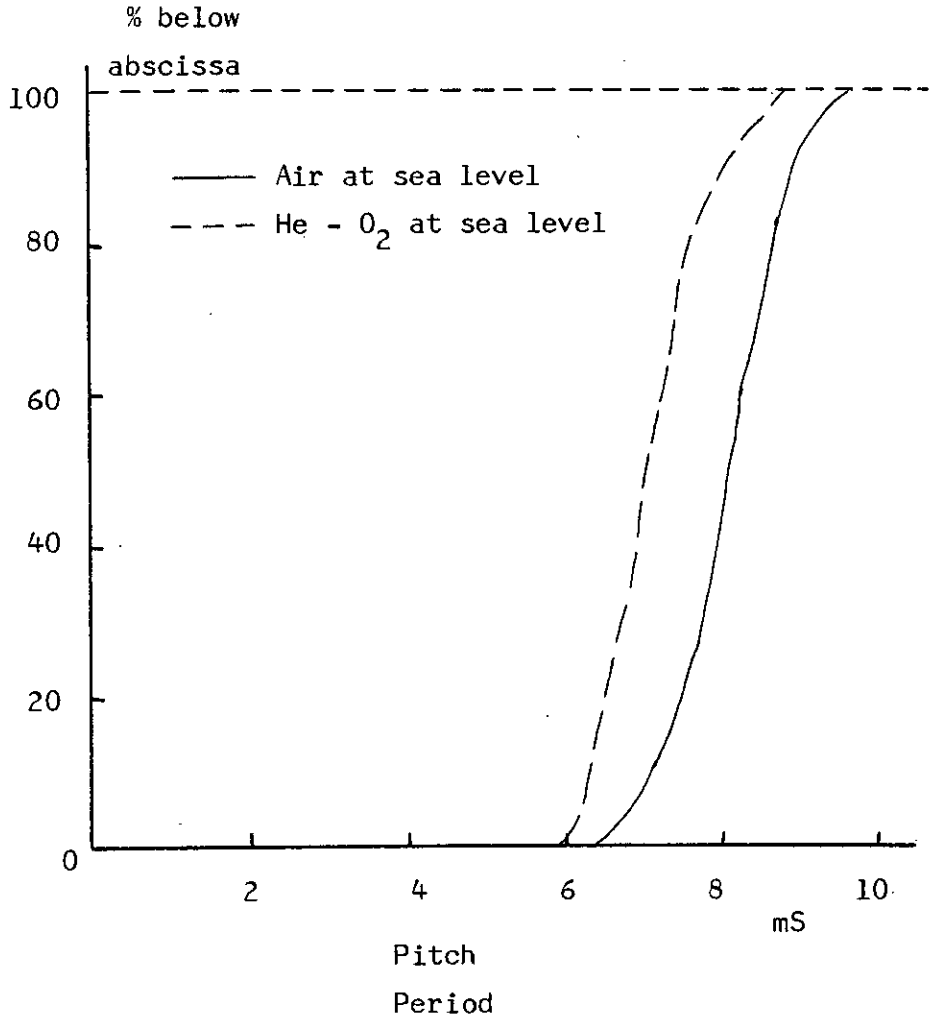


Fig. 2.10 Cumulative distributions of pitch period in air and heliox at sea level (from (77)).

due to an increase in ambient pressure, as indeed might be expected from what is in the main a muscularly controlled event. Published data [78] relating the pitch period for the same utterances made in air and in a pressured helium-oxygen (heliox) environment are shown in Fig. 2.11, and demonstrate a close correlation between the pitch distribution in both atmospheric air and a heliox atmosphere, simulated at pressure in a surface-based compression chamber, on the same day. In contrast, the pitch distribution measured some three months later for the same utterance in air at sea level shows a noticeable change. On the other hand, results shown in Fig. 2.12 demonstrate that pitch period in a pressured heliox atmosphere is reduced in comparison to the pitch period in air [78]. However, notice that the reduction in pitch period does not vary directly with depth since the reduction is greater for a depth of 70ft than for 200ft, and in any case such changes fall within the expected range of pitch variations for normal speech in air. Several possible causes for these observed changes in pitch period at depth have been suggested. Firstly, at depth and especially in a diving habitat, it has been observed that divers tend to speak with increased vocal intensity in an attempt to overcome background noise levels experienced in the diving chamber environment [79]. Such increases in vocal intensity are normally accompanied by an increase in fundamental frequency. Secondly, environmental effects on the diver's speech mechanism, such as changes in the acoustic loading of the vocal cords due to increased gas density, might be expected to produce some slight change in pitch period [80]. Finally, the diver may invoke modifications to his speech to alter his pitch in an attempt to enhance the intelligibility of his speech as he himself judges.

Although changes in pitch period have been measured at depth, changes

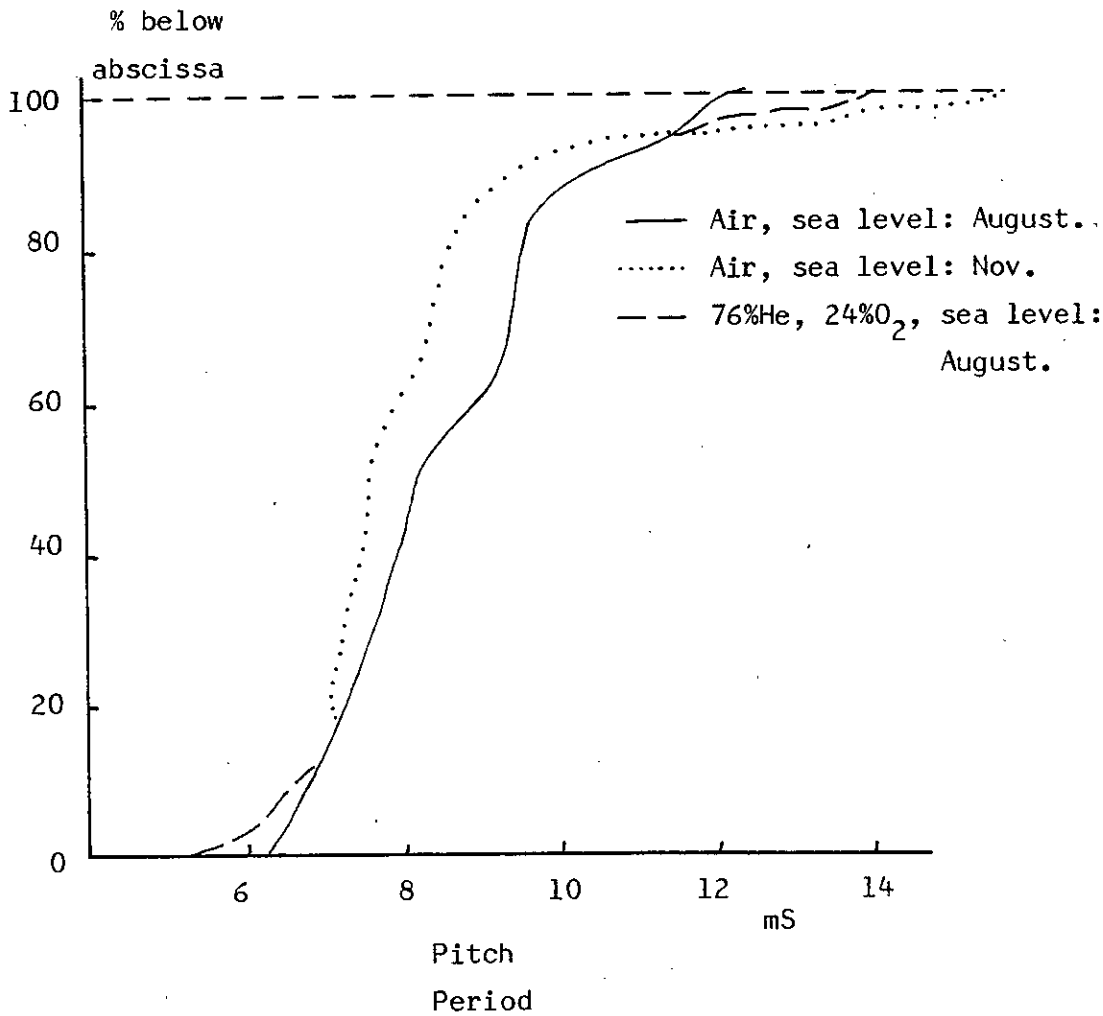


Fig. 2.11 Cumulative distributions of pitch period in air and heliox at sea level (from (78)).

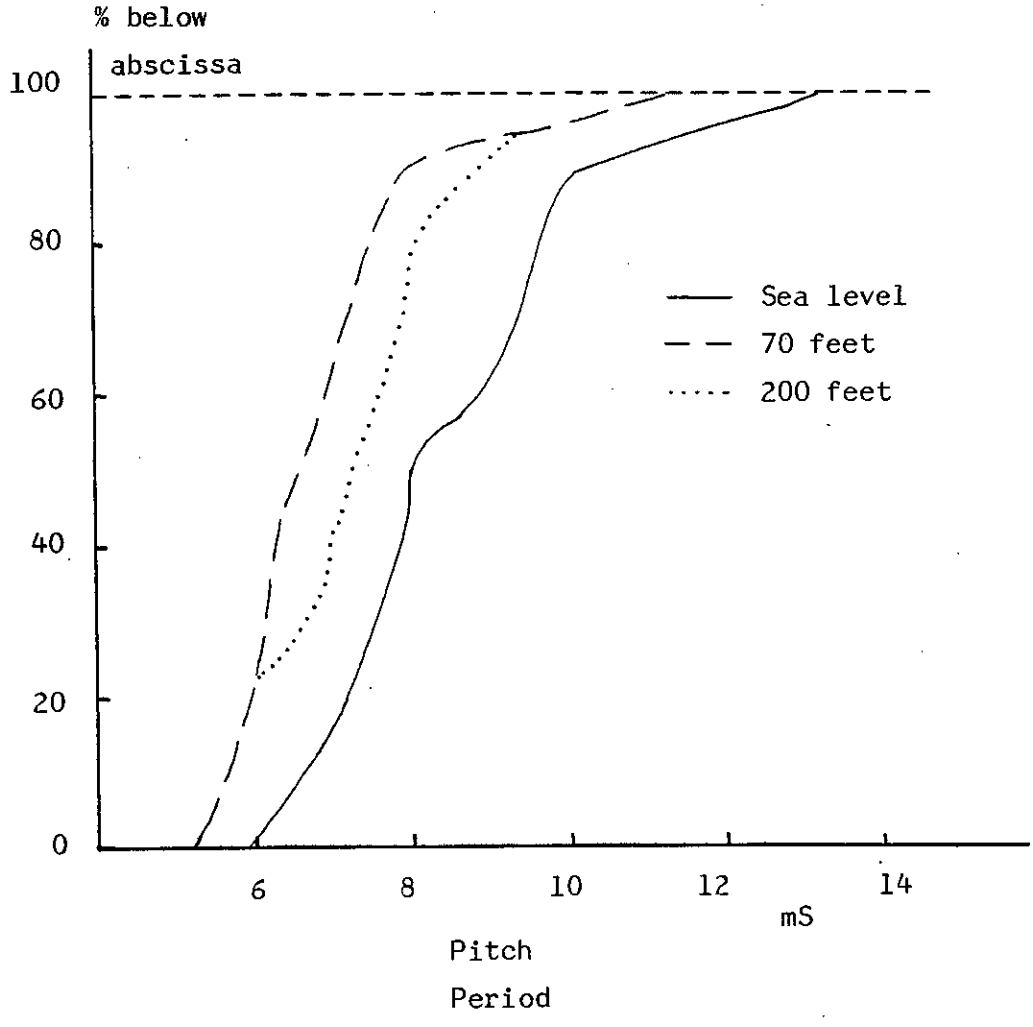


Fig. 2.12 Cumulative distributions of pitch period in a 76% helium, 24% oxygen mixture at various depths (from (77)).

of the magnitude demonstrated in Fig. 2.12 have minimal effect on speech intelligibility [81,82]. In conclusion, variations of the fundamental period of repetition of vocal cord vibration are relatively independent of both ambient pressure and respiratory gas composition. Furthermore, pressure and gas composition appear to produce little effect on the spectral characteristic of the vocal cord source [83].

2.2.3 THE EFFECTS OF PRESSURE AND GAS MIXTURE ON VOICED SPEECH

In a similar manner to speech in high pressure air, attempts have been made to quantify the effect of a pressured atmosphere containing a high percentage volume of helium on the speech spectrum, with particular emphasis on the characteristics as applied to voiced speech [74,80]. Recalling equ.(2-4) relating the velocity of sound, c, of a gas to its pressure P, density ρ and adiabatic constant γ, then the ratio R of the velocity of sound of a respiratory mixture containing helium gas, c_h, to the velocity of sound in air at the same pressure, c_a, is given by:

$$R = \frac{c_h}{c_a} = \left(\frac{\gamma_h \rho_a}{\gamma_a \rho_h} \right)^{\frac{1}{2}} \qquad (2\text{-}8)$$

where γ_h is the adiabatic constant for the mixture with helium, and:

$$\gamma_h = \sum_i Q_i \gamma_i \qquad (2\text{-}9)$$

where Q_i and γ_i are the percent volume and corresponding adiabatic constant for the ith gas in the mixture, and similarly:

$$\rho_h = \sum_i Q_i \rho_i \qquad (2\text{-}10)$$

Since the frequency of resonance of any rigid resonator is proportional to the velocity of sound, then the closed vocal tract resonance Fwh for the mixture with helium at pressure P can be found by combining equs.(2-6) and (2-8), that is:

$$Fwh = Fwa.R.\left(\frac{\rho_h P_h}{\rho_a P_a}\right)^{\frac{1}{2}} \qquad (2\text{-}11)$$

where Fwa is the closed tract resonance in air at normal pressures. Furthermore, substituting equs.(2-5b) and (2-8) into equ.(2-11) gives:

$$Fwh = Fwa.\left(\frac{\gamma_h P_h}{\gamma_a P_a}\right)^{\frac{1}{2}} \qquad (2\text{-}12)$$

Thus, assuming Fwa is measured in air at one atmosphere pressure, then for a given gas mixture, $Fwh \propto P^{\frac{1}{2}}$, and the general relationship relating formant frequency in a high pressure helium mixture, Fhn, to original formant frequency in air, Fn, is therefore given by [74]:

$$Fhn = (R^2.Fn^2 + Fwh^2)^{\frac{1}{2}} \qquad (2\text{-}13)$$

Notice that since γ varies only slightly with gas composition, then Fwh is approximately independent of gas mixture and varies as a function of pressure only. Basic assumptions relating to equ.(2-13) are (a) that the formant frequency in air Fn is that as observed by, e.g. spectrographic analysis, and (b) that the vocal tract walls act as perfect acoustic reflectors. However, since the closed tract resonance is measurable in air at normal pressures, then there is no reason to suggest that equ.(2-7) should not apply equally to normal speech in air [80], that is :

$$Fn^2 = Frn^2 + Fwa^2 \qquad (2\text{-}14a)$$

$$\Rightarrow \quad Frn^2 = Fn^2 - Fwa^2 \qquad (2\text{-}14b)$$

where Frn is the resonance (not directly observable) assuming the vocal tract to be a perfect rigid resonator, and Fwa is the closed tract resonance frequency as measured, the implication being that the term Fn in equ.(2-13) should actually be replaced by Frn as defined by equ.(2-14b). Thus, replacing the variable Fn in equ.(2-13) by Frn as defined in equ.(2-14b) and replacing explicitly the term Fwh by

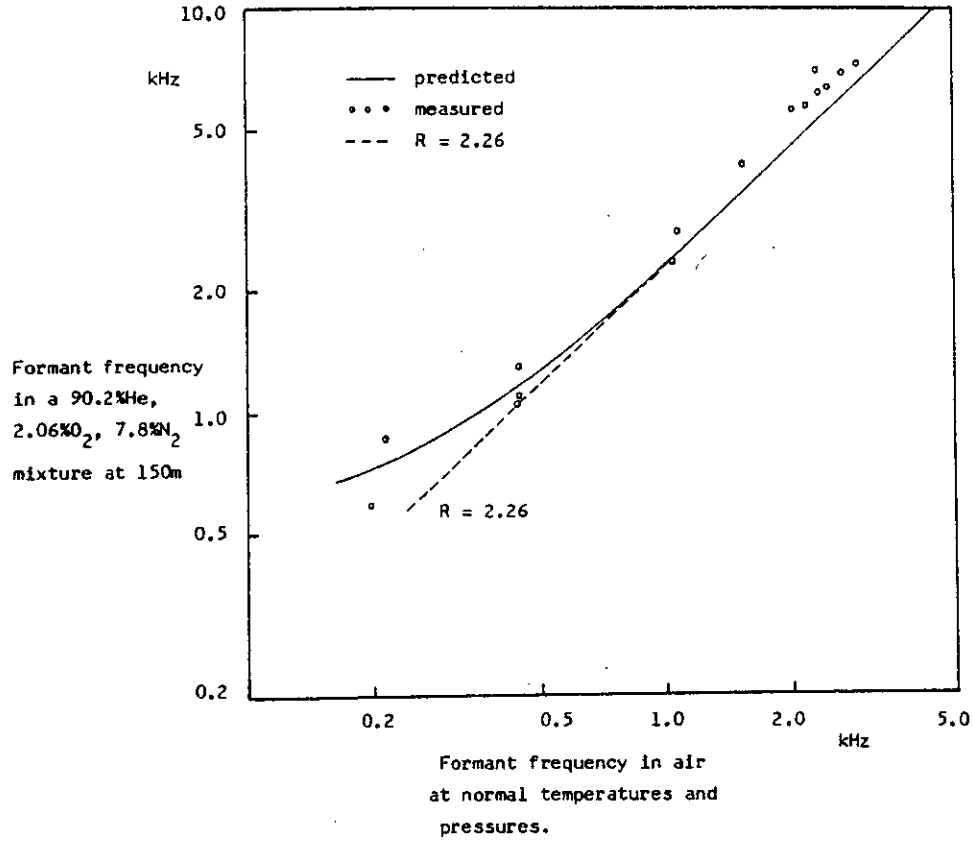


Fig. 2.13 Formant transposition for one subject in a high-pressure helium-oxygen-nitrogen mixture (from (74)).

equ.(2-12) gives:

$$Fhn = R \times \left[F_n^2 + \left(\frac{c_a^2 \gamma_h P_h}{c_h^2 \gamma_a P_a} - 1 \right) F_{wa}^2 \right]^{\frac{1}{2}} \qquad (2\text{-}15)$$

Figure 2.13 demonstrates the application of equ.(2-15) to formant data, measured for one subject breathing a pressured helium-oxygen-nitrogen mixture at depth. The value for Fwa is 180Hz in this case.

While certain published results regarding formant frequency analysis in helium [81,84,85] are claimed to approximate closely to the theoretical relationship of equ.(2-15), not all authors concur as regards

the relative effect this has on speech perception. Those whose analytical results approximate to equ.(2-15) have postulated that (a) the nonlinear formant shift characteristic, which is most pronounced below approximately 700Hz in a pressured helium environment, has a large role to play in the decrease of intelligibility of voiced helium speech [86]; others have argued (b) that it is the large shift of formant frequncies upwards in the spectrum which most affects intelligibility, since the formant shift characteristic is grossly linear[87,88,89], (c) that estimates of the closed tract resonance in air of between 150 and 200Hz are too high, and (d) that it is only at pressured helium to normal air velocity ratios of >2.0 that intelligibility is impaired[80].

There is also disagreement as to the effect of the spectral shift characteristic in respect of formant bandwidths. Some authors [80,86,90] report that formant bandwidth is unaffected by the formant frequency shift characteristic, whereas more recent research [91,92] has found that the lower formant bandwidths increase by a ratio of the order of R^2 with the bandwidths of higher formants increasing by the ratio R or even less.

There is, however, overall agreement as regards the impairment of intelligibility due to (e) high frequency attenuation in the voiced spectrum [88,93] and (f) the nasal quality of helium speech [74,88,94]. High frequency attenuation, effective beyond approximately 5-7kHz, has been attributed in the main to the fact that, whereas the vocal tract formant frequencies are shifted upwards in frequency, the vocal tract excitation source spectral characteristic remains unaltered[83,95]. Therefore, since this characteristic exhibits a power spectrum roll-off of -12dB/octave [17,96], then high frequency spectral events in the voiced speech waveform are relatively more attenuated than in a normal

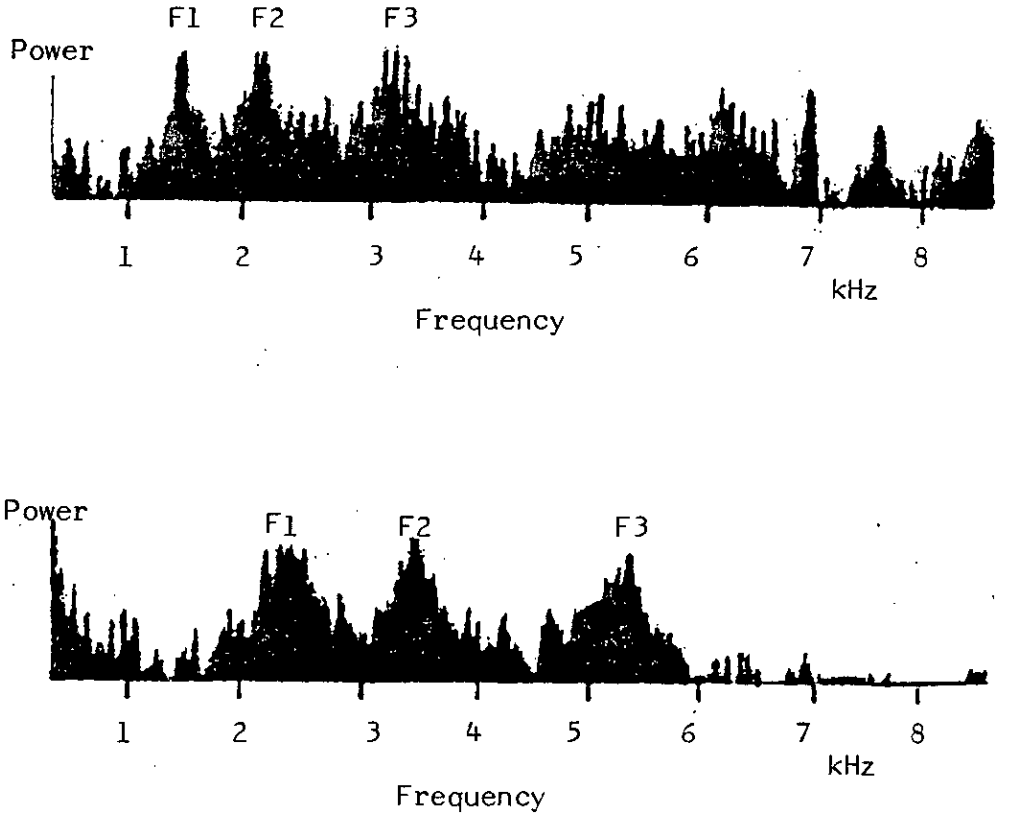


Fig. 2.14 Vowel spectra for one subject (a) in air and (b) in heliox at sea level (from (88)).

air environment. It has also been suggested that acoustic loading of the vocal cords themselves may indeed affect the high frequency region of the vocal cord source characteristic [80]. Figure 2.14 shows two speech spectra of the same vowel uttered by a speaker in air (Fig. 2.14(a)), and in a helium environment (Fig. 2.14(b)) at normal atmospheric pressure (79% Helium, 16% Oxygen). The frequency transposition ratios for formant frequencies F1, F2 and F3 are approximately 1.65:1, 1.57:1 and 1.56:1 respectively, illustrating the non-uniformity of frequency translation especially for the first formant. An increase of formant bandwidth is also apparent in Fig. 2.14(b) and, in particular, there is severe attenuation of high frequency components above 6kHz.

Thus, in terms of the effect of a high-pressure helium-rich atmosphere on the mathematical model of the vocal tract transfer function, the general form of equ.(2-1) is once again valid, in that the vocal tract can still be described by N poles, although their frequency locations are roughly transferred upwards by the ratio R. The damping factors, and hence the bandwidths of certain poles, may or may not be affected, although certainly those situated towards the higher end of the speech spectrum would appear to suffer increased damping.

Explanations as to the nasal quality of helium speech vary widely, and as yet a precise definition has not been proffered. It has been tentatively related to both (1) the increase in overall formant frequency, and in particular nonlinear shifts of low frequency formants [81,84]; and (2) increased transmission of acoustic energy into the nasal cavity itself, resulting in actual nasal resonance, with the most likely transmission path being through the tissue of the soft palate [80].

2.2.4 EFFECTS OF HIGH PRESSURE HELIUM MIXTURES ON UNVOICED SOUNDS

Although it is generally agreed that the intensity of unvoiced consonant speech sounds is severely attenuated compared to voiced vowel intensity in a high pressure helium atmosphere, the mechanism of this phenomenon is not clearly understood. Explanations offered to the present relate to high frequency attenuation linked to the physical properties of the air stream in the vocal tract constriction as compared to voicing produced by the vocal cords [74,97]: an increase in the vocal effort of the diver, which would alter the vowel-consonant intensity ratio, has also been proposed as a possible explanation [71].

The enhancement of vowel sound intensity relative to the intensity

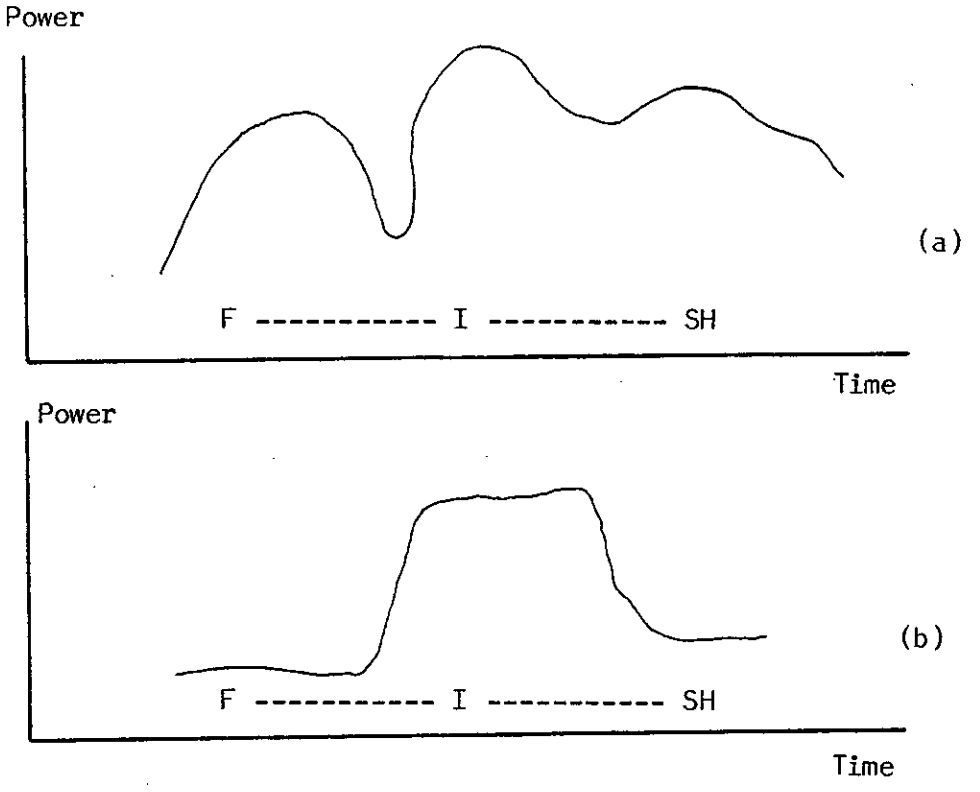


Fig. 2.15 The word 'fish' spoken (a) in normal air and (b) in heliox at a depth of 300ft (from (99)).

of consonant sounds is an important feature in the degradation of helium speech intelligibility, since consonants and the transitions between consonants and vowels have been shown to provide important cues to the next sound to be produced by a particular speaker [60]. Hence, if such cues are missing or degraded, the listener's perceptual system may be caught unawares and will effectively lag behind changes in the speech as it tries to assimilate information from the helium speech waveform, and may misinterpret and confuse vowel formant transitions, thereby reducing intelligibility [76,98].

An example of consonant attenuation in helium speech [99] is demonstrated in Fig.2.15, which compares vocal intensity as a function of

time for the word "fish" spoken at the surface in air, Fig. 2.15(a) and then at 300ft in a pressured helium atmosphere, Fig.2.15(b).

2.2.5 SELF-INTELLIGIBILITY IN A HIGH PRESSURE HELIUM ATMOSPHERE

The quantification of the effects of a high pressure helium atmosphere on the self-intelligibility of the diver is a somewhat complex problem. Whilst it has been estimated that overall spectral sensitivity in the aural canal is attenuated [100], it has also been shown that the diver's auditory characteristic has a tendency to emphasise by +10dB/octave those frequencies above 5kHz, and attenuate those below 5kHz [80], which in some respects compensates for the attenuation of high frequency voiced formants due to the source spectral fall-off characteristic, and may also help boost unvoiced consonant intensity. Auditory feedback, however, does not depend solely on the speech signal path through the surrounding respiratory medium, but is a rather convoluted process involving acoustic pathways through the facial tissue and bone, and the processing characteristics of the human brain [101]. Under normal circumstances in speech in air, the interaction of the signals transmitted through each of these several acoustic pathways produces recognisable signal patterns which indicate to the talker that the sound he has uttered was produced in the manner he desired [102]. Experimentation [103] has shown that disruption of the auditory feedback mechanism causes the talker to change his voice output in a manner which, presumably, renders the speech signal more intelligible to himself as he judges, but can be detrimental in terms of the effect upon intelligibility to the listener. In addition, different subjects alter their speech in an individual manner to the

same change in the auditory feedback system [104].

It has been reported that some divers appear to voluntarily adapt their voices (usually over a matter of days) in order to sound more intelligible to themselves, with varying opinions as to the success or otherwise of such manipulations with respect to the listener [71,90,94,105]. This is demonstrated in Fig. 2.16, which shows the formant frequency ratios, which are important in the perception of vowel sounds, for four different vowels uttered by the same subject. These results suggest that relative formant frequency ratios, and to some extent absolute formant frequency values, can be varied in a concious manner by the diver in a helium environment. The subjects spent several days in a deep diving chamber in a pressured helium environment, and their respective vowel formant frequencies were measured in air before the experiment, in the chamber a short time after the dive had commenced, and finally prior to leaving the chamber after several days in the helium environment.

The trends of Fig. 2.16 show that the first formant frequency F1 has ultimately tended towards its value in air, and that the relative formant frequency ratios (F1:F2, F1:F3) have also demonstrated a consistent trend towards their values in air.

2.3 ACOUSTIC PROPERTIES OF HELIUM SPEECH RELATED TO UNSCRAMBLER DEVICES

The acoustic properties of the distorted speech produced as a result of breathing in a pressured environment containing high percentages of helium gas, although not completely quantified to date, can be broadly summarised in comparison to speech in an air environment at normal pressures, as follows:

	Vowel 1			Vowel 2			Vowel 3			Vowel 4		
	F1 (Hz)	F1:F2	F1:F3	F1 (Hz)	F1:F2	F1:F3	F1 (Hz)	F1:F2	F1:F3	F1 (Hz)	F1:F2	F1:F3
Early in experiment	1000	1.9	3.25	1100	1.91	3.27	950	2.3	3.3	1450	1.93	2.97
Late in experiment	900	2.0	3.28	950	2.0	3.47	800	2.6	3.56	1100	2.09	3.36
In air	550	3.09	4.45	600	3.0	4.17	400	3.0	5.75	700	2.0	3.36

Fig. 2.16 Self-adaptation of formant frequency ratios in a high-pressure heliox environment (adapted from (94)).

(1) fundamental (pitch) frequency is relatively unchanged with respect to a normal air atmosphere [77,78,79];

(2) voiced formant centre frequencies in air below approximately 500Hz are shifted by an amount greater than that expected from the ratio R of the velocity of sound of the pressured mixture to that of air;

(3) formant frequencies in air beyond 500Hz are shifted approximately linearly by the sound velocity ratio R;

(4) the overall effect of properties (3) and (4) is to produce a non-linear voiced formant shift in high pressure helium, the nonlinearity being confined to the low frequency region of the speech spectrum[81,84];

(5) formant bandwidth may [91,92] or may not [80,86,90] be increased in a helium environment;

(6) there is enhanced attenuation of high frequencies above 5kHz in the spectrum of voiced sounds [90,93];

(7) the difference in sound pressure level between voiced and voiceless speech sounds is aggravated in a helium environment [99].

All commercially available real time helium speech unscramblers assume approximations to certain of the above properties in order to simplify system architecture and produce cost-realisable devices. These devices fall into two distinct categories of system architecture; (a) those based on signal processing in the time domain and (b) those based on signal processing in the frequency domain. Both system architectures involve roughly the same assumptions as regards the helium speech signal in order to favour linear processing strategies, viz:

(1) fundamental frequency is unaltered;

(2) the overall speech spectrum is shifted linearly upwards in frequency by the sound velocity ratio R;

(3) there is no requirement to differentiate between voiced and unvoiced

sounds in order to provide differential amplification of voiceless segments of speech.

Depending on the individual system, there may be simple prefiltering in order to provide limited amplification of select regions of the frequency spectrum.

Historically, the earliest attempts to unscramble helium speech involved the use of tape recorders where the helium speech was recorded at a fast speed, then subsequently played back at a slower speed at which the resulting speech was more intelligible [87]. A modified version of this technique, which enabled real time processing, involved the use of a continuous tape loop on which the helium speech was recorded by one fixed recording head, then subsequently read by a set of pick-up heads mounted on a capstan which was itself rotating past the tape loop [90]. However, these tape methods inherently involved bulky moving mechanisms and, more importantly, were limited in their ultimate fidelity since in addition to shifting the formant frequencies of the speech signal, the fundamental frequency was also shifted.

The direct electronic descendent of the tape loop technique involves processing directly the time-varying helium speech signal, and is the most widely used technique employed in commercial unscramblers today [93,106,107,108]. The basic time-domain technique involves segmentation of the helium speech signal either asynchronously or in synchronism with the pitch period, followed by time expansion of each speech segment by the required correction ratio: the basic assumption here, of course, is that formant bandwidths increase in a helium atmosphere proportional to the velocity of sound ratio R. Note too that the time-domain processing approach inherently requires special precautions to maintain pitch information. A detailed account and

simulation of one particular time-domain implementation [108] is provided in Chapter IV of this thesis.

2.3.1 FREQUENCY DOMAIN UNSCRAMBLING TECHNIQUES

Frequency-domain techniques involve unscrambling of the helium speech signal by frequency manipulation. In commercial unscramblers requiring real time processing, this manipulation is achieved by bandpass filtering.

In unscrambling by frequency subtraction, for example, the input helium speech is first converted upwards in frequency by a balanced modulator, and is subsequently down-converted by a balanced demodulator of different (and variable) frequency [78], thereby catering for variations in the heliox mixture. This technique can be extended by providing several bandpass filter channels at the input, each with its own modulator-demodulator, thereby allowing each frequency band to be shifted independently. In this way, it is possible to provide correction for the nonlinear formant shift criterion by a piecewise-linear approximation. The system is ultimately limited in terms of cost, physical size and power consumption by the number of modulation channels to be provided. Note that since each frequency within any one passband is shifted down by exactly the same amount, then original formant bandwidths are conserved.

Another frequency domain technique which has been implemented in real time involves unscrambling by frequency multiplication, in which the input helium speech is again filtered to provide several passbands: in this case, however, each passband is scaled down in frequency by multiplication by a geometric ratio. The basic strategy is similar to that of analysis-synthesis vocoding [18], in which the input bandpass

filters analyse the spectral power in each passband, whose value is then passed to the synthesis filter, where it is reconvolved with the excitation source. All synthesis filter outputs are then summed together to provide the resulting output speech. In the particular application of helium speech, the synthesis filter centre frequency is related to the corresponding analysis filter centre frequency by the required geometric ratio. Although it is theoretically possible to relate each analysis-synthesis filter by a slightly different ratio and so provide nonlinear frequency scaling, this has proved difficult to implement cost-effectively, therefore commercially available unscramblers employing this technique relate each analysis-synthesis filter pair by the velocity of sound ratio R as defined in equ.(2-8), permitting only linear shifting of the speech spectrum [109,110]. Additionally, in direct contrast to the frequency subtraction method, original formant bandwidths will also be compressed by the ratio R.

Frequency domain techniques, although technologically more complex, offer several advantages over their time domain counterparts in terms of improving the intelligibility of the unscrambled speech [111]: (1) by suitable bandpass filtering, the fundamental frequency of excitation can be conserved and reconvolved with the remaining spectrally shifted channel outputs, so maintaining the naturalness of the speech. This is particularly expedient since it obviates the need for any pitch extraction circuitry; also, since the whole of the pitch waveform is conserved, then there is no loss of information in the otherwise discarded trailing end of the waveform inherent in time domain processing; (2) frequency domain techniques, in theory, permit nonlinear shifting of areas of the speech spectrum through piecewise approximations; (3) depending on the particular technique employed,

original formant bandwidths are either conserved or compressed and (4) by selective manipulation of the amplitude response of each filter passband, it is possible to compensate to some extent for the high frequency losses observed in a heliox environment.

2.3.2 WAVEFORM CODING TECHNIQUES

While the analog frequency domain techniques outlined above offer important improvements over time domain techniques, further improvement yet is possible by the use of waveform coding techniques, which are still wholly based in the frequency domain but which demand digital signal processing.

Waveform coding involves the transformation of the time-sampled waveform to some other set of parameters which fully describe the signal, such as sampled-data poles and zeros through the z-transformation. The derived transform variables are then manipulated according to some predetermined algorithm such that applying the inverse transformation produces a resulting time waveform whose spectral content has been corrected for the helium speech distortion. Virtually any type of spectral frequency and amplitude correction is possible with these techniques, depending only upon the algorithm or set of algorithms used for correction in the transform domain and the signal bandwidth as defined by the sampling rate. To date, processors based on homomorphic deconvolution [112] and linear predictive coding [75] have been applied to the helium speech problem, with reportedly excellent results in terms of the intelligibility of the unscrambled speech. Historically, however, the principal disadvantage of waveform coding systems has been that they are difficult to implement in

real time due to conflictions between processing bandwidth requirements and state-of-the-art digital processing. Recent advances in array processing and microelectronics have, however, permitted the implementation in real time of an unscrambler system based on the short-time Fourier transform [92], and this technique is examined in depth in Chapter IV of this thesis.

In summary, whilst time domain unscrambling techniques are readily and cheaply implemented in real time, their performance in terms of intelligibility remains non-optimal by the very nature in which they correct helium speech: although improvements in the technology are possible, such as miniaturisation [108], improvements in the helium speech correction characteristic are severly restricted if not impossible to achieve. Frequency domain unscramblers based on bandpass filtering provide great improvements in intelligibility over the time domain techniques, since no explicit pitch detection is required and signal integrity is conserved. Although these devices theoretically permit nonlinear frequency and amplitude correction, for reasons of device complexity and cost, this feature has not been implemented to date on any commercially available unscrambler. Also, being based on analog filtering techniques, these devices are inflexible once assembled, i.e. changing the helium speech correction characteristic as advances are made in the understanding of the helium speech phenomenon is likely to incur physical component changes in the device. Waveform coding techniques, on the other hand, offer the potential of providing high performance unscrambler systems through digital signal processing. Nonlinear spectral correction is relatively facile, and moreover the device is flexible to future changes in the helium

speech correction characteristic, requiring changes only in the digitally-implemented correction algorithm.

A comparison and detailed description of unscrambler devices in each of the three categories of unscrambler techniques described in sections 2.3.1-2 is contained in Appendix A.

CHAPTER III

AN ACOUSTIC ANALYSIS OF VOICED HELIUM SPEECH AND AMBIENT NOISE.

Notwithstanding the technological complexity of the real-time electronic systems which are at present commercially available for helium speech unscrambling, there has still been cause to vilify the quality of the unscrambled helium speech they produce [90,113,114,115]. There are two corollaries into which the root causes for the poor performance of these systems may be divided: (1) there may be assumptions pertaining to the acoustic events in helium speech which these systems make which are false, or indeed the systems may be deficient in their provisions for processing certain acoustic attributes of the signal which have an important bearing upon intelligibility; and (2) the unscrambling algorithms which are implemented in these devices may affect the resultant unscrambled speech in some manner which is antagonistic to good intelligibility.

In this chapter, proposition (1) is explored, with proposition (2) being investigated in Chapter IV. In order both to confirm and elucidate certain of the classical assumptions as to the effect of a pressured heliox atmosphere upon the acoustic events of the speech signal, as detailed in Chp.II, section 2.2, new recordings of helium speech have been procured, and the first part of this chapter describes the choice of speech material used in the recordings and the experimental scenario. The main body of Chapter III, however, is concerned with the acoustic analysis of the recorded speech. Firstly, in order to vindicate the assumption that pitch is invariant from air to a heliox environment,

there is an analysis of the pitch of voiced speech using homomorphic pitch determination [116]. This is followed by an extensive spectral analysis of the spectral correlates of the perceptual features directly affecting the intelligibility of voiced helium speech, with specific reference to formant frequency, amplitude and bandwidth. The spectral analysis itself employs autoregressive spectral modelling [117,118,119], and this technique is described in detail. The chapter concludes with a discussion of the results coupled with an acoustic analysis of the major noise source present in the recordings and its likely effects upon the foregoing speech analyses.

3.1 SPEECH MATERIAL TO PROVIDE AN ANALYSIS DATABASE AND THE EXPERIMENTAL SCENARIO

3.1.1 CHOICE OF RECORDING MATERIAL

The bases on which the material to be recorded was chosen were (a) that the material should not be linguistically difficult to pronounce and should lend itself to as natural a pronunciation as possible; (b) it should provide a high confidence in terms of constancy of performance of the subject from reading to reading and (c) if possible, it should be structured acoustically such that, temporally, the various acoustic events required for analysis can be subsequently segmented and extracted for analysis with relative ease. Since there was no guarantee that the subject would have had prior experience of any phonetically-oriented experimentation beforehand, the material ultimately chosen consisted of lists of carrier phrases of the form (1) "I say fa sometimes" and (2) "I say heed sometimes" [120]. The complete lists used for the

recordings are displayed in Fig. 3.1. In each case, the target word, which in turn contains the target speech sound for subsequent analysis, is between the words "say" and "sometimes" in the carrier phrase, such as the fricative sound /ʃ/ in the target word "ʃa" in example (1) above. The strategy of using carrier phrases of this type, whilst helping with naturalness and aiding subsequent segmentation and isolation of the target speech sounds, also gives some control over regulation of the overall speaking tempo. The lists can basically be resolved into two groups: those containing consonant targets (lists 1 and 3) and those containing vowel targets (lists 2 and 4). List 1 contains the known plosive and stop (voiced and unvoiced) consonants to be found in English, such as the unvoiced /t/ in "ta", and its voiced counterpart /d/ in "da". Note that the phonetic environment of each consonant is the same, each being preceded by the vowel "ay" in "say" and followed by the vowel /a/, both of which are of greater intensity than the corresponding plosive, thereby aiding with the identification and extraction of the required portion of the speech signal. An exactly similar device is used in list 3, which contains the known fricatives such as /s/ and /z/ and nasals (/m/,/n/) together with the liquid consonants such as /l/ and /r/. List 2 contains the voiced vowel diphthongs, that is, vowels which involve formant transitions during their production in order to give the impression of the desired phoneme, such as the "a" to "ee" transition to produce the "ɪ" in "hɪde". Note here that the target sounds are contained between the unvoiced aspirate /h/ and stop consonant /d/, which again helps with identification and extraction. List 4 contains the known English monophthongs, or steady-state vowels. The bulk of the analysis of speech in this chapter refers directly to the speech sounds in List 4.

Carrier Phrase: "I say item sometimes."

List 1	List 2	List 3	List 4
item	item	item	item
(1) *pa*	(1) h*ay*ed	(1) *z*ah	(1) h*ee*d
(2) *la*	(2) h*i*de	(2) *y*ah	(2) h*oo*d
(3) *ta*	(3) h*ow*ed	(3) *w*ah	(3) h*i*d
(4) *da*	(4) h*oe*d	(4) *l*a	(4) h*ea*d
(5) *ka*	(5) h*oy*ed	(5) *n*ah	(5) h*a*d
(6) *ga*	(6) h*ee*red	(6) *s*ah	(6) h*ar*d
(7) *j*ar	(7) h*ea*rd	(7) *m*a	(7) h*u*d
(8) *ch*a	(8) h*oa*rd	(8) *th*ank	(8) h*o*d
		(9) *n*ah	(9) h*aw*ed
		(10) *sh*ah	(10) who'd
		(11) *l*a	(11) h*er*d
		(12) *v*ah	
		(13) *th*an	

List 5

Passage:

"When the sunlight strikes raindrops in the air, they act like a prism, and form a rainbow. The rainbow is a division of white light into many beautiful colours. These take the shape of a long, round arch with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no-one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow."

Fig. 3.1 Material used in recordings of helium speech.

Finally, a passage, termed "The Rainbow Passage" [121], concluded each recording. This passage is not used directly in the analysis of helium speech presented here, but rather is used as a means of providing a subjective basis on which to judge the performance of simulated unscrambler systems as described in subsequent chapters of this thesis.

3.1.2 EXPERIMENTAL SCENARIO

One volunteer diver only was used during the recordings. The subject was male, Caucasian, age 30, height approximately 5ft 11ins, slim build, and indicated that his accent is typical of the Rotorua region, New Zealand, where he spent his childhood. The recordings were made in a diving simulation chamber, situated on the surface on dry land, during the course of a welding training exercise, with all lists being spoken only once at, nominally, 100ft depth intervals during the compression phase of the dive, down to a maximum depth of 500ft. The subject was together with four other divers for the purpose of the training exercise, and all 5 had already experienced a hyperbaric heliox atmosphere several times before. The subject, who had no phonetic training whatsoever, was instructed to hold the recording microphone about 6ins away from his mouth and to read the lists in the order they came in his own time. All readings, including the reading in air at surface pressures, were made within the chamber, with the subject standing in a free atmosphere, unencumbered by personal breathing apparatus. A profile of the dive during the recording phase is shown in Fig. 3.2, with the corresponding table of environmental parameters in Fig. 3.3. In order to minimise the effect on recordings from microphone characteristics, the microphone chosen was an electrostatic

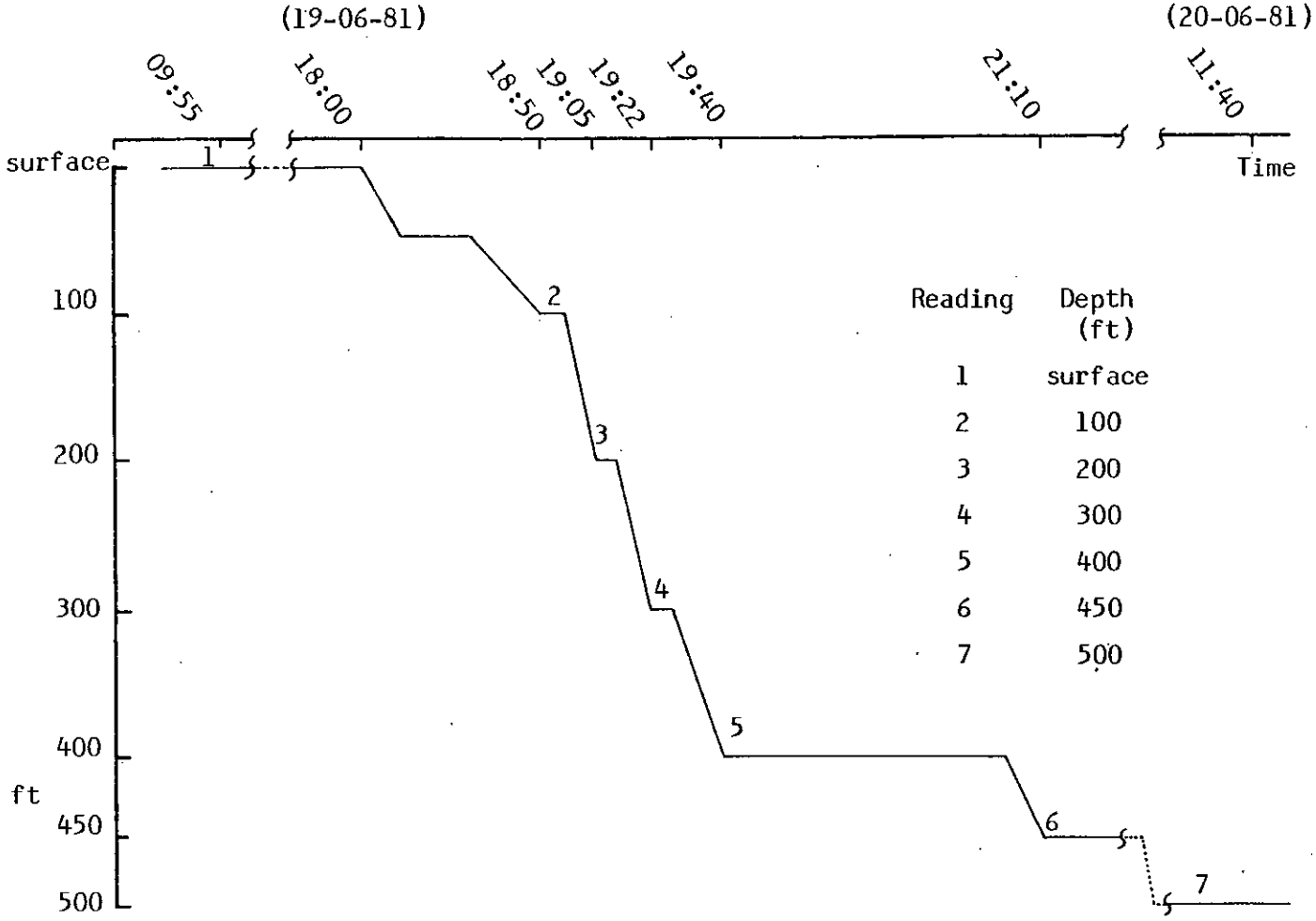


Fig. 3.2 Profile of dive during recording of helium speech.

Depth (ft)	Temp. (°C)	Press. (Bar)	He %	O_2 %	γ_{mix}	$\rho^{(1)}_{mix}$ (kg/m^3)	$\frac{C_{mix}}{C_{air}} = R$
0	23.3	1.0	Air	Air	1.4	1.293	-
100	23.9	3.99	91.6	8.4	1.638	0.2835	2.31
200	25.6	6.98	95.2	4.8	1.6475	0.2385	2.53
300	26.7	9.97	96.7	3.3	1.6514	0.2198	2.634
400	28.9	12.96	97.5	2.5	1.6535	0.2098	2.698
450	28.3	14.45	97.55	2.45	1.6536	0.2091	2.7023
500	28.9	16.3	97.6	2.4	1.6538	0.2085	2.7065
-	-	-	100	-	1.66	0.1785	2.9307
-	-	-	-	100	1.4	1.429	-

Fig. 3.3 Table of environmental parameters during recording of helium speech.

Note (1): All densities given in this column are referred to S.T.P.. For gas density at depth, multiply by associated absolute pressure.

condenser microphone, Bruel and Kjaer (B&K) 4133, with associated preamplifier. This microphone has a guaranteed flat pressure-frequency response to within +1dB in air from 20Hz to 18kHz, with the 3dB bandwidth being from 10Hz to 20kHz. From information supplied by the manufacturer, the overall sensitivity of the microphone is attenuated by a pressured heliox atmosphere such that the frequency response effects basically a level shift. Microphone, preamplifier and self-contained power supply were all located in the diving chamber. The recorder used was a Racal Store 4D using F.M. recording techniques and tape speeds of 30i.p.s.(air) and 60i.p.s.(helium) giving bandwidths of D.C. to 10kHz and D.C. to 20kHz respectively. A 'Krohn-Hite' 3rd order bandpass filter (low cut-off 15Hz) was placed between the preamplifier output and the recorder to block-off a 200v D.C. polarising voltage from the B&K preamplifier. Thus, the effective recording bandwidth both for air and heliox was of the order of 20Hz to 10kHz. Note that both recorder and filter were located in air external to the chamber. For the purposes of the analyses which follow, which were all carried out on digital computers, the speech data was digitised, inclusive of necessary anti-aliasing requirements, to 12-bit resolution on a Digital PDP11/40 computer.

3.2 ANALYSIS OF FUNDAMENTAL FREQUENCY

3.2.1 ANALYSIS PROCEDURE

For the ensuing pitch analysis, the eleven monophthong vowels of list 4 were considered. An example of the characteristic intonation

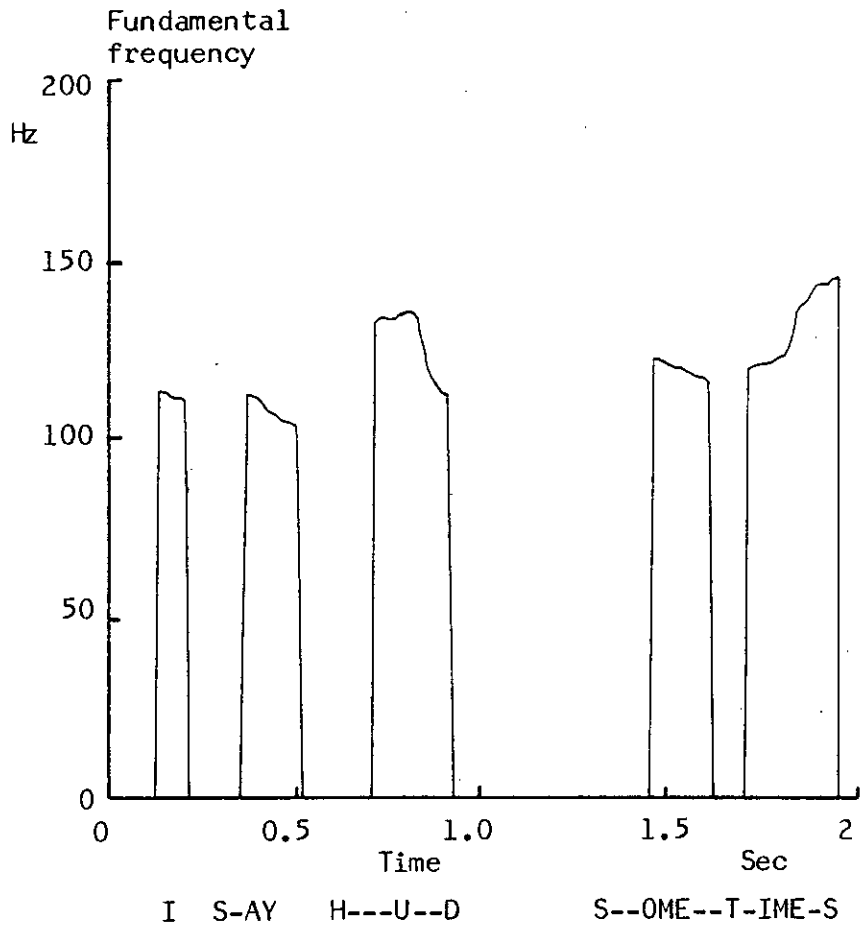


Fig. 3.4 Typical pitch intonation contour adopted by subject during reading of list 4 (see Fig. 3.1).

contour adopted by the subject during the readings is shown in Fig. 3.4. The basis for the calculation of this contour was a pitch extraction algorithm based on time-domain processing involving several pitch determinations, computed in parallel, on the input speech waveform [122]. The analysis window for each fundamental frequency estimate was 25mS, with a 15mS overlap between successive estimates. Finally, a 5-point median smoothing operator followed by a 3-point Hanning window [123] was applied to the resulting pitch estimates to produce Fig. 3.4.

The pitch contour of Fig. 3.4 verifies that the subject is treating

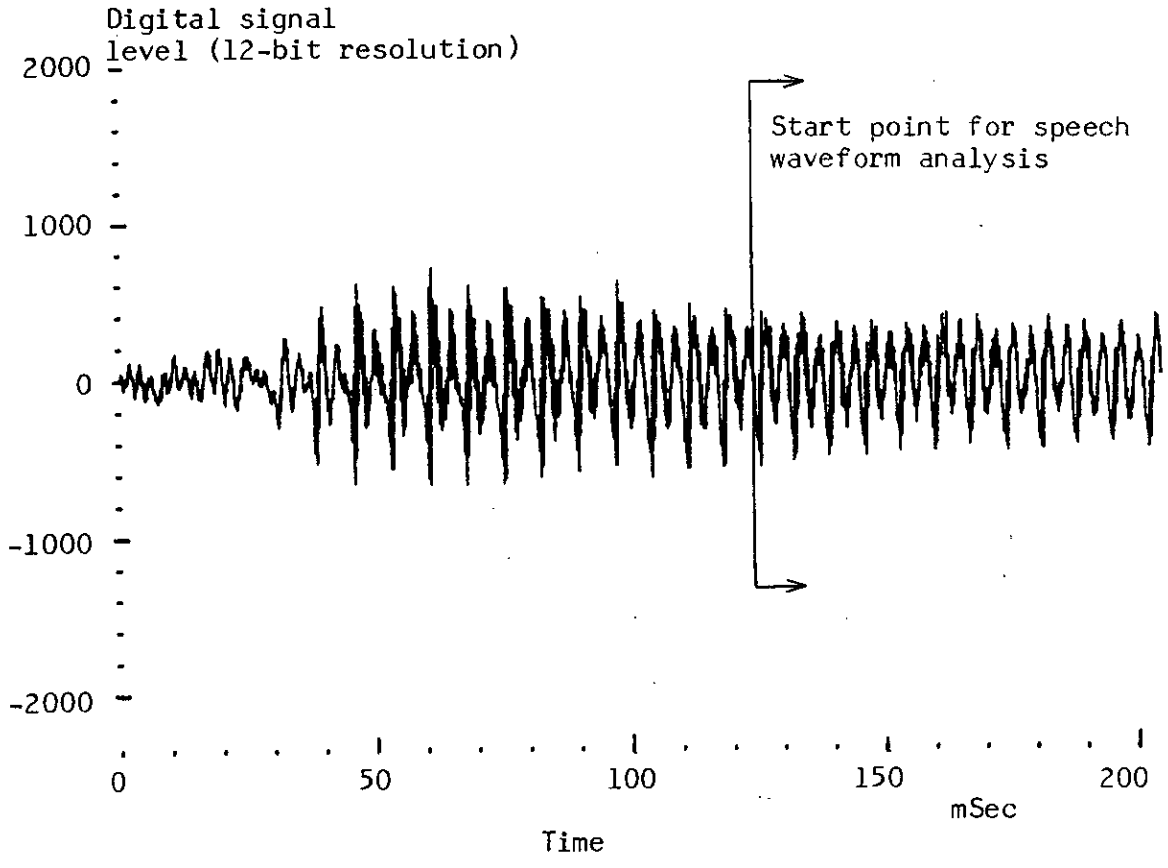


Fig. 3.5 Typical choice of start point on speech waveform for acoustic analysis.

each list of phrases with a 'listing' intonation; that is, the subject's pitch increases towards the end of each phrase except for the very last phrase of each list, which is pronounced with falling intonation. Recording A2(a) demonstrates list 4 as spoken by the subject in air, and recording A2(b) is the same list spoken at 100ft depth.

The starting point on the digitised speech waveform for the fundamental frequency analysis was determined by hand, and was taken to be the point on the waveform after which it appeared that initial intensity transients had subsided, see Fig. 3.5. Estimates of voiced pitch frequency were made using a modified cepstrum analysis technique [124,125]. The sequence

of calculations for the cepstral pitch determination process is shown in Fig. 3.6. From the point on the waveform from which the analysis was to commence, a 25mS frame of the signal was stored in a buffer. For this frame of data, the fundamental frequency was determined, and then the next frame of 25mS was taken from a point on the waveform approximately 3.2mS later in time. Thus, each successive frame overlapped the last by 22.8mS. In all, 9 frames were processed in this way, and finally, the fundamental frequency for each test vowel was determined from the average of all 9 estimates. In the cepstral domain, cepstral tailoring was applied through the use of a trapezoidal window with centre quefrency 1/125Hz, minimum quefrency 1/75Hz, maximum quefrency 1/400Hz, and tilt factor = 70%. Verification as to the voiced/unvoiced excitation status of the speech segment under analysis was made using estimators based on (a) information on the time-signal zero crossing rate, (b) the residual error power from autoregressive modelling and (c) the first reflection (PARCOR) coefficient from the recursive filter analogue of the vocal tract [126,127]. When the excitation was judged to be voiced, pitch determination itself was done by a quefrency peak selection operation.

In order to verify the estimates from the cepstral pitch determination routine, several of the results were checked at random by visual measurement of fundamental frequency from the stored speech waveform, and it was found that the hand-picked pitch estimates differed at most by approximately 0.5% from the cepstrally-picked estimate.

The results of this analysis are displayed in Fig. 3.7, which shows fundamental frequency for the 11-element vowel space of list 4, Fig. 3.1, as a function of depth. The dotted line joins the mean values of the representative fundamental frequency from depth to depth, and the solid

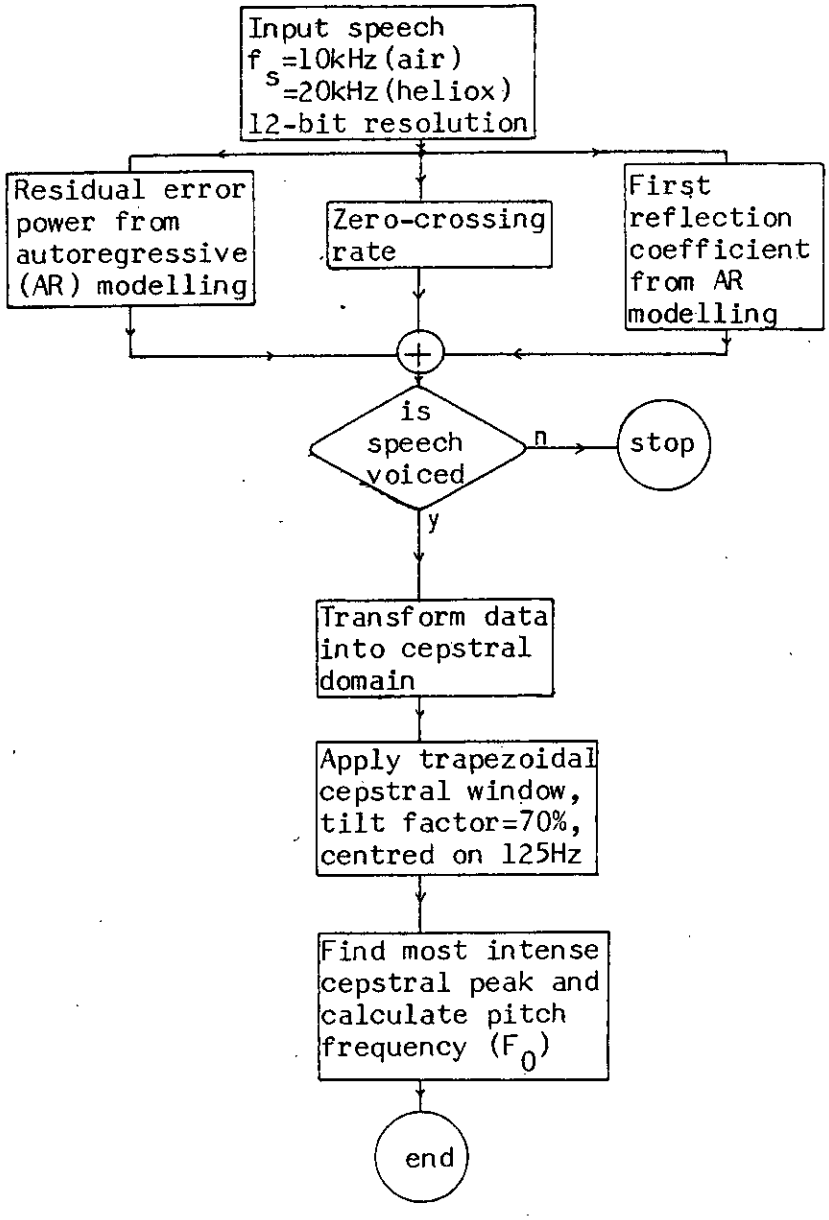


Fig. 3.6 Fundamental frequency determination by cepstral analysis.

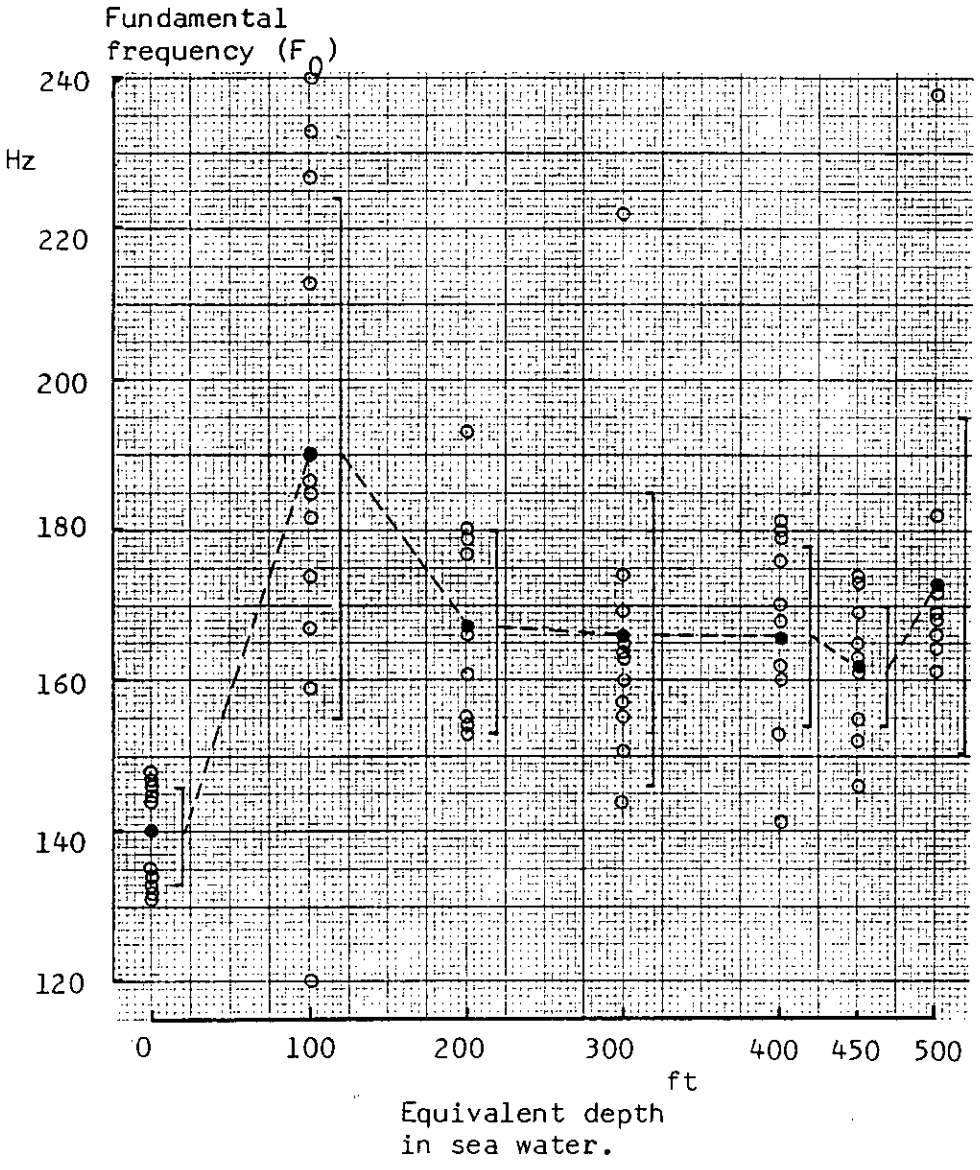


Fig. 3.7 Fundamental frequency v. depth. ● = mean Fo over vowel space; - - - trace of mean value from depth to depth;] standard deviation about mean.

vertical bars about the mean indicate the standard deviation for the pitch at each depth.

3.2.2 DISCUSSION OF RESULTS

The trends of Fig. 3.7 strongly suggest that there is no correspondence between fundamental frequency and the increasing helium content and pressure of the respiratory environment. This is borne out by the very sharp increase in fundamental frequency at 100ft and its immediate fall then stabilisation with increasing depth. This observation is in fact in general agreement with early results [77,88], as indeed might be expected from what is in the main a muscularly controlled event. What remains to be explained, however, is the very sharp increase in pitch at 100ft and indeed the fundamental frequencies at lower depths which are, in general, some 30 or so Hertz higher than the corresponding value for the recording made in air at surface pressure.

Dealing first with the sharp rise at 100ft; several indicators point to this being due to psychological stress. The start of the dive had already been delayed for several hours due to technical problems, therefore tension was already high at this time. In addition, this was the first time that the subject had enunciated the word lists in the presence of the other divers, and indeed the first time all five had been together for several days, and hence the subject was exposed to a certain amount of ribaldry on their part. It was also the first time for several months that the subject had been in a heliox atmosphere. All these factors are liable to be contributory to a state of nervous tension. This explanation is made more plausible by the reduction in mean pitch some 20 minutes later at a depth of 200ft, where the subject

was perhaps becoming more accustomed to the heliox atmosphere, and to his companions. It is also strengthened by the large value of standard deviation of pitch frequency at 100ft compared with other depths, signalling a more erratic performance in terms of fundamental frequency, perhaps due to physiological manifestations of psychological stress, such as an increase in vocal cord muscle tension [128].

Below 300ft, the measured pitch remains within constant limits, although still higher than might be expected. There are again several possible explanations for this. The diver's own hearing, affected by the heliox atmosphere, will cause frequency attenuation within the aural canal [80], in which case his vocal intensity is likely to increase in an effort to auto-compensate for the drop in the overall input of acoustic energy to the auditory system. Additionally, the subject may have been speaking louder in the apparent belief that those outside the chamber could not hear him adequately. This effect is evidenced by most people on the telephone, in that if the subscriber becomes faint due to a technical fault, then although he may be hearing the caller perfectly well, the tendency is for the caller to raise his voice. Similarly, if the effect of the pressured atmosphere is such that it promotes enhanced damping of the communications loudspeaker within the compression chamber, then those outside the chamber will sound progressively fainter and fainter, hence causing the subject to unnecessarily raise his voice. This would normally lead to an increase in pitch frequency of perhaps 10-20Hz. It was also noted that the subject gave linguistic emphasis to the target word from time to time. This too is likely to increase fundamental frequency.

3.3 AUTOREGRESSIVE (AR) SPECTRAL ANALYSIS

The analysis of formant centre frequency, amplitude and bandwidth which follow are all based on the parametric spectral analysis technique of autoregressive (AR) spectral modelling [129]. The discussion and theorems presented in this section have also an important relevance to the unscrambling technique based on linear prediction which is developed in Chapter V.

3.3.1 BASIC PRINCIPLES

The basic assumption in AR spectral analysis is that the speech signal is produced by a pulse-excited all-pole filter [130] whose transfer function is similar to that of equ.(2-1), and the recursive digital equivalent of the speech mechanism in this respect is shown in Fig. 3.8. Note that z^{-1} represents unit time delay and a_n the coefficient associated with each delay stage, with v(nT) the vocal tract excitation function and s(nT) the resulting speech signal.

In parametric spectral analysis in general, the aim is to represent the input signal by a limited set of derived parameters which describe the signal spectrum in some optimum manner [131,132]. In the case of AR spectral modelling, the required parameters are the z-plane poles of the digital signal s(nT) which, from a consideration of the duality between the discrete Fourier and z-transformations, will under certain conditions uniquely specify the optimum estimation of the signal spectrum [133]. Since the optimum spectral fit in this case relates to the poles of the speech signal, then those features of the signal spectrum which are directly due to the influence of signal poles, that

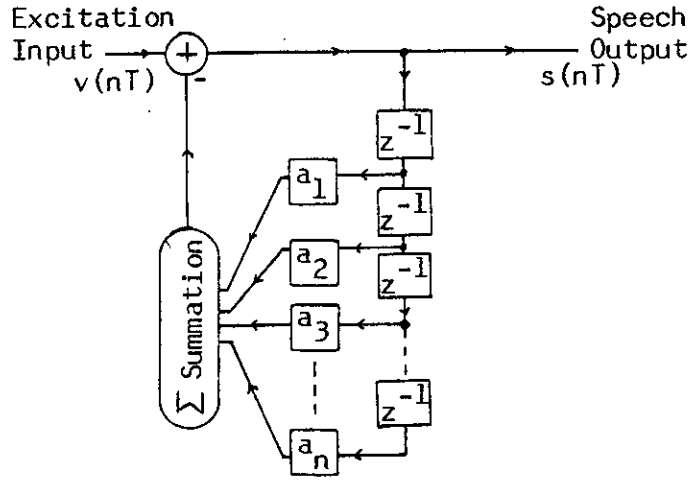


Fig. 3.8 Recursive digital filter model of the speech mechanism.

is spectral peaks and areas of resonance, will be modelled best, whereas features due to the influence of signal zeros, that is troughs or antiresonances, will not be faithfully represented [12]. However, although the characteristic transfer function of certain speech sounds, such as nasals, do indeed contain zeros of transmission, the important aspects of speech perception still relate to formant information, that is, the relative locations, amplitudes and bandwidths of spectral peaks [10], so vindicating the use of AR spectral analysis.

Since the aim of the appropriate filter which will comprise the AR analysis system is to extract as much information as possible from the input signal in respect of its characteristic poles and hence its intrinsic spectral features, then the output signal resulting from processing by this analysis filter must, theoretically, be devoid of all spectral detail, and so will consist of either Gaussian noise or a series of Dirac impulses, since the spectral characteristic of both of

these series is flat (white) with no discernable features over the observable frequency range of interest.

In the case of human speech, note that the vocal tract cannot be considered to be excited by a perfect Dirac impulse source, but rather by a complex glottal pulse source [14] of the form of Fig. 2.3(a). However, since the AR analysis filter deliberately seeks to produce an output signal which is a perfect Dirac impulse, then the analysis filter parameters must model in some way both the glottal source and vocal tract characteristic, but as one convolved entity: that is, there is no deconvolution of source and transfer function spectral features as a result of applying the analysis filter. To summarise, the aim of the analysis filter can be considered as being to produce at its output a residual signal e(nT) which is spectrally whitened compared to the input signal s(nT), in which case the analysis filter impulse response h(nT) must reflect information regarding the spectrum of the input signal: in fact, the analysis filter coefficients must correspond in some way to the inverse spectrum of the input speech signal.

Let the corresponding discrete Fourier transforms or spectra of input s(nT) and residual e(nT) be S(mF) and E(mF) respectively, and let the frequency response of h(nT) be H(mF). From convolution theory:

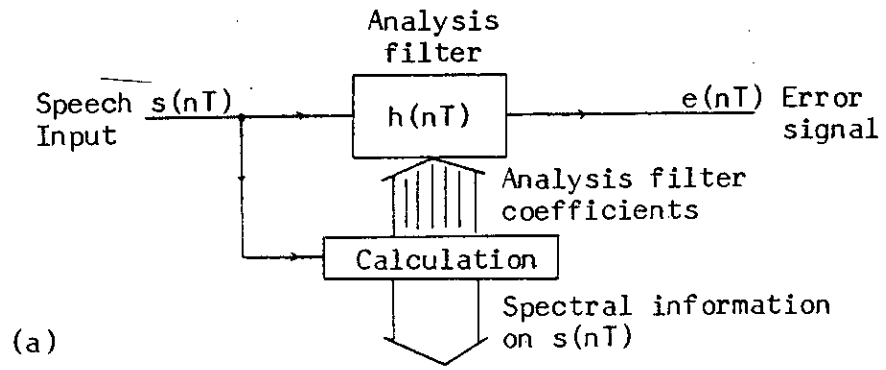
$$s(nT) * h(nT) = e(nT) \rightleftharpoons S(mF) \times H(mF) = E(mF) \qquad (3\text{-}1)$$

where * denotes convolution. From the right-hand side of equ.(3-1),

$$S(mF) = E(mF)/H(mF) \qquad (3\text{-}2)$$

That is, if the analysis filter frequency response is obtained together with the spectrum of the residual signal, then the input spectrum can be derived. Equation (3-2) can be considered as the kernel of autoregressive spectral analysis.

In the case of speech as the input signal for analysis, the spectral



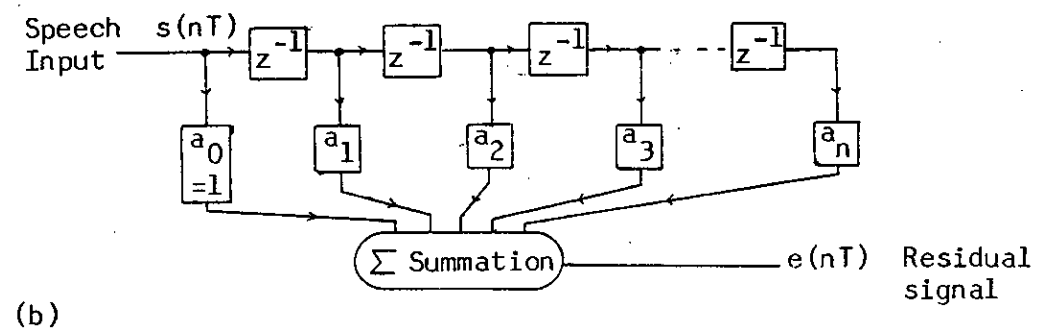


Fig. 3.9 (a)Analysis filter to provide spectral information based on stochastic characteristics of the input signal s(nT).
(b)Block structure of spectral analysis (prediction error) filter based on linear prediction.

characteristics of the signal may be changing approximately every 10-30mS [12], and so the analysis filter parameters must also be changed periodically to provide adequate spectral whitening of the residual. Since there is virtually no a priori information regarding the speech signal from instant to instant, that is, there is no deterministic relationship dependent on time only which will satisfactorily describe the input signal, then the analysis filter coefficients must be calculated using stochastic characteristics of the input signal itself [134], see Fig. 3.9(a).

3.3.2 CALCULATION OF ANALYSIS FILTER COEFFICIENTS

The basis for the calculation of the analysis filter coefficients rests on linear regression, with particular reference to linear prediction in this case [119]. Specifically, the analysis filter attempts to model the present input signal sample s_n based on a weighted sum of input points which occurred earlier in time, that is :

$$\hat{s}_n = -\sum_{k=1}^{p} a_k \, s_{n-k} \qquad (3\text{-}3)$$

where $\hat{s}_n$ is the predicted estimate of the current speech sample s_n based on the p values of the analysis filter coefficients a_k and the p most immediate past samples of the input signal s(nT). The constraint upon the system is defined as being that the filter output residual or error signal,

$$e(nT) = s(nT) - \hat{s}(nT) \qquad (3\text{-}4)$$

should be such that its power content, the error power, E, is minimised [119], that is, from equs.(3-3) and (3-4):

$$E = \sum_n e_n^2 = \sum_n (s_n + \sum_{k=1}^{p} a_k s_{n-k})^2 \qquad (3\text{-}5)$$

then for minimum error power, for any single coefficient a_i,

$$\frac{dE}{da_i} = 2 \sum_n (s_n + \sum_{k=1}^{p} a_k s_{n-k}) \cdot s_{n-i} \qquad 1 \leqslant i \leqslant p \qquad (3\text{-}6)$$

$$= 0 \text{ for minimum error power.}$$

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-i} = -\sum_n s_n s_{n-i} \qquad 1 \leqslant i \leqslant p \qquad (3\text{-}7)$$

Explicit expansion of equ.(3-5) followed by the substitution of equ.(3-7) yields the relationship:

$$\sum_n s_n^2 + \sum_{k=1}^{p} a_k \sum_n s_{n-k} s_n = E \qquad (3\text{-}8a)$$

and, by extending k to cover the range $0 \leqslant k \leqslant p$,

$$\sum_{k=0}^{p} a_k \sum_n s_{n-k} s_n = E, \qquad a_o = 1 \qquad (3\text{-}8b)$$

From equs.(3-3), (3-4) and (3-8b), the block structure of the spectral analysis or prediction error filter is defined by Fig. 3.9(b), and moreover is the inverse of the autoregressive filter model of the speech mechanism shown in Fig. 3.8. However, it is equ.(3-7) which provides the key of the calculation of the coefficients a_k. Absorbing the -ve sign into the coefficient series a_k and realising that the ith autocorrelation lag is given by:

$$r(i) = \sum_{n} s_n s_{n-i} \tag{3-9}$$

then equ.(3-7) can be written as:

$$\sum_{k=1}^{p} a_k r(|i-k|) = r(i) \qquad 1 \leqslant i \leqslant p \tag{3-10}$$

which yields a set of p equations which may be written in matrix form as:

$$\begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(p-1) \\ r(1) & r(0) & r(1) & \cdots & r(p-2) \\ r(2) & r(1) & r(0) & \cdots & r(p-3) \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ r(p-2) & \cdot & \cdot & & r(1) \\ r(p-1) & r(p-2) & \cdots & & r(0) \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ a_{p-1} \\ a_p \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \cdot \\ \cdot \\ r(p-1) \\ r(p) \end{bmatrix} \tag{3-11}$$

The set of p equations in equ.(3-11) are referred to as the Yule-Walker normal equations [135,136], and these can be augmented, by considering equ.(3-8b), to incorporate the error power E:

$$\begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(p) \\ r(1) & r(0) & r(1) & \cdots & r(p-1) \\ \cdot & \cdot & \cdot & & \cdot \\ r(p-1) & \cdot & \cdot & \cdots & r(1) \\ r(p) & r(p-1) & \cdot & \cdots & r(0) \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ \cdot \\ a_{p-1} \\ a_p \end{bmatrix} = \begin{bmatrix} E \\ 0 \\ \cdot \\ 0 \\ 0 \end{bmatrix} \tag{3-12}$$

The autocorrelation matrix on the left-hand side of equs.(3-11) and (3-12) is of the Toeplitz form, i.e. the elements along each diagonal are equal, and is positive definite. The Toeplitz form lends itself

readily to solution, in this case for the a_k, by a recursive formulation known as Levinson recursion [137], and does not entail explicit matrix inversion. These recursive methods are much faster than matrix inversions since their computing time is proportional to p^2 instead of the p^3 for traditional solution methods, where p is the number of analysis coefficients to be computed. The Levinson algorithm constructs the analysis filter of length p recursively, adding one new ℓth filter stage at a time, where ℓ defines the computation pass number and also the length of the filter at the ℓth pass, $\ell \leqslant p$. Thus, each ℓth pass in the total set of p passes solves for the ℓ updated values of the $a_{\ell k}$ in equ.(3-11) together with the error power of the residual signal of the ℓth filter stage, E_{ℓ}, in equ.(3-12), $k \leqslant \ell$, from which the coefficient sets (a_{11},E_1), (a_{21},a_{22},E_2) etc. are generated. The algorithm is initialised by setting:

$$a_{11} = -r(1)/r(0) \qquad (3\text{-}13a)$$

$$E_1 = (1-|a_{11}|^2).r(0) \qquad (3\text{-}13b)$$

with the following recursion relations for $2 \leqslant \ell \leqslant p$:

$$a_{\ell\ell} = -\,r(\ell) + \sum_{j=1}^{\ell-1} a_{\ell-1,j}\, r(\ell-j) \,/E_{\ell-1} \qquad (3\text{-}14a)$$

$$a_{\ell i} = a_{\ell-1,i} + a_{\ell\ell} a^{*}_{\ell-1,\ell-i} \qquad (3\text{-}14b)$$

and

$$E_{\ell} = (1-|a_{\ell\ell}|^2).E_{\ell-1} \qquad (3\text{-}14c)$$

Since successively higher order models are being estimated with this algorithm, examination of the error power E_{ℓ} at each stage can assist in determining a suitable model order p for the data being analysed, since for a perfect AR process, E_{ℓ} will first reach its minimum value at the correct model order.

The resulting coefficients $(a_0 \rightarrow a_p)$ are termed the analysis or

prediction error filter coefficients, and the coefficients (a_{11},a_{22}.. .. $a_{\ell\ell}$ .. a_{pp}) resulting at each ℓth recursion are termed the reflection coefficients, written as (k_1,k_2, .. k_p), and are useful in monitoring that a stable model of the signal is being produced [125], in that a necessary and sufficient condition for stability is that $|k_i| < 1$ for $i = 1,2,...p$.

3.3.3 DERIVATION OF THE SPECTRAL ESTIMATE

From equ.(3-2) relating the spectrum or power spectral density (PSD) of the input signal s(nT) to the power spectral densities of the residual error signal e(nT) and impulse response h(nT) of the analysis filter, then if the input signal s(nT) is the result of a perfect AR process, then the value of the spectral power of each component frequency present in the PSD of the residual, e(nT), must be equal to some constant K, since its characteristic PSD is, ideally, flat. Moreover, from the Weiner-Kinchine theorem [138], which relates the autocorrelation function $R(\tau)$ of a signal to its power spectral density, P(f) through the relationship:

$$R(\tau) = \int_{-\infty}^{+\infty} P(f)e(j2\pi f\tau)df \qquad (3\text{-}15a)$$

or, in discrete form,

$$r(n) = \sum_{m=0}^{M} P(m)e(j2\pi mn/M) \qquad (3\text{-}15b)$$

then, for the 0th autocorrelation lag at time n=0;

$$r(0) = \sum_{m=0}^{M} P(m) \qquad (3\text{-}16)$$

i.e. the value of r(0) is the sum of the spectral power of each component frequency. With particular reference to the residual signal e(nT), the zeroth autocorrelation lag, from equ.(3-9), is related to the residual

error power E through

$$r(0) = \sum_{n} e_n^2 = \sum_{m} P(m) = E = K, \text{ a constant} \qquad (3\text{-}17)$$

Thus, the PSD of the input signal, S(f), from equ.(3-2), is:

$$S(f) = E/H(f) \qquad (3\text{-}18)$$

where H(f) is the PSD of the frequency response of the analysis filter. Furthermore, since the analysis filter structure shown in Fig. 3.9(b) is non-recursive, then its impulse response corresponds exactly to the predictor error filter coefficient series $(a_0, a_1, a_2 \ldots a_p)$. Thus, the PSD can be obtained by taking directly the Fourier transform of this series in order to produce H(f). The PSD of the input signal, S(f), is then found by simply inverting the values of each power spectrum component of H(f), and multiplying it by the error power E, under the implicit assumption that the corresponding PSD of the residual is spectrally flat. In particular, note that since the AR spectrum is obtained from (effectively) a time series resulting from the one-shot impulse response of the analysis system, then the resulting spectrum is devoid of any lines which would otherwise be present in the Fourier transform of the (periodic) input signal itself. This is perhaps the most important aspect of AR spectral analysis, since it allows a much clearer picture of formant frequencies and spectral detail which would otherwise be obscured by the spectral lines due to time periodicity in the conventional Fourier transform of the input signal.

Whilst the choice of the analysis filter length, p, should be such that the error power E is minimised, note too that the z-transformation of a filter impulse response which is of the form $a_0 + a_1 z^{-1} + a_2 z^{-2} + \ldots a_p z^{-p}$ will contain p roots in the z-plane. Thus, the choice of filter length corresponds directly to the amount of poles which are to be modelled into the required spectral estimate.

3.4 SPECTRAL ANALYSIS OF VOICED SPEECH

3.4.1 ANALYSIS PARAMETERS AND PROCEDURE

The determination of the order of the autoregressive process to be modelled, or equivalently, the length of the prediction error filter, is an important point in the estimation of the AR spectrum. This length determines both the smoothness of the spectral estimate and, together with the autocorrelation sequence length, its degree of approximation to the actual signal spectrum. There are various conditions which have been proposed in order to aid assessment of the correct filter order and autocorrelation sequence length depending on the statistical properties of the data [139,140,141,142]. The approach here, however, has been to specify the required parameters from a consideration of the properties of the speech acoustic waveform.

From section 2.1.2, the first 3 formants of speech in air occur below 3.5kHz and from Fig. 3.3, which shows that for a 100% helium environment the expected linear frequency shift is approximately x3, then the first 3 formants of speech recorded in the heliox environments of this experiment should be located below 10kHz, corresponding to the bandwidth of the recording system, in fact. The same system bandwidth, however, will preclude any analysis of unvoiced speech sounds such as fricatives, since significant acoustic events occur in their spectra up to 6-8kHz in normal air speech [43], which is expected to translate to approximately 15kHz and beyond in heliox and is therefore far in excess of the available recording bandwidth. Thus, the discussions which follow are confined to exploring details of voiced speech, with particular reference to the first 3 formants only.

Assuming each formant to be the result of a complex pole pair in the vocal tract transfer function of equ.(2-1), then a filter order of at least p=6 is required. Experimentation has shown, however, that a model order of p=10 provides a more adequate spectral fit [143], allowing for anomalous spectral detail which occurs from time to time in speech. Note at this point that helium speech is characterised by the same number of poles as normal air speech, since poles are theoretically only shifted upwards in frequency with no extra poles introduced.

In the present experiment, the speech data for analysis is also contaminated to some extent by a noise source, located within the diving chamber, relating to equipment for cleansing excess carbon dioxide from the respiratory mixture. Since the noise characteristic may affect the spectral estimate by introducing extraneous spectral peaks, then the model order chosen for the analysis was increased to p=14 for both air speech, with sample frequency fs=10kHz, and helium speech at depth, fs=20kHz. Note that an extensive discussion of the likely effects of this noise upon the voiced speech analysis presented here is contained in section 3.5 later in this chapter.

As regards the autocorrelation sequence or frame length, N, needed to provide the p values of autocorrelation lag necessary in the recursive estimation of the predictor error filter coefficients, this is of necessity at least N=p, and ideally as long as possible since the theoretical definition of autocorrelation in the continuous time domain assumes an infinite sequence of points, $N \rightarrow \infty$. This definition, however, assumes long-term statistical stationarity of the signal, which is certainly not the case for speech, whose stochastic characteristics can be expected to change on average roughly every 10-30mS [12]. Thus, since from equs.(3-15) it is seen that power spectral density information

is reflected in the autocorrelation sequence, intuitive reasoning suggests that this sequence should be calculated over a time frame of some 10-30mS, with preferably the inclusion of at least one pitch period to account for as much spectral detail as possible.

From a consideration of the pitch analysis results in Fig. 3.7, the average rate of change of the perceptual features of speech as related to the target vowels of list 4, Fig. 3.1, and the sample rates of the digitised data, an autocorrelation sequence length N=256 was chosen, corresponding to time frames of 25.6mS (fs=10kHz) for air speech and 12.8mS (fs=20kHz) for helium speech.

Tracking of the individual formants of each vowel from depth to depth was performed manually by visual identification of formant peaks. This task was found to be rendered difficult by the fall-off with increasing frequency of the spectral slope characteristic. Speech acoustic theory (see section 2.1.1) dictates that this fall-off should be at the rate of -6dB/octave above the fundamental frequency, comprising the -12dB per octave fall-off in the glottal excitation source spectrum and the +6dB/octave emphasis due to lip radiation. Thus, in order to cancel the -6dB/octave fall-off and thereby aid formant identification, a preemphasis of +6dB/octave was applied to the data [125] prior to processing. This was furnished by an approximate first-order digital differentiator with transfer function $1-\mu z^{-1}$, where the preemphasis factor, μ, is defined by the +3dB frequency, fp, of the spectral emphasis slope characteristic and by the sampling frequency, fs;

$$\mu = \frac{fs}{fs + fp} \qquad (3\text{-}19)$$

with fp for each depth being taken to be the average value of pitch frequency as given in Fig. 3.7.

Although the tracking of formants for any particular vowel sound was performed manually from depth to depth, numerical evaluation of formant centre frequency, amplitude and bandwidth was determined algorithmically by a peak-searching routine which identified resonance features in each spectral estimate. This algorithm basically assumes each peak to be the result of a complex pole pair in the transfer function of the vocal tract, and fits a parabola through both the peak frequency, f(c), and the two points about either side of this, f(c-1) and f(c+1). It is the resulting parabolic curve - whose apex may not necessarily lie on the point f(c) itself but will be within the region f(p)=f(c)+rf, where rf is the frequency resolution of the Fourier transform used to generate the power spectrum - which is used to provide estimates of formant centre frequency, amplitude and bandwidth. Note that fast Fourier transform (FFT) techniques [144], with a resolution of approximately 20Hz, were used to generate each spectrum.

Lastly, in an effort to mitigate the fluctuations in formant frequency and amplitude which inevitably occur over the duration of the voiced speech segment [64], several contiguous power spectra resulting from the analysis of each input speech frame of length N=256 points were added together for each individual vowel sound to produce one representative average spectrum to which the peak-picking algorithm described above was applied. The spectra from 9 contiguous frames in all were summed and averaged in each analysis, with a time spacing between successive spectral estimates of 3.2mS for both speech in air and helium speech. Thus, the overall time of analysis for each vowel segment is (25.6+8x3.2)=51.2mS for air speech and 38.4mS for helium speech. This is admissible in the vowel sections analysed since they relate to the group of voiced monophthongs, which are considered to be among the

more steady-state speech sounds, demonstrating long-term stability of formant frequencies and being of long duration (some 70-100mS), as is evidenced in Fig.3.4. The averaging of spectral frames in this way also has an important bearing upon the enhancement of the AR spectral estimate for input signals contaminated by Gaussian noise, and this topic is detailed in section 3.5.

A diagram of the sequence of operations for the AR spectral analysis of speech in this experiment is shown in Fig. 3.10. The starting point for the analysis on the digitised speech waveform of each vowel was determined by hand in an exactly similar manner to that for the analysis of fundamental frequency, see section 3.2.1 and Fig. 3.5.

3.4.2 CLASSICAL INTERPRETATION OF FORMANT SHIFT CHARACTERISTICS

An example of the AR spectrum produced as a result of applying the sequence of operations in Fig. 3.10 to the vowel "*er*" in "h*er*d" (uttered in air) is shown in Fig. 3.11, together with data, generated by the parabolic peak-searching algorithm, in respect of the first 3 formant peaks as chosen and subsequently tracked from depth to depth by visual identification based on a 'nearest approximation' criterion: that is, from a knowledge of the theoretical linear frequency shift commensurate with the heliox to air sound velocity ratio R as given for each depth in Fig. 3.3, then the nearest likely spectral peak to the frequency position f=FnxR, where Fn is the nth formant frequency in air, was taken to be the corresponding formant frequency Fhn in heliox. Note that it was found difficult to track the formants of the vowel "*awe*" in "h*awe*d", item 9, list 4, Fig. 3.1, with any confidence, and so the following discussions on formant frequency, bandwidth and amplitude

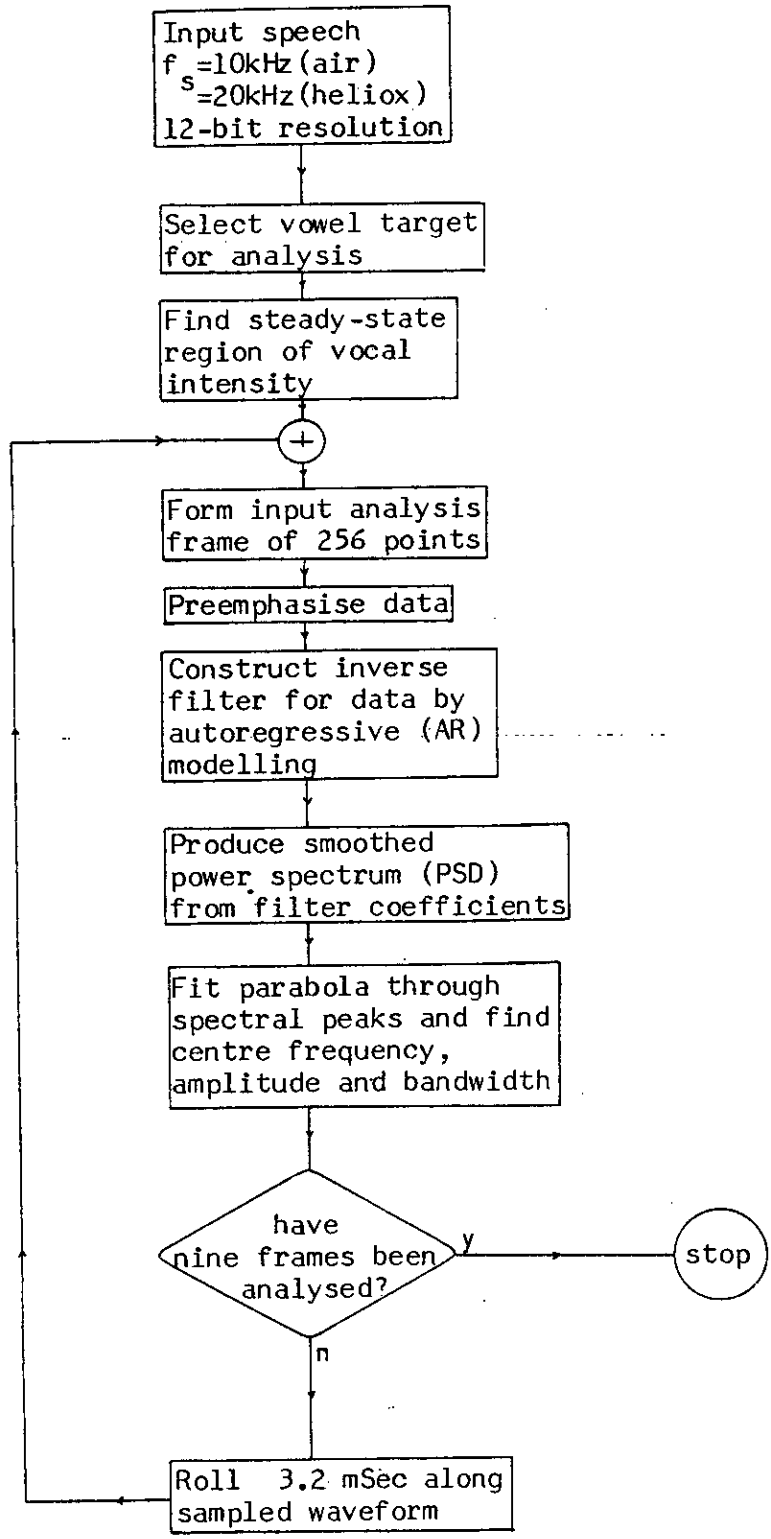
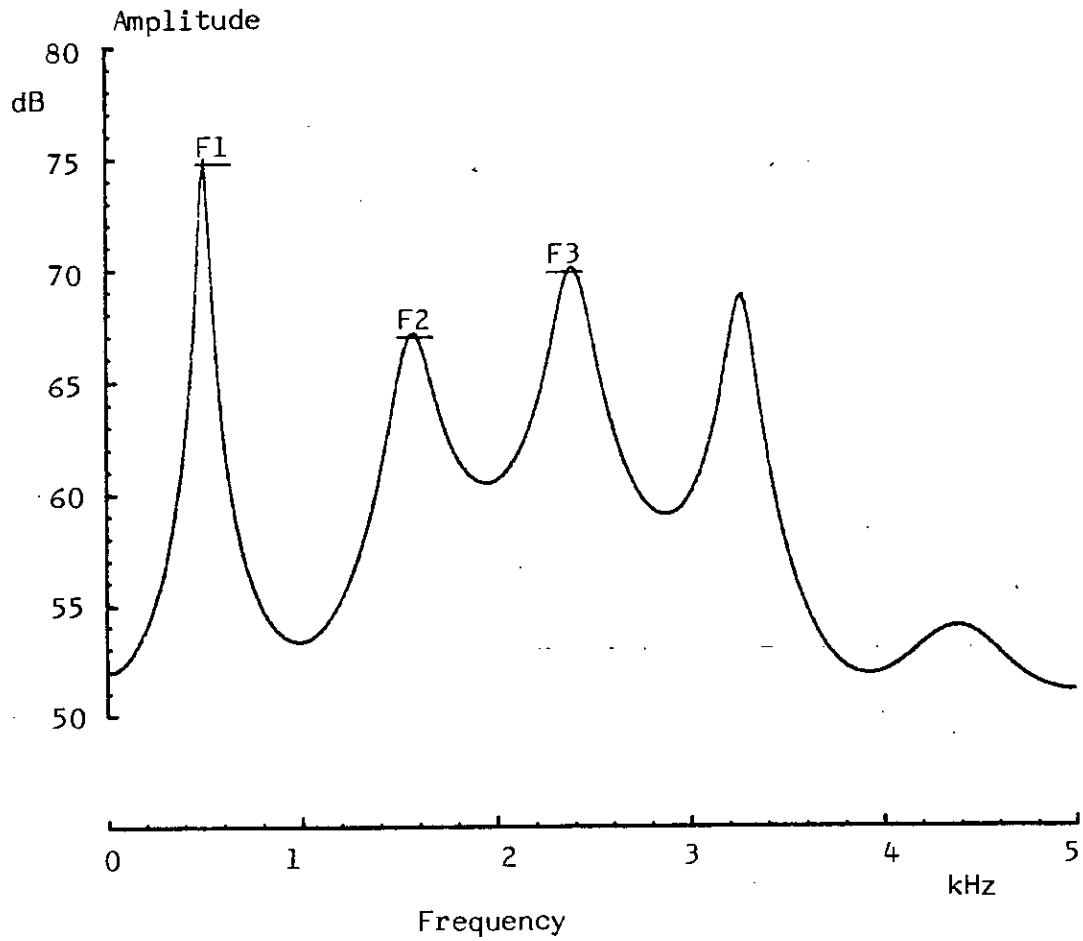


Fig. 3.10 Sequence of operations in the autoregressive (AR) spectral analysis of each vowel segment of list 4, Fig. 3.1.



Formant	F1	F2	F3
Frequency (Hz)	504	1572	2387
Amplitude (dB)	75.1	67.1	70.1
Bandwidth (Hz)	47	150	141

Fig. 3.11. AR spectrum of the vowel "*er*" in "h*er*d" uttered in air, demonstrating formant data as generated by the parabolic peak searching routine.

exclude this item.

Each of the 6 graphs of Fig. 3.12(a-f) shows the collective results for all vowel formant centre frequencies in air against corresponding formant centre frequencies in the respective heliox respiratory mixtures for each depth as tabulated in Fig. 3.3. The dotted curve shown on each graph corresponds to the theoretical nonlinear shift characteristic as defined by equ.(2-15), with closed vocal tract resonance Fwa=180Hz.

A first assessment of the results indicates that the theoretical nonlinear curve approximates roughly to the general perspective of the data. In order to test the nonlinear theory further, however, the results on any one graph in Fig. 3.12 were divided into two groups and a least mean squares (l.m.s) straight line approximation fitted to each group. The first group consisted of first formant frequency (F1) results only, that is , low frequency formants in heliox below roughly 1kHz, and the second group contained second and third formant frequencies (F2 and F3). The l.m.s fit through each group produces the equations:

$$fhn_1 = m_1 fn_1 + c_1 \qquad (3\text{-}20a)$$

and

$$fhn_2 = m_2 fn_2 + c_2 \qquad (3\text{-}20b)$$

where fhn is formant frequency in heliox, fn is formant frequency in air, m_1 and m_2 are the slopes of the derived best-fit lines for groups 1 and 2 respectively, and c_1 and c_2 are constants.

From a consideration of the rate of change of slope m of the nonlinear shift curve in the low frequency region of the spectrum, it can be deduced that for the data at 100ft depth (Fig. 3.12(a)), with decreasing frequency the slope m increases from m=R towards $m=\infty$, whereas at depths of 200ft and beyond, the slope decreases from m=R towards m=0, and therefore the l.m.s fit should produce values for slopes m_1 and m_2 such that $m_1 < m_2$ at these latter depths. The results

continued on pg.97

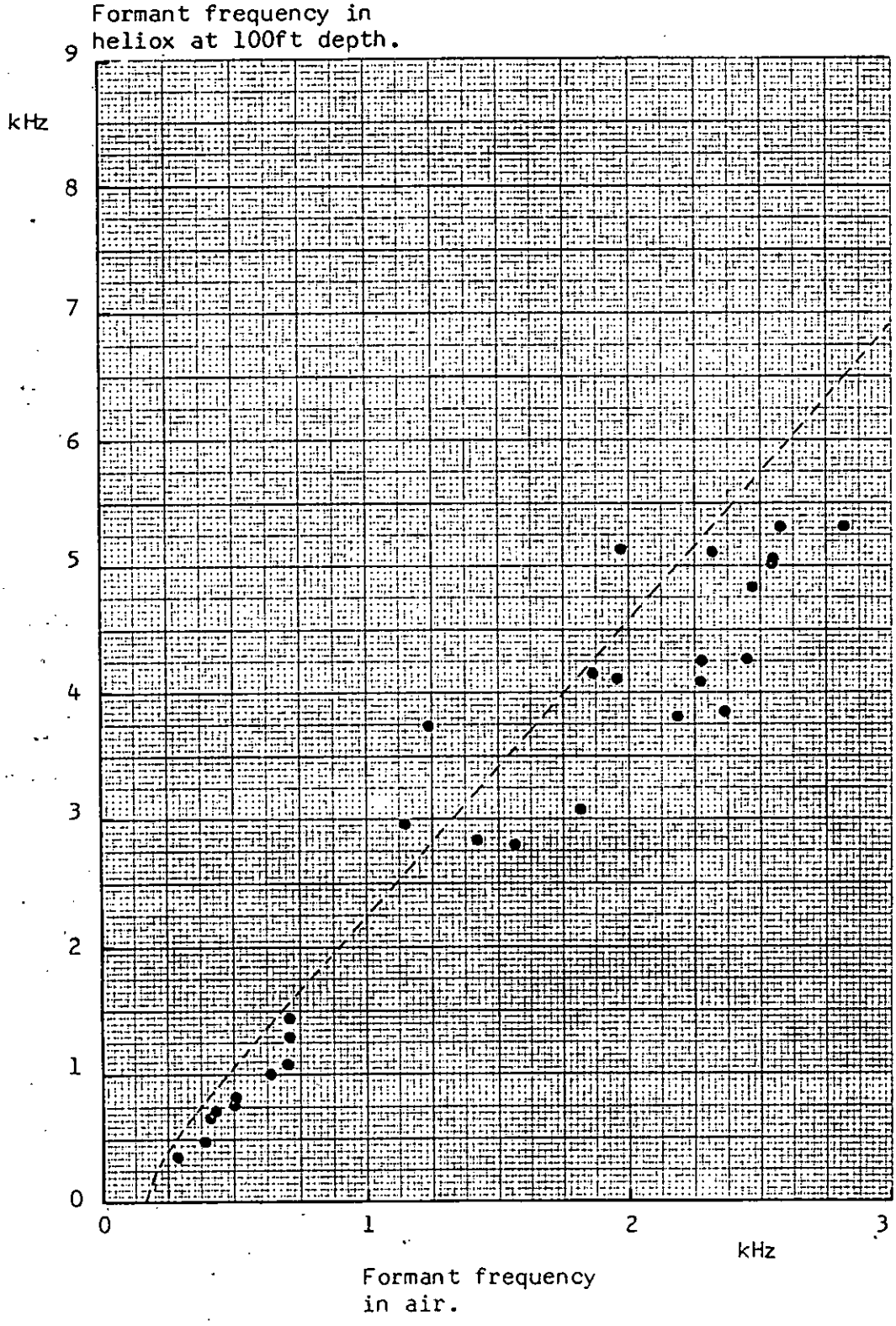


Fig. 3.12(a) Formant frequency in heliox at 100ft v. formant frequency in air. - - - theoretical nonlinear shift characteristic for Fwa = 180Hz.

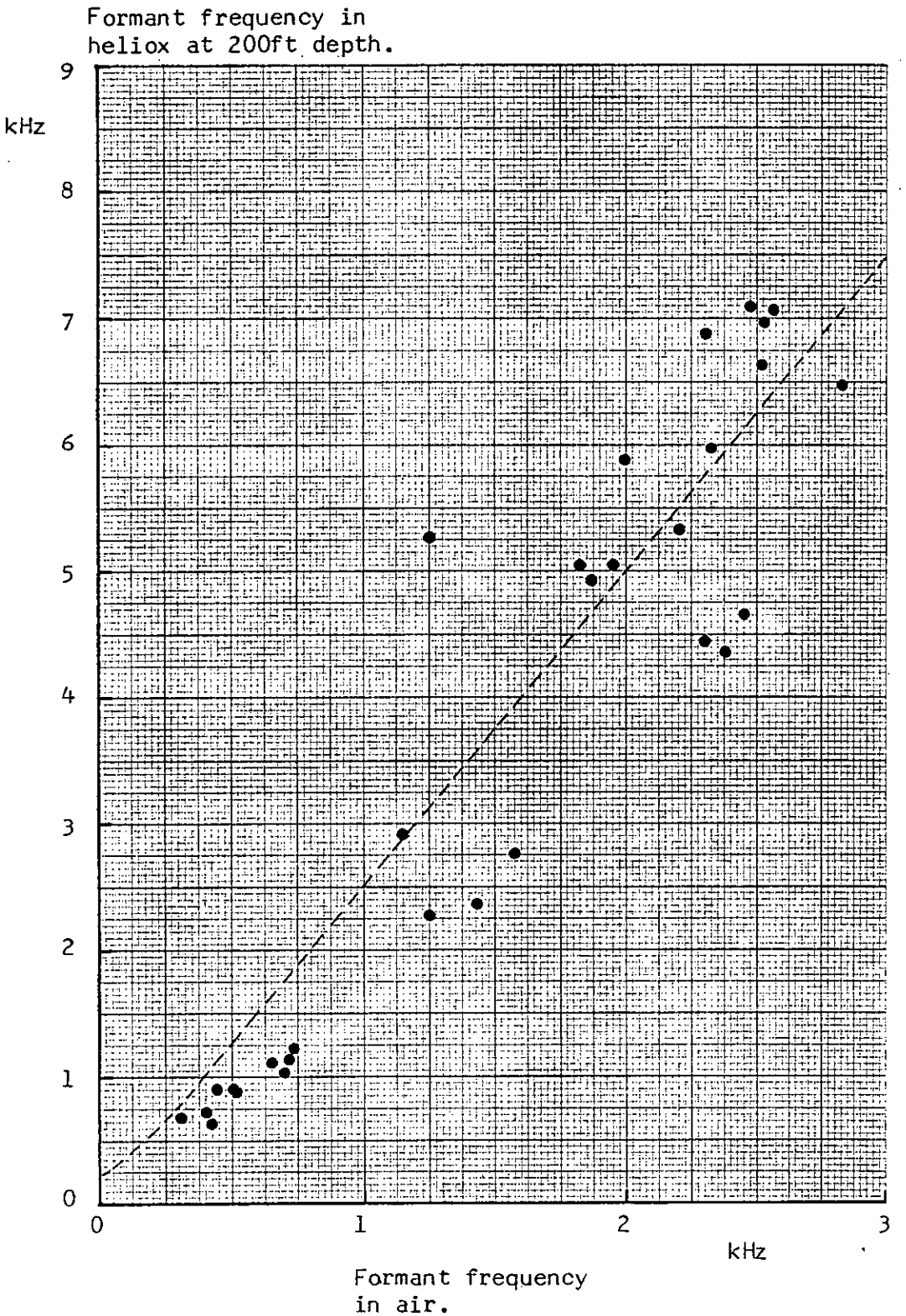


Fig. 3.12(b) Formant frequency in heliox at 200ft v. formant frequency in air. - - - theoretical nonlinear shift characteristic for Fwa = 180Hz.

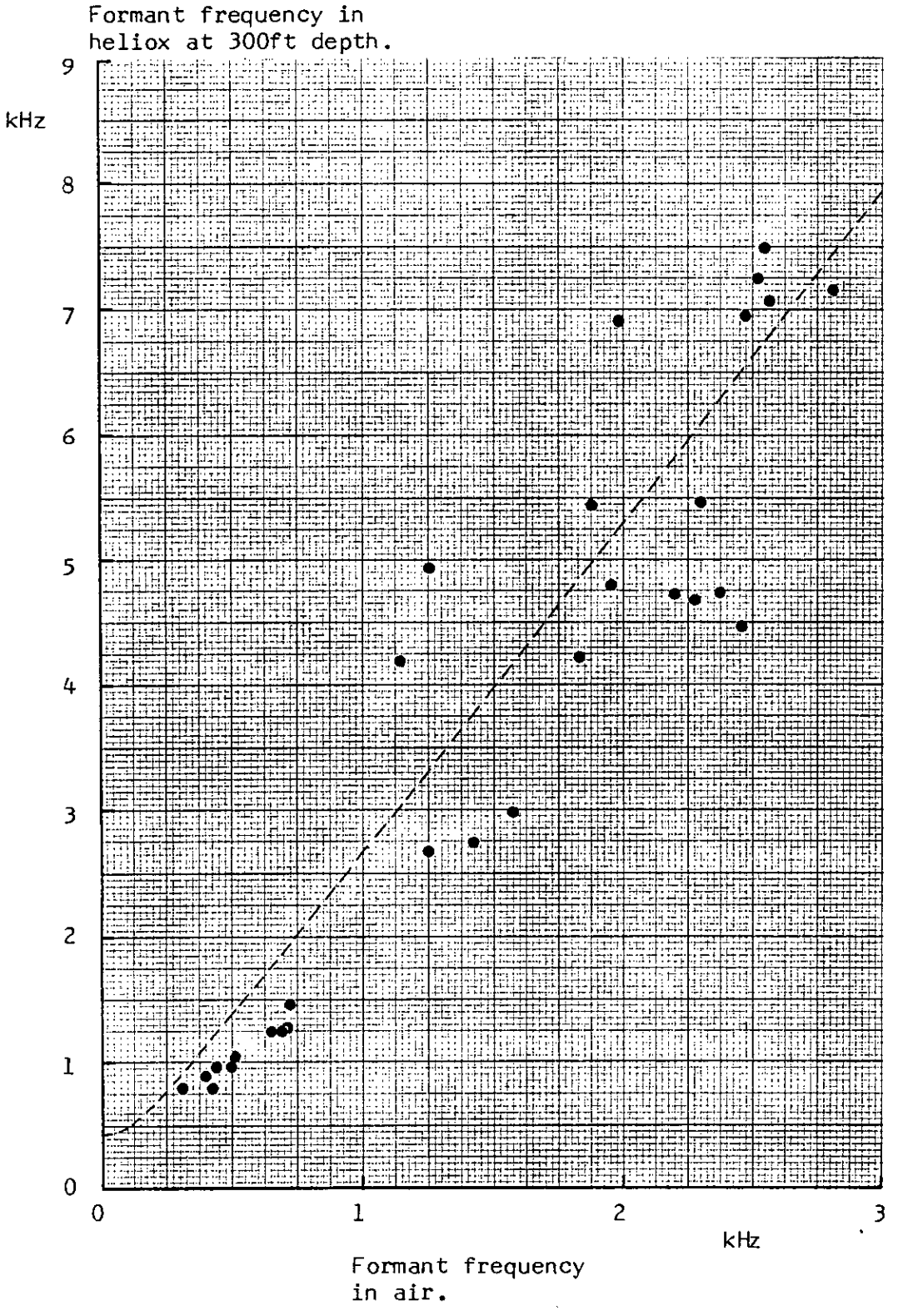


Fig. 3.12(c) Formant frequency in heliox at 300ft v. formant frequency in air. - - - theoretical nonlinear shift characteristic for Fwa = 180Hz.

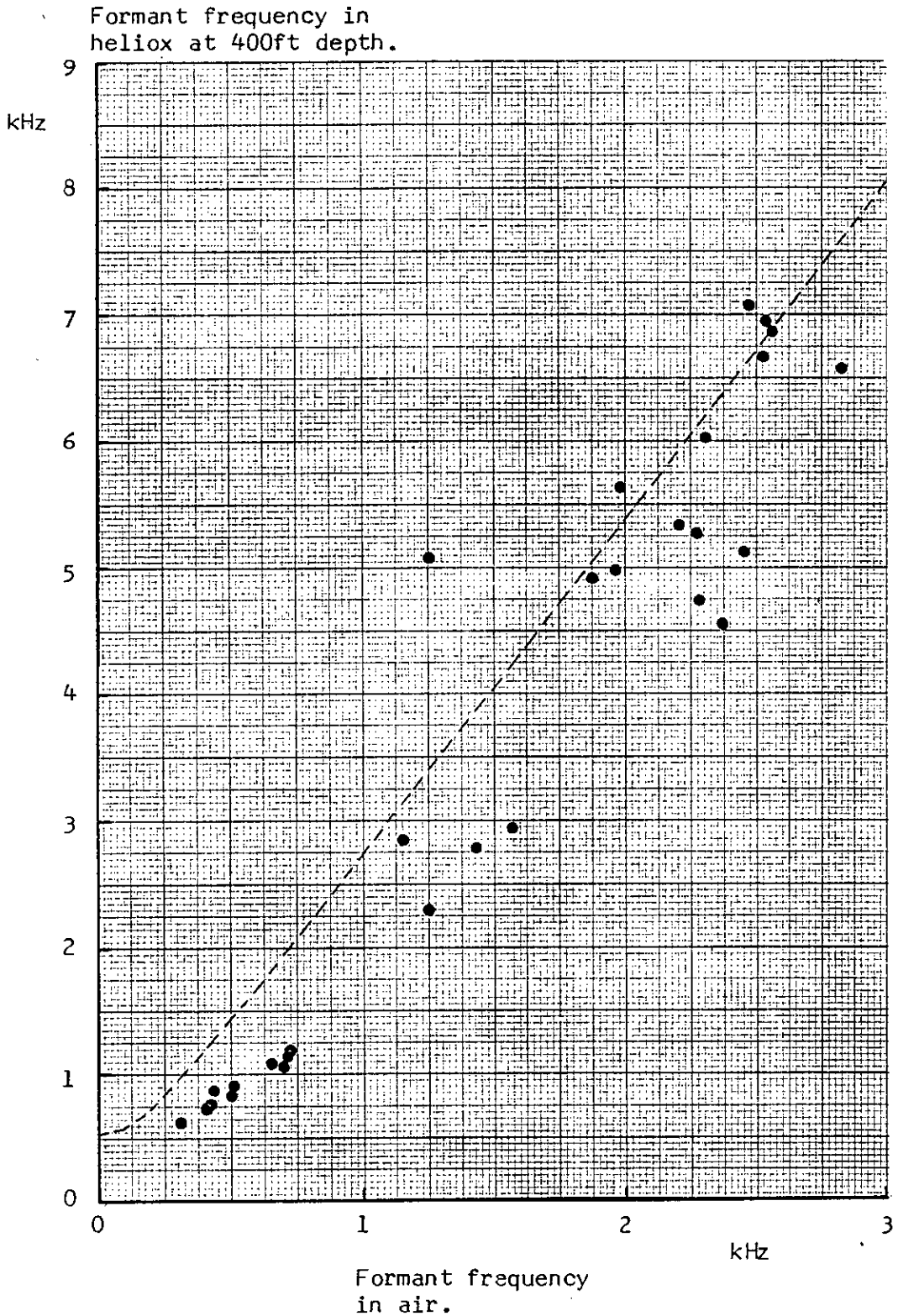


Fig. 3.12(d) Formant frequency in heliox at 400ft v. formant frequency in air. - - - theoretical nonlinear shift characteristic for Fwa = 180Hz.

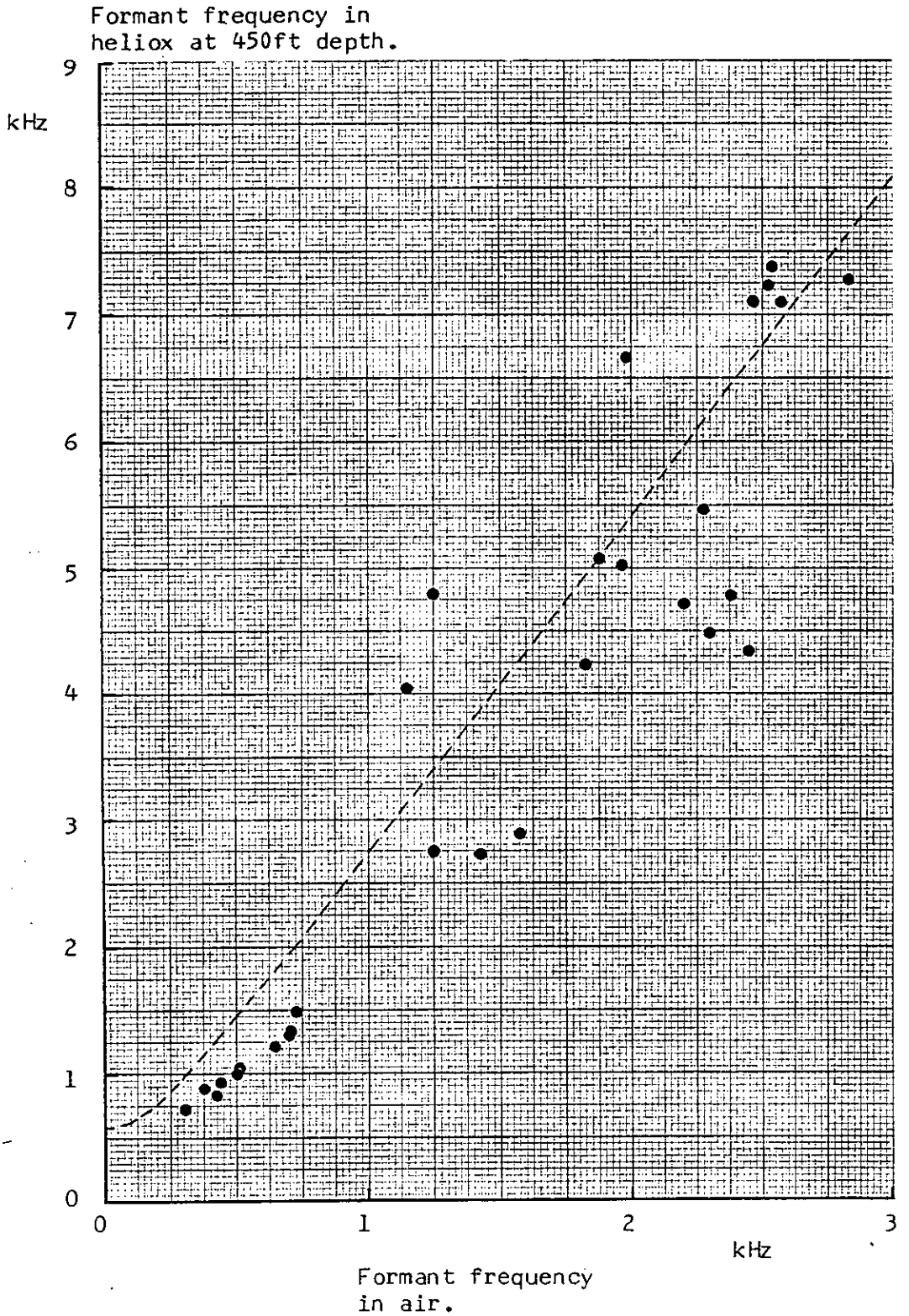


Fig. 3.12(e) Formant frequency in heliox at 450ft v. formant frequency in air. - - - theoretical nonlinear shift characteristic for Fwa = 180Hz.

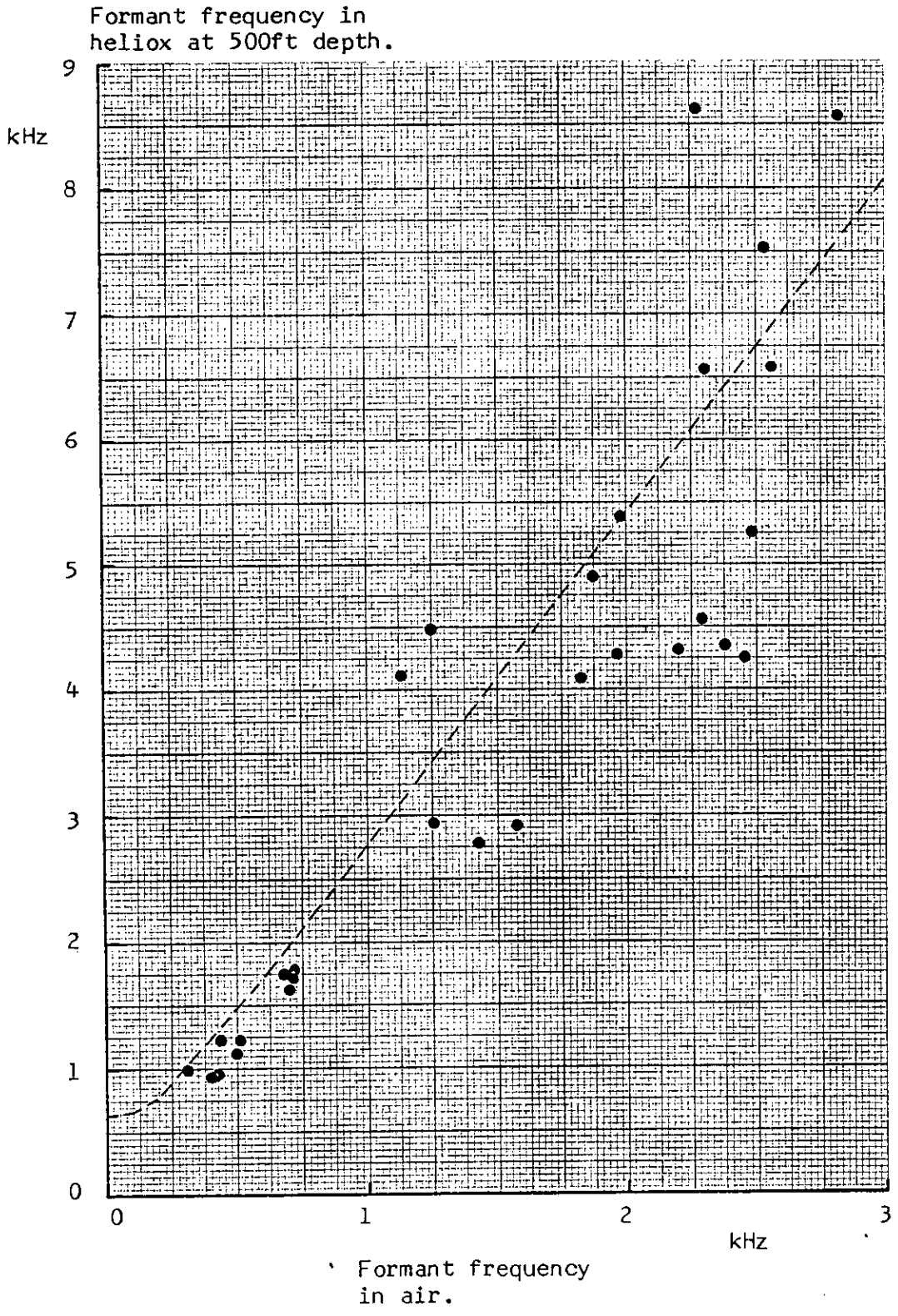


Fig. 3.12(f) Formant frequency in heliox at 500ft v. formant frequency in air - - - theoretical nonlinear shift characteristic for Fwa = 180Hz.

Least squares fit : $\boxed{f_{hn} = mf_n + c}$;

f_{hn} = formant frequency in heliox

f_n = formant frequency in air

Depth (ft)	F1 formant region m_1	F1 formant region c_1	F2 & F3 formant region m_2	F2 & F3 formant region c_2
100	2.19	-324	1.31	1370
200	1.2	267	2.2	494
300	1.52	246	2.13	723
400	1.27	228	2.24	312
450	1.6	202	2.2	532
500	2.16	156	1.63	1321

Fig. 3.13 Piecewise-linear least mean squares (l.m.s) fit through formant data at each depth.

of applying this piecewise linear fit to the data at each depth are shown in Fig. 3.13, from which it can be seen that for 100ft depth, $m_1 > m_2$ as predicted, and for all other depths, with the exception of the results for 500ft depth, it is seen that $m_1 < m_2$. Moreover, note that consistent with the nonlinear curve, the value of c_1 at 100ft depth is -ve and values at lower depths have +ve values, indicating that there may indeed be resonance of the vocal tract wall tissue as outlined in section 2.2.3. Thus, the initial appreciation of the results appears to confirm the nonlinear characteristic proposed by equ.(2-15), i.e. it is the low frequency region of the speech spectrum which is shifted nonlinearly due to the effects of vocal tract wall vibration. Before proceeding further with this matter, however, consider the large spread of formant frequency values in each graph of Fig. 3.12(a-f), which is particularly evident in the high frequency area of the spectrum.

Similar results have been noted before [71,105] and have been put down to random adaptation by the diver of his own speech to render it more intelligible as he himself judges. However, in the new interpretation presented here, it has been possible to demonstrate consistent trends directly relating to this apparently random spread of formant frequencies.

3.4.3 PHONEME-SPECIFIC FORMANT SHIFT PROFILES IN HELIOX

The vowel formant profiles shown in Fig. 3.14(a-j) are a reinterpretation of the formant frequency results of Fig. 3.12. Basically, the formants of individual vowel phonemes have been deconvolved from each of the collective results of Fig. 3.12 and collated to form one new graph: thus each graph in Fig. 3.14 pertains to one vowel sound only, and the formant frequencies at each depth or heliox mixture are now plotted together on the same graph.

In Fig. 3.14(a), the data for F1, F2 and F3 corresponding to any single depth have been coded by the same letter (e.g. 'b' identifies F1, F2 and F3 formants at 200ft depth); tracing of the letter code from F1 to F2 to F3 provides a characteristic formant shift profile for each depth, and reveals a significant trend: each formant profile demonstrates a similar characteristic shape irrespective of depth, that is, composition of the heliox respiratory mixture. Moreover, the profile is of a different shape to the predicted quadratic relationship of equ.(2-15). A similar phenomenon has been observed in the case of each vowel analysed, with the formant data at each depth for any one vowel producing a characteristic profile. Thus, for clarity, in Fig. 3.14(b-j), the dotted line shown in each graph joins the mean values of F1, F2 and F3 averaged over all depths, and can be justifiably taken to represent the

continued on pg.103

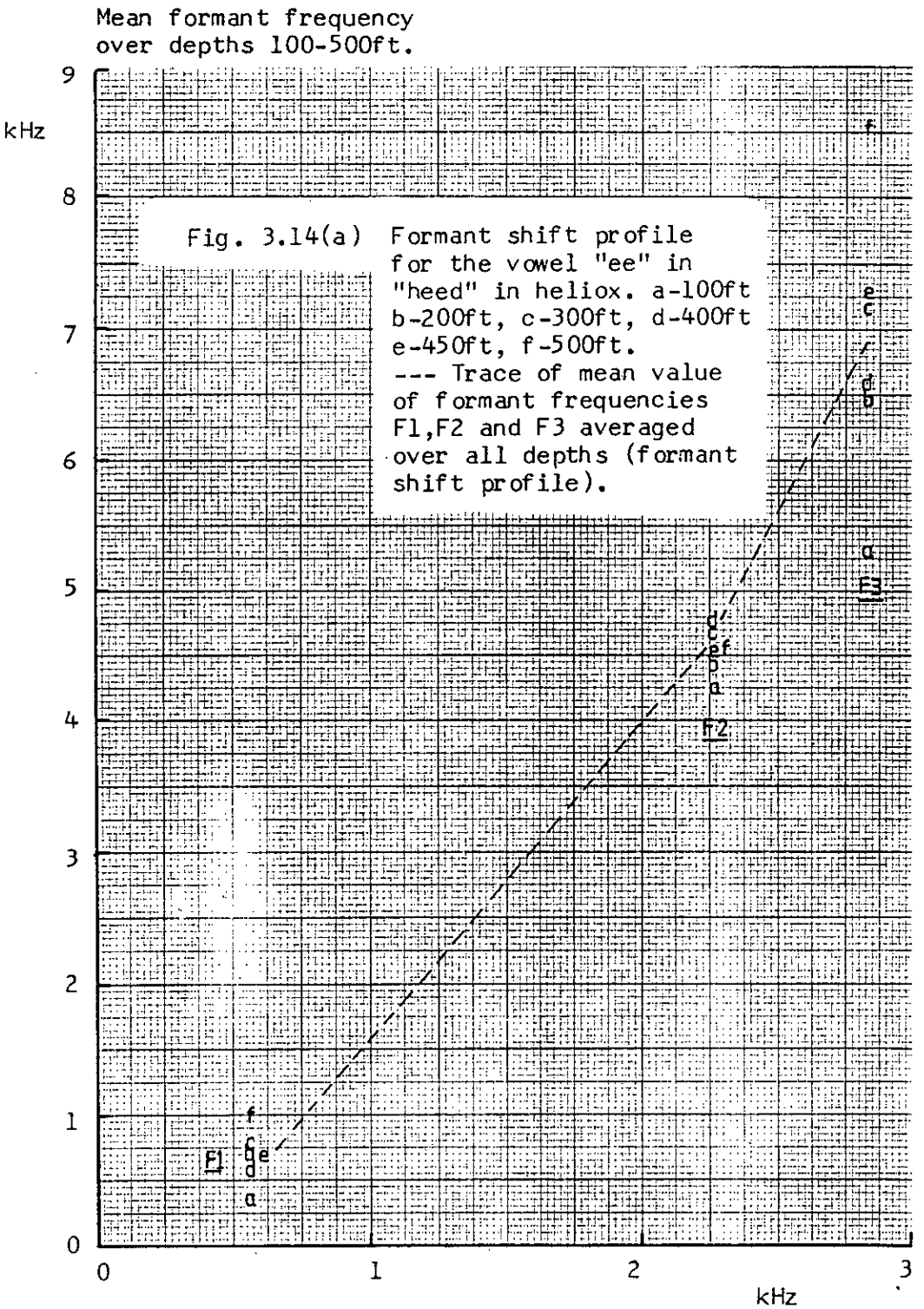
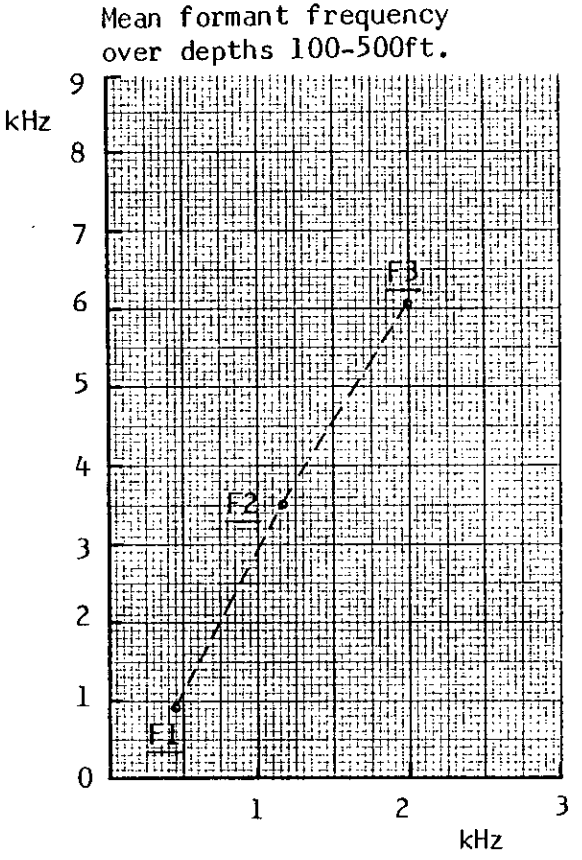
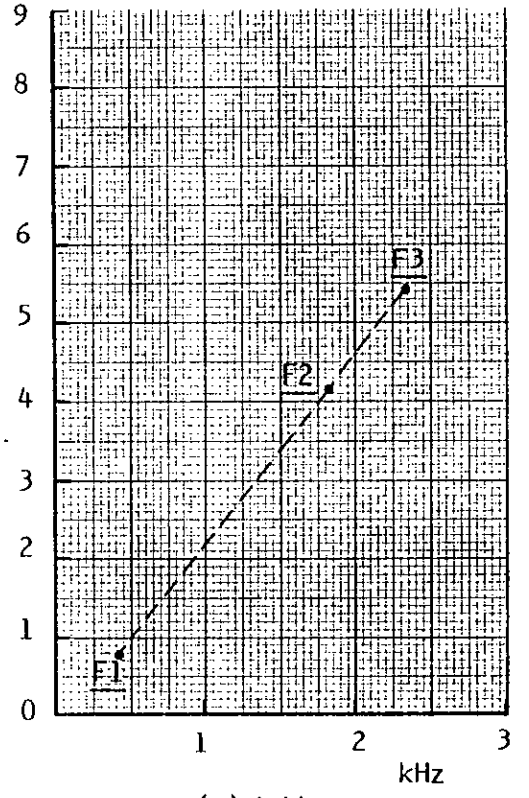


Fig. 3.14(a) Formant shift profile for the vowel "ee" in "heed" in heliox. a-100ft b-200ft, c-300ft, d-400ft e-450ft, f-500ft. --- Trace of mean value of formant frequencies F1,F2 and F3 averaged over all depths (formant shift profile).

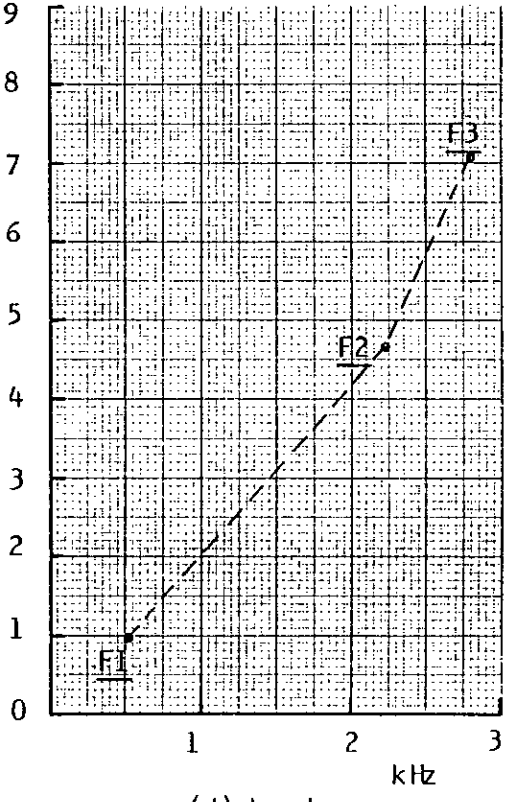
(a) heed



Formant frequency
in air.
(b) hood



(c) hid



(d) head

Fig. 3.14(b-d) Formant shift profiles
in heliox.

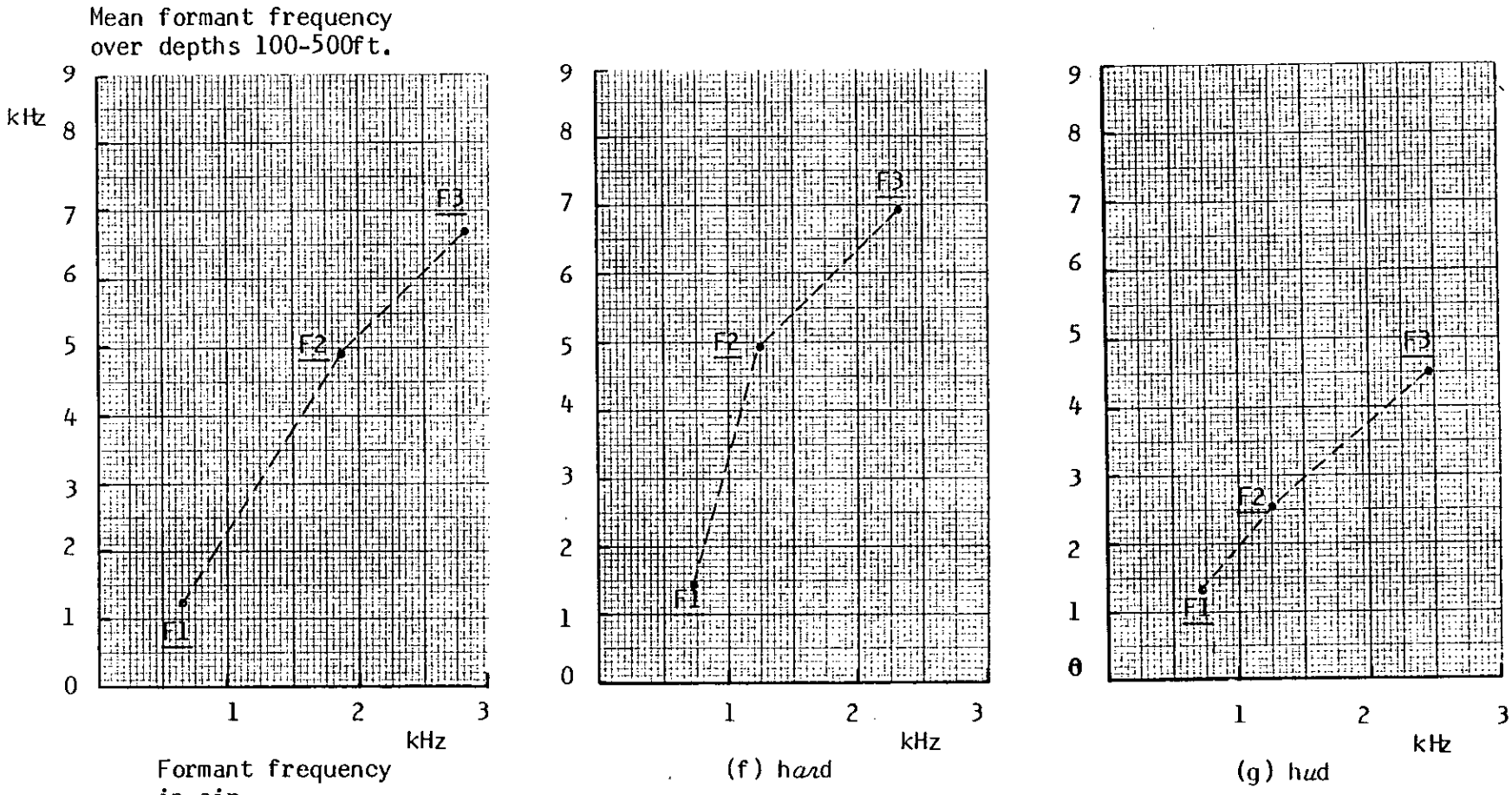
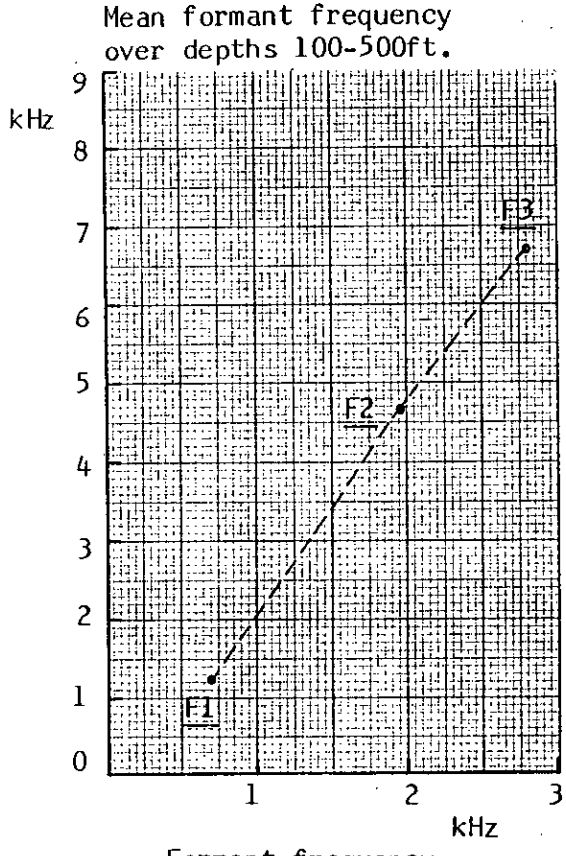
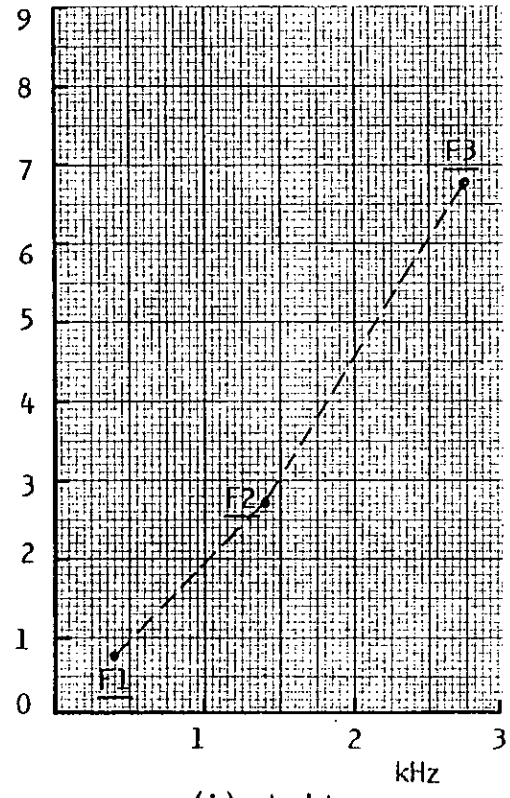


Fig. 3.14(e-g) Formant shift profiles in heliox.

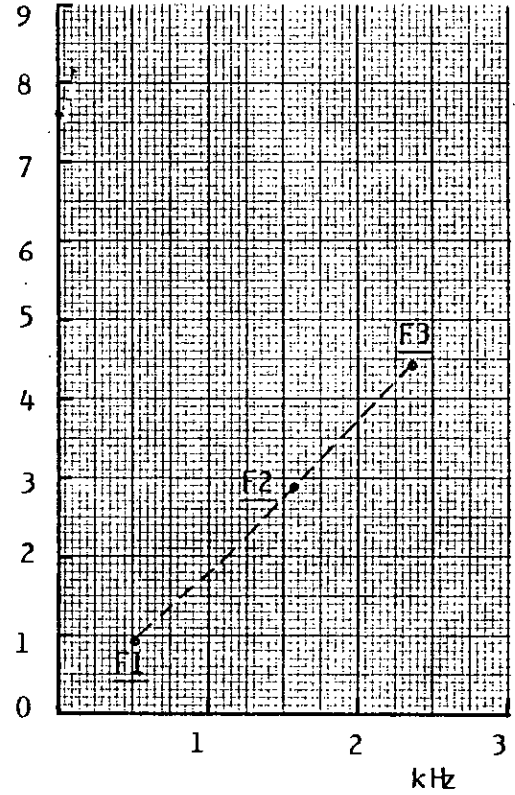


Formant frequency
in air.

(h) hod



(i) who'd



(j) heard

Fig. 3.14(h-j) Formant shift profiles
in heliox.

characteristic formant shift profile for each vowel phoneme in question.

A comparison of each of the profiles of Fig. 3.14(a-j) against each other indicates that there is an important disparity in profile shape from one vowel phoneme to another. Indeed, the results of these graphs imply that each vowel has a tendency to produce an individual formant frequency shift signature in a pressured heliox respiratory mixture.

A possible cause for this effect is that (a) the subject may be trying to consciously alter his speech in a consistent manner for each phoneme so as to palliate the effect of formant frequency translation due to the physical properties of the heliox gas mixture. Since each speech sound involves different articulatory members in its production (see section 2.1), then this may explain the different profiles for each phoneme and the consistency of profile shape with depth. It is also possible (b) that the helium speech waveform may in itself be acceptable to the talker, but that the feedback of his own voice through the heliox environment may induce him to make changes in his speech output. The reasoning offered here is similar to that relating to the disruption of the auditory feedback mechanism discussed in section 2.2.5, in that the arrival time at the ear of the speech waveform transmitted through the heliox atmosphere will be different compared to that of air, whereas the arrival time of the signal transmitted by, for example, bone conduction will remain unchanged. The talker may therefore be adapting his speech to minimise the effect of the unusual combination of speech waveforms presented to his own perceptual system. Which, if any, of the above causes is responsible for producing the formant profiles has important implications in terms of helium speech unscrambler architecture, and this is discussed further in section 3.6.

3.4.4 UNIDENTIFIABLE FORMANT FREQUENCIES AT DEPTH

In the course of tracking vowel formants from depth to depth, it was noticed that there was a consistent occurrence of formant features for certain vowels, within the frequency range specified by F1-F3, for which no corresponding formant correlates in normal air speech could be found. The previous discussions of formant frequency shift characteristic, by their very nature, preclude the inclusion of such data. Certainly no other research into the characteristics of voiced helium speech could be found which broached this subject, and so the results presented here relate to a newly observed and hitherto unreported phenomenon in helium speech.

Figures 3.15(a-j) are another interpretation of the formant frequency shift data, but now allow inclusion of the unidentifiable formant data. The solid lines join formants F1, F2 and F3 whose origins are directly traceable to normal air speech, and the dotted lines indicate the new formants U1 and U2 which have appeared at depth. The table in Fig. 3.15(k) displays the formant shift ratio with respect to air relating to the directly traceable formant data F1, F2 and F3.

Figure 3.15(ℓ) is a tabulation of the translation back to air of unidentifiable formants U1 and U2, assuming the linear formant shift ratios given by the velocity of sound ratio R as defined in Fig. 3.3. It can be seen that unidentifiable formants appear at depth in 70% of the vowels analysed, and the translated values fall roughly into three frequency regions, around 1kHz, 1.5kHz and 2kHz.

Similar phenomena have already been noted in the study of the nasalisation of vowels in a normal air environment [145], in which 'extra' formants were observed to occur between the known formants for non-nasal vowels for subjects with a pronounced nasal quality to their voice,

continued on pg.11

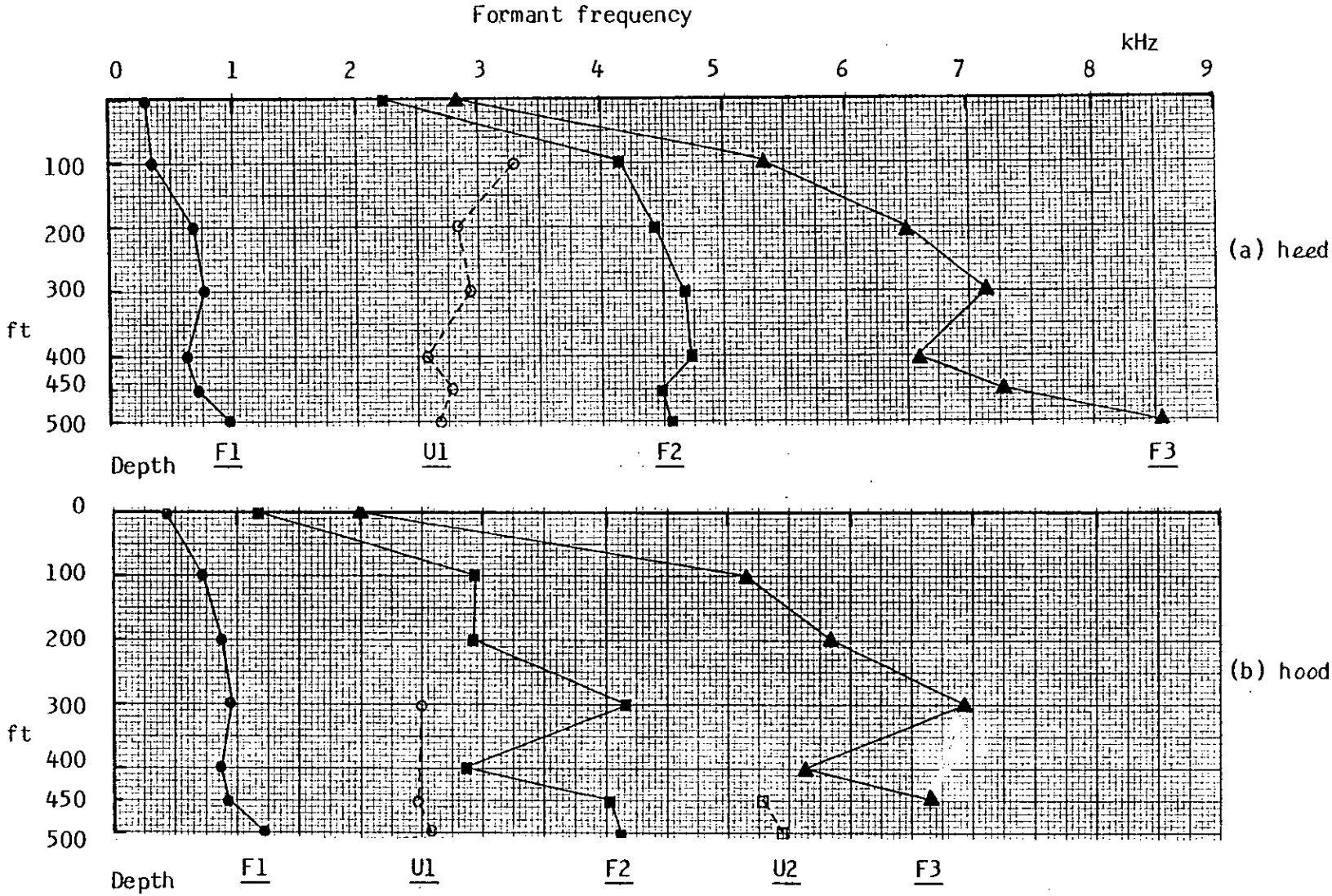


Fig. 3.15(a-b) Trace of formant frequencies with depth.

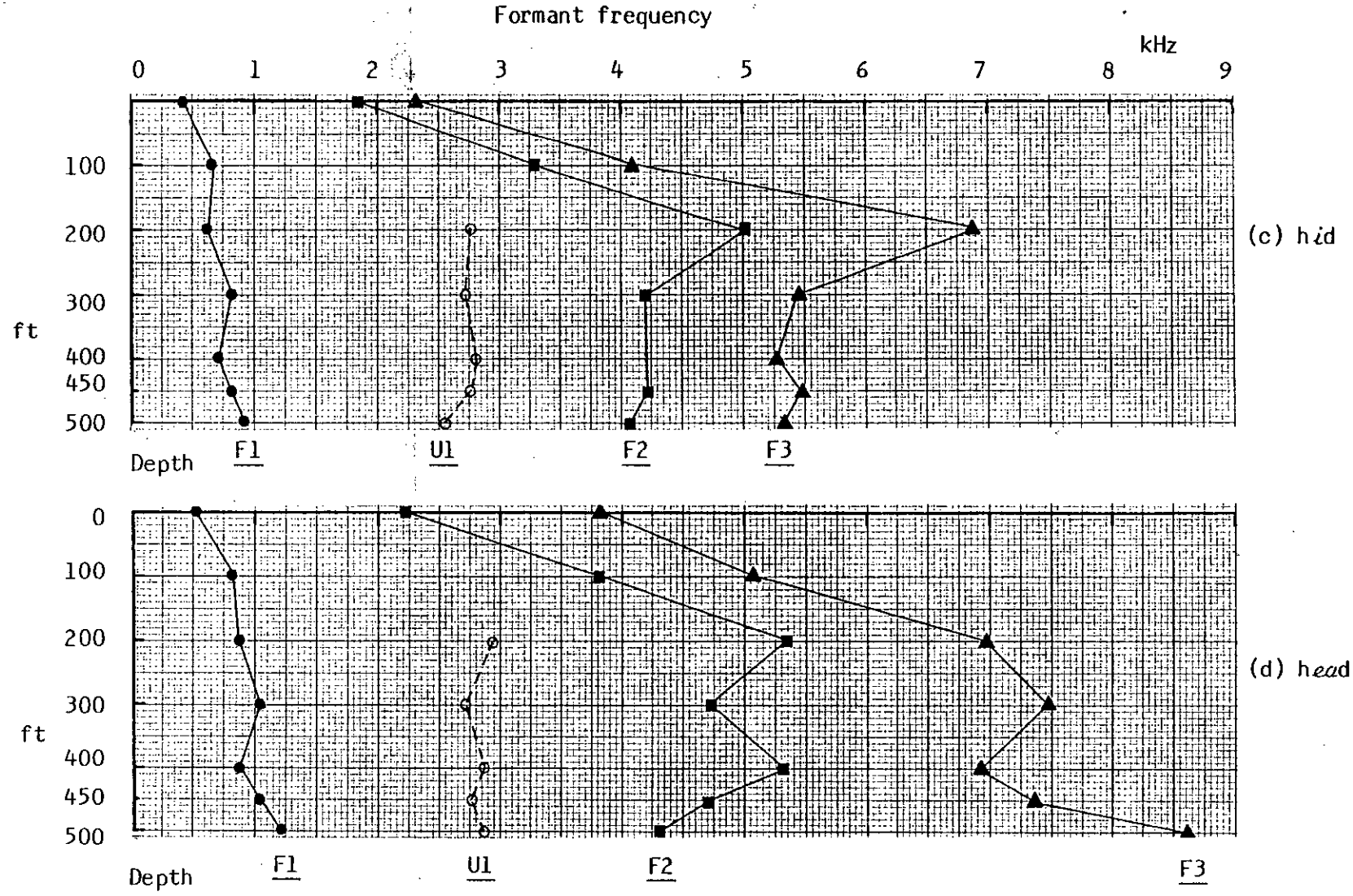


Fig. 3.15(c-d) Trace of formant frequencies with depth.

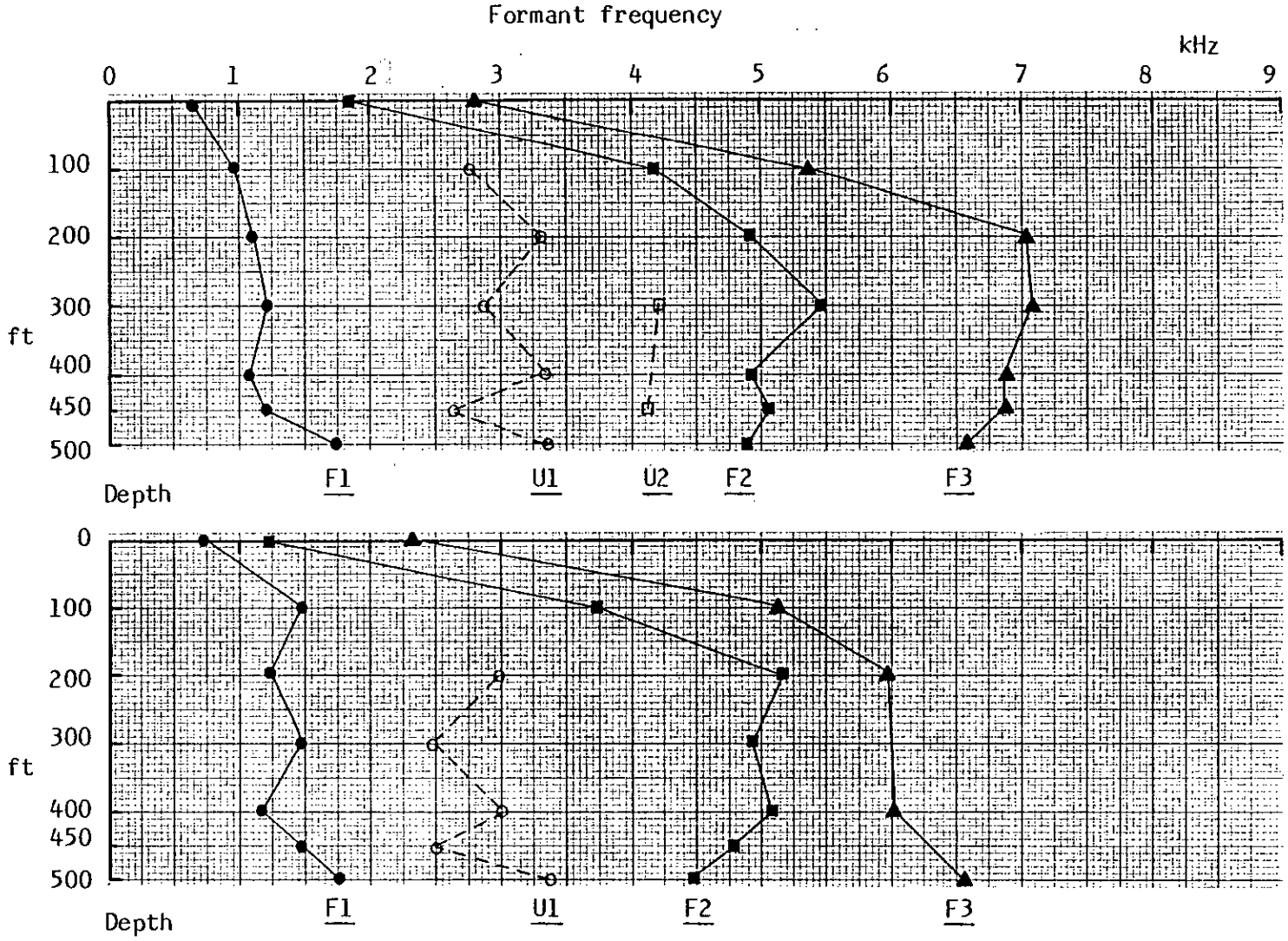


Fig. 3.15(e-f) Trace of formant frequencies with depth.

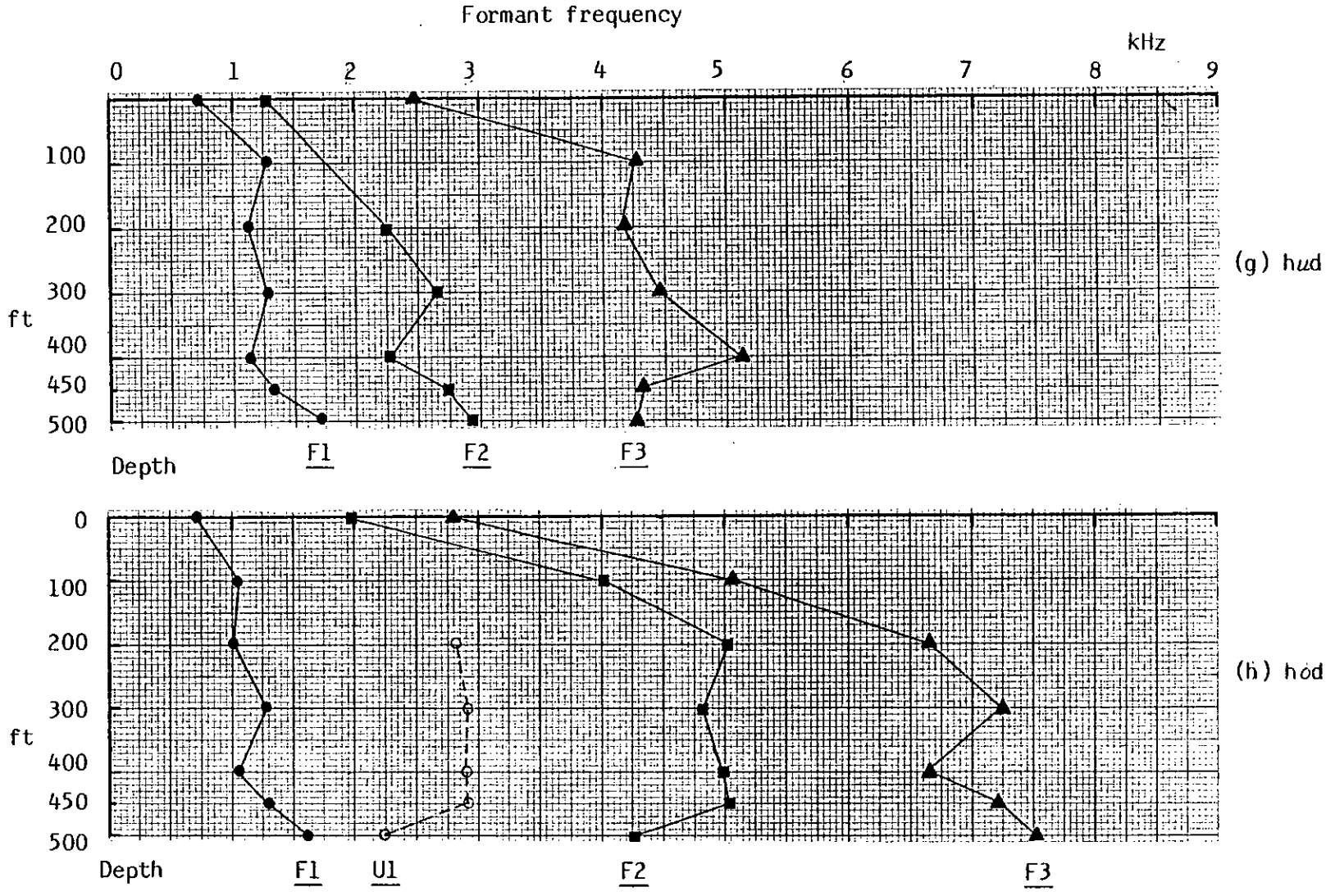


Fig. 3.15(g-h) Trace of formant frequencies with depth.

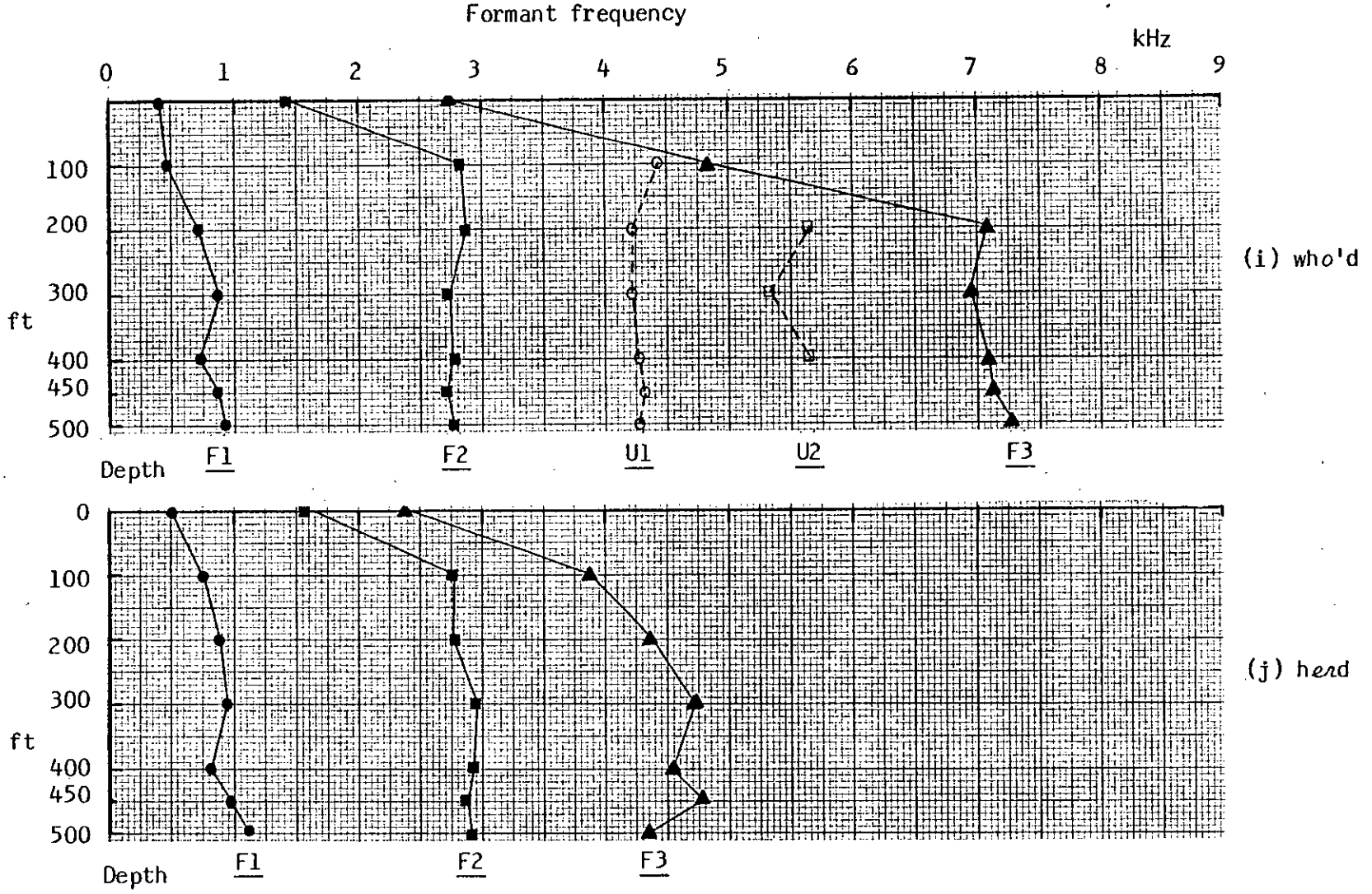


Fig. 3.15(i-j) Trace of formant frequencies with depth.

First formant (F1) shift ratios in heliox

Item — Depth (ft)	heed	hood	hid	head	had	hard	hud	hod	who'd	herd
100 -	-1.16	1.63	1.55	1.63	1.46	2.04	1.8	1.51	1.19	1.5
200 -	-2.27	2.0	1.42	1.71	1.66	1.66	1.56	1.46	1.8	1.78
300 -	-2.57	2.14	1.88	2.01	1.86	2.04	1.8	1.84	2.14	1.88
400 -	-2.06	1.95	1.67	1.72	1.65	1.65	1.59	1.5	1.79	1.65
450 -	-2.35	2.1	1.9	2.01	1.8	2.06	1.83	1.85	2.18	1.97
500 -	-3.21	2.78	2.17	2.36	2.63	2.83	2.37	2.29	2.3	2.21

Second formant (F2) shift ratios in heliox

Item Depth (ft)	heed	hood	hid	head	had	hard	hud	hod	who'd	herd
100 -	-1.86	2.53	1.69	1.73	2.24	2.2	-	2.04	1.98	1.77
200 -	-1.94	2.51	2.75	2.41	2.64	4.21	1.8	2.55	2.01	1.76
300 -	-2.05	3.6	2.31	2.13	2.92	3.94	2.12	2.45	1.92	1.88
400 -	-2.06	2.46	-	2.41	2.64	4.07	1.83	2.53	1.94	1.86
450 -	-1.96	3.47	2.31	2.13	2.71	3.82	2.19	2.55	1.9	1.83
500 -	-1.99	3.53	2.24	1.95	2.62	3.58	2.35	2.17	1.94	1.86

Third formant (F3) shift ratios in heliox

Item Depth (ft)	heed	hood	hid	head	had	hard	hud	hod	who'd	herd
100 -	-1.88	2.58	1.77	1.81	1.9	2.21	1.74	1.81	1.77	1.63
200 -	-2.29	2.94	2.99	2.48	2.5	2.57	1.89	2.37	2.6	1.82
300 -	-2.53	3.47	2.37	2.67	2.51	-	1.82	2.59	2.54	1.97
400 -	-2.33	2.82	2.3	2.47	2.44	2.59	2.08	2.38	2.59	1.9
450 -	-2.58	3.34	2.37	2.63	2.52	-	1.76	2.58	2.6	2.0
500 -	-3.02	-	2.34	3.07	2.36	2.83	1.73	2.68	2.66	1.82

Fig. 3.15(k) Table of formant shift ratios with respect to formant frequency in air for F1 - F3.

Item	heed	hood	hid	head	had		hard	hod	who'd	
Depth (ft)	U1	U1	U1	U1	U1	U2	U1	U1	U1	U2
100	1442	-	-	-	1229	-	-	-	1914	-
200	1127	-	1098	1164	1316	-	1186	1123	1676	2222
300	1126	1378	1035	1034	1094	1601	946	1112	1600	2019
400	949	-	1043	1070	1248	-	1248	1077	1592	2080
450	1030	923	1022	1019	979	1526	928	1084	1592	-
500	984	958	944	1069	1238	-	1255	827	1586	-

Fig. 3.15(ℓ) Table of unidentifiable formant frequencies translated to air values by the velocity ratio R (see Fig. 3.3).

such as those with cleft palates. The appearance of these unidentifiable formants at depth may therefore be related to some extent to the pronounced nasal quality of helium speech, and this topic is explored further in the discussion in section 3.6.

3.4.5 FORMANT AMPLITUDE CHARACTERISTICS IN HELIOX

The collective formant amplitude data for all vowels pooled together at each depth, including the data in air at the surface, is shown in Fig. 3.16(a-g). The amplitude data displayed here has been corrected for spectral preemphasis, which is possible, from equ.(3-19), if both the preemphasis factor, μ, and sampling frequency, fs, are known. Thus, the data reflects the true amplitude of formant peaks in the power spectrum of the acoustic waveform. A 5-point running median filter was applied to the data, and the resulting output is given by the dotted line on each graph. Lastly, as a guide, an amplitude decay slope of

continued on pg.11

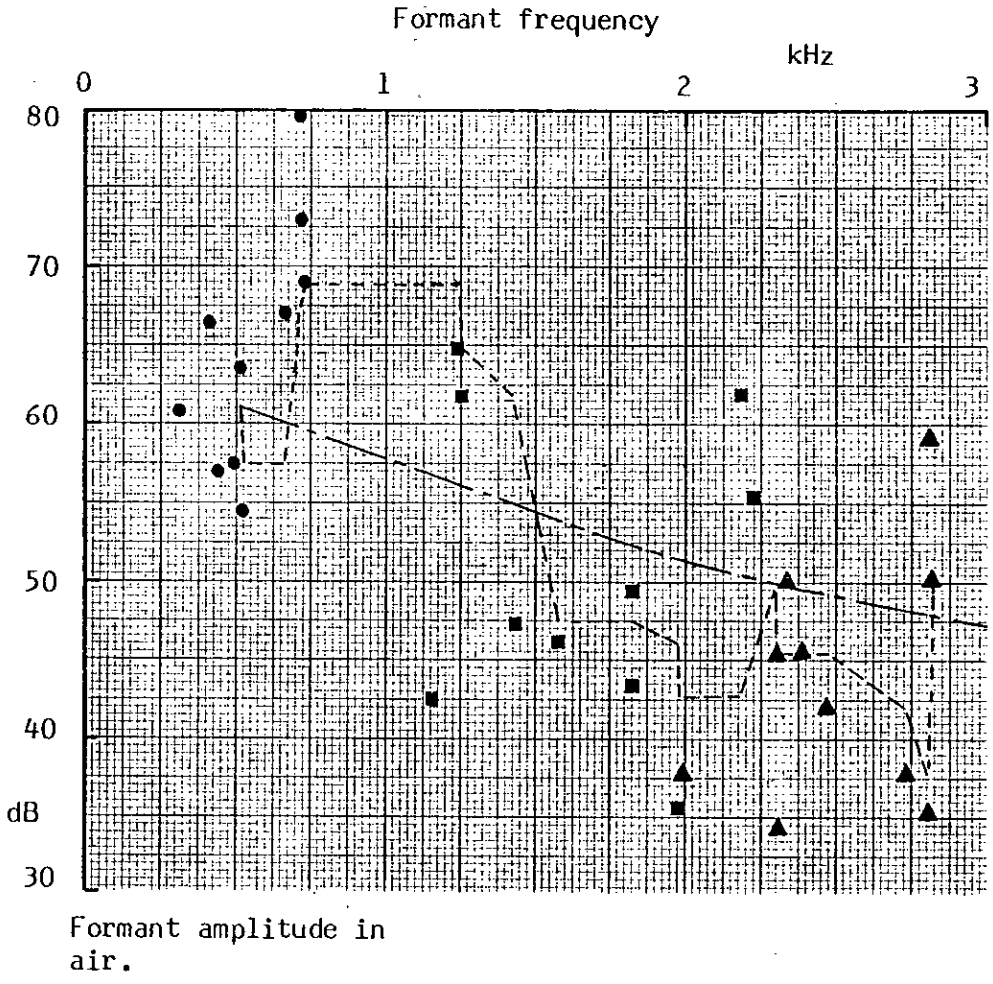


Fig. 3.16(a) Formant frequency v. amplitude in air.
●=F1, ■=F2, ▲=F3
--- Running median (5 point)
—·— -6dB/octave slope.

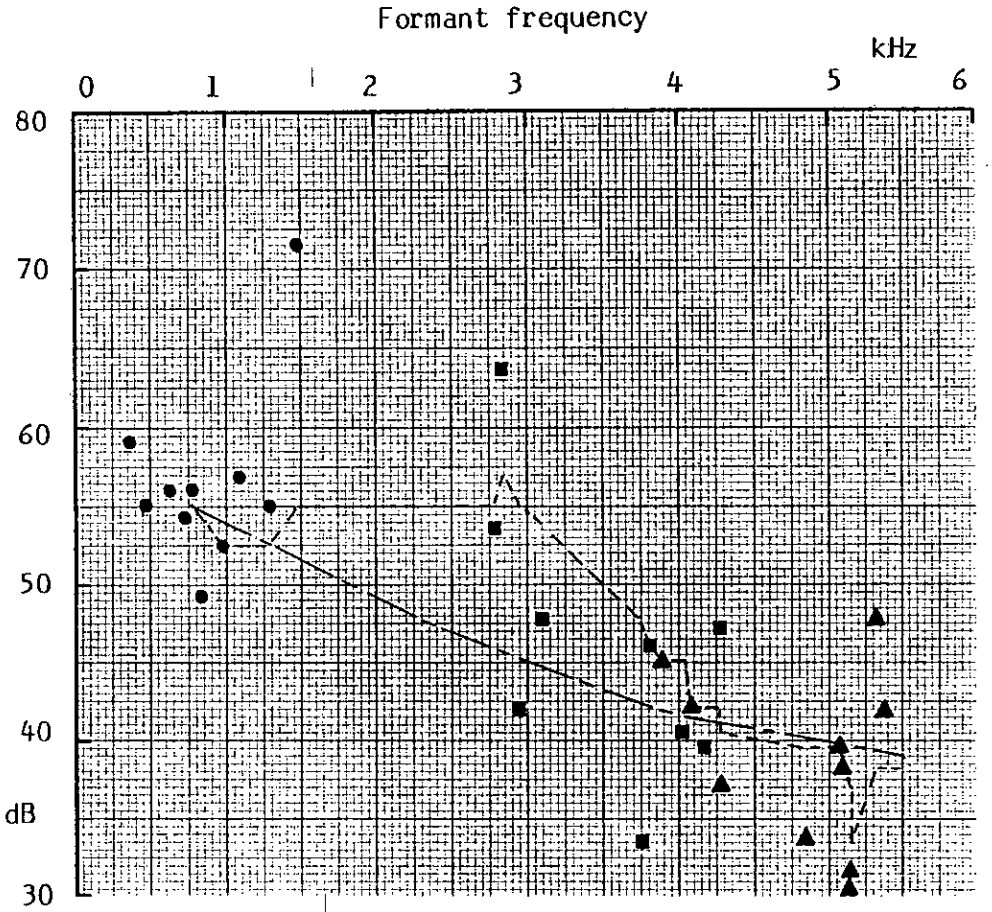


Fig. 3.16(b) Formant frequency v. amplitude at 100ft.
●=F1, ■=F2, ▲=F3.
--- Running median (5 point)
—·—. -6dB/octave slope.

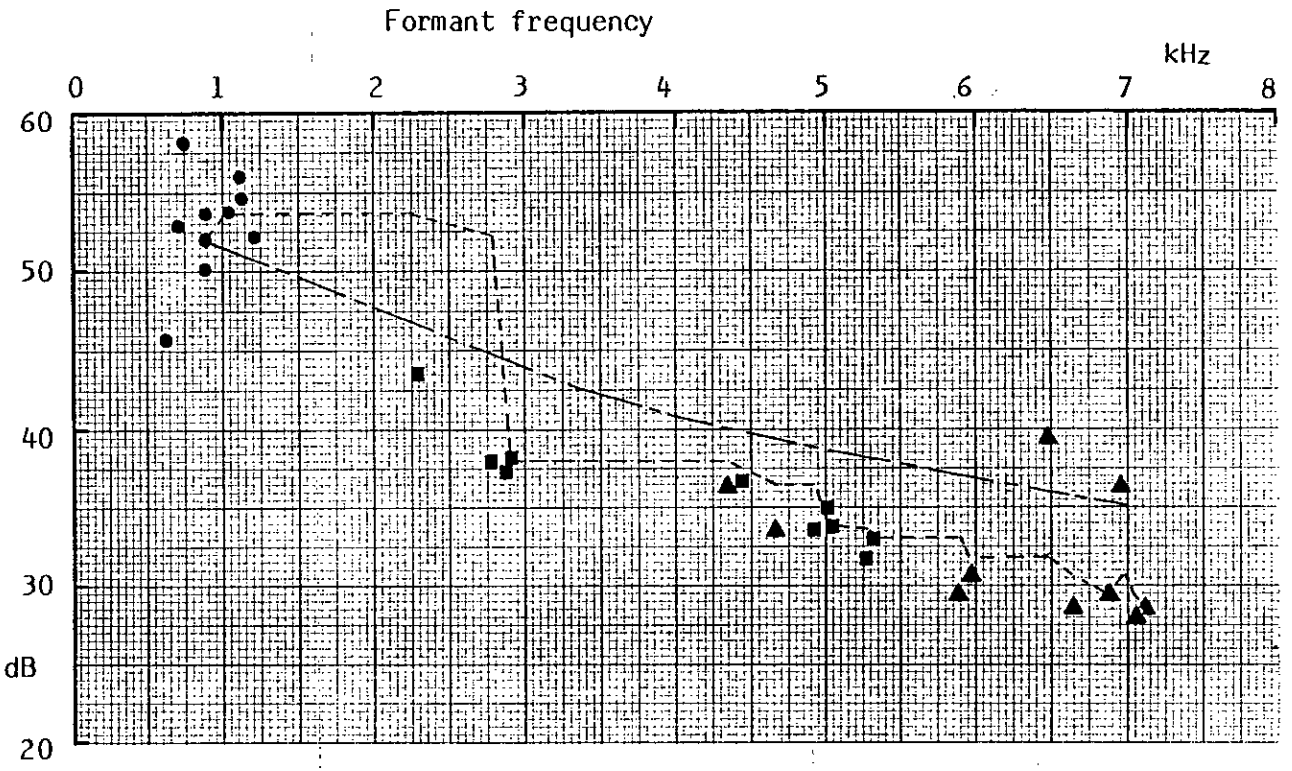


Fig. 3.16(c) Formant frequency v. amplitude at 200ft.
●=F1, ■=F2, ▲=F3.
--- Running median (5 point)
——·—— -6dB/octave slope.

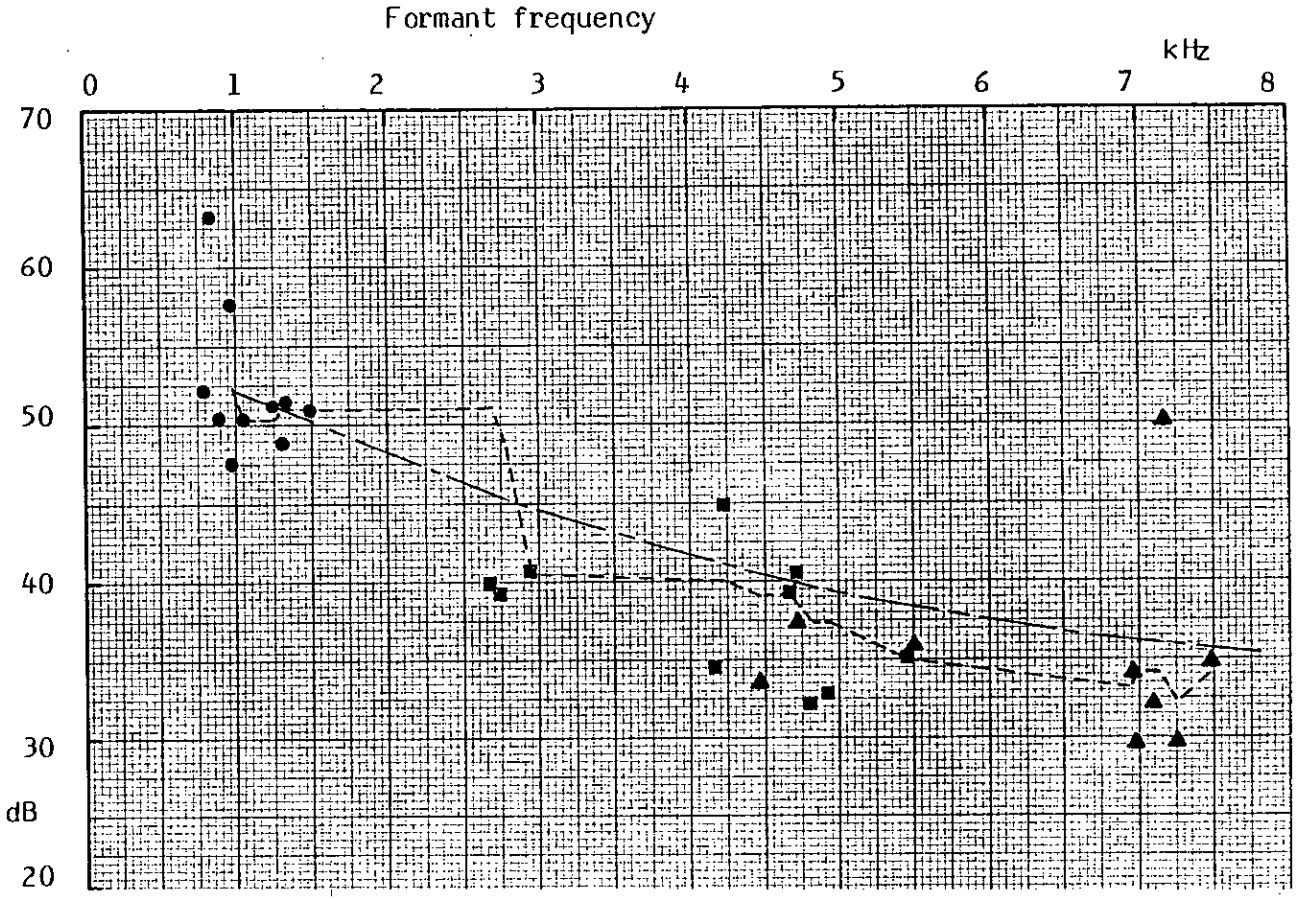
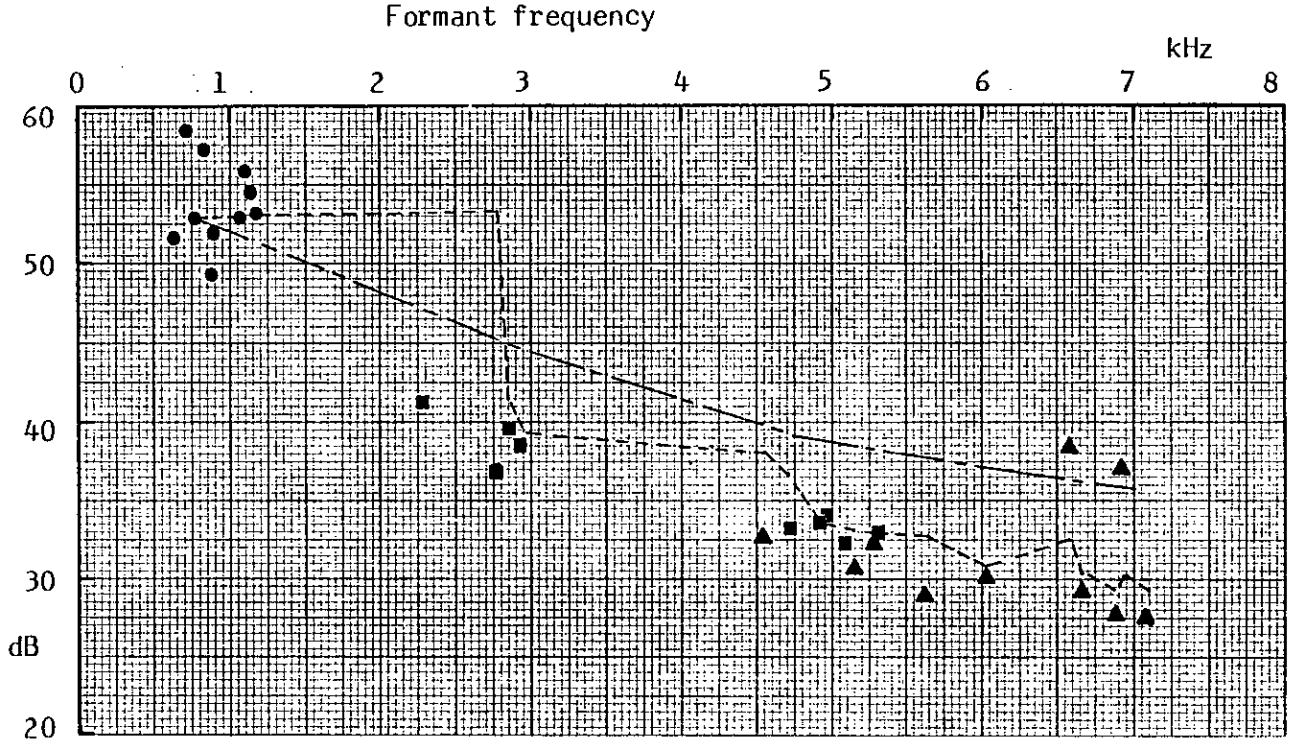


Fig. 3.16(d) Formant frequency v. amplitude at 300ft.
●=F1, ■=F2, ▲=F3.
--- Running median (5 point)
—·— -6dB/octave slope.



Formant amplitude in
heliox at 400ft depth.

Fig. 3.16(e) Formant frequency v. amplitude at 400ft.
●=F1, ■=F2, ▲=F3.
--- Running median (5 point)
—·— -6dB/octave slope.

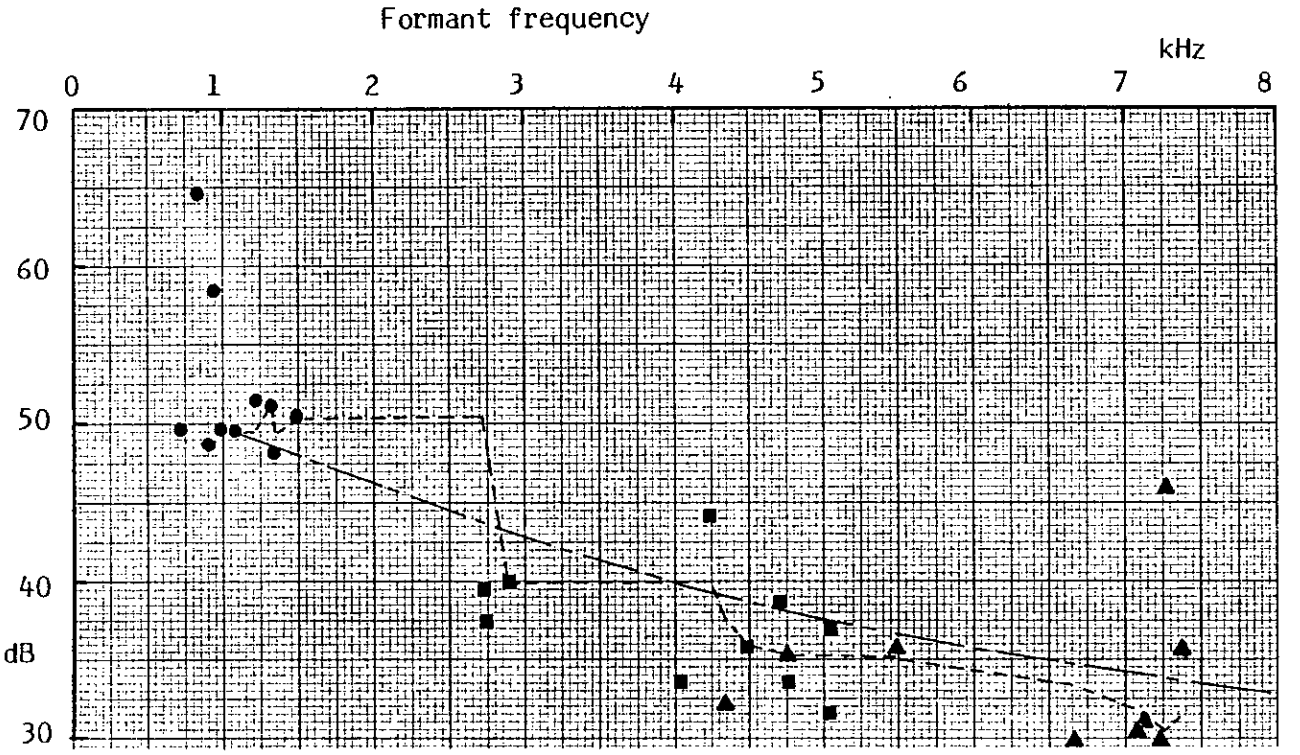
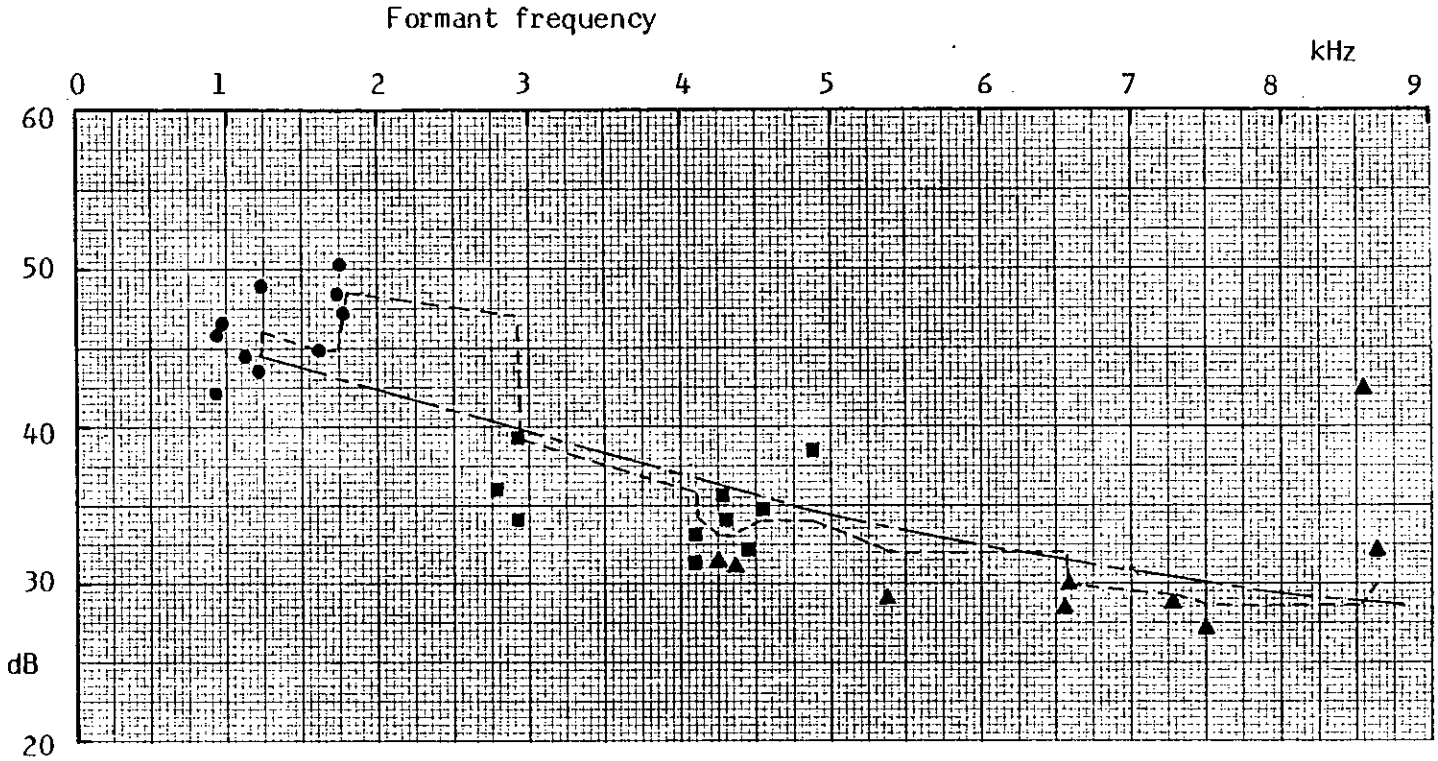


Fig. 3.16(f) Formant frequency v. amplitude at 450ft.
●=F1, ■=F2, ▲=F3.
--- Running median (5 point)
—·— -6dB/octave slope.



Formant amplitude in
heliox at 500ft depth.

Fig. 3.16(g) Formant frequency v. amplitude at 500ft.
●=F1, ■=F2, ▲=F3.
--- Running median (5 point)
——— -6dB/octave slope.

-6dB/octave is shown by the chained line on each graph, starting from the first output point of the median filter.

The median characteristic in each graph indicates a steady decrease in formant amplitude with increasing frequency. Note that the important feature in this respect is not the absolute values of formant amplitude from depth to depth, but rather the relative slope of the amplitude decay. Absolute formant amplitude reflects the overall sensitivity of the recording microphone frequency response in a particular environment, and also the position of the microphone with respect to the subject. As regards the relative amplitude decay, it can be seen that in the main, the data at each depth, including the formant amplitude data from recordings on the surface in air, corresponds well to a slope decay of around -6dB/octave.

3.4.6 FORMANT BANDWIDTH CHARACTERISTICS IN HELIOX

The collective bandwidth data for all vowels pooled together at each depth, including the data in air at the surface, is shown in Fig. 3.17(a-g). The data here is interpreted in terms of formant Q-factor, given by:

$$Q = \frac{\text{formant centre frequency}}{\text{formant bandwidth}} \qquad (3\text{-}21)$$

Once again, a 5-point running median filter has been applied in an effort to expose any significant trends in the data, and the filter output is represented in each graph of Fig. 3.17 by the dotted line.

Scrutiny of the data in Fig. 3.17(a) relating to air speech shows that, in addition to a fairly large spread of data, the underlying trend is a fairly low Q for first formant data, gradually increasing towards the higher-frequency second formants, and then a drop in

continued on pg.124

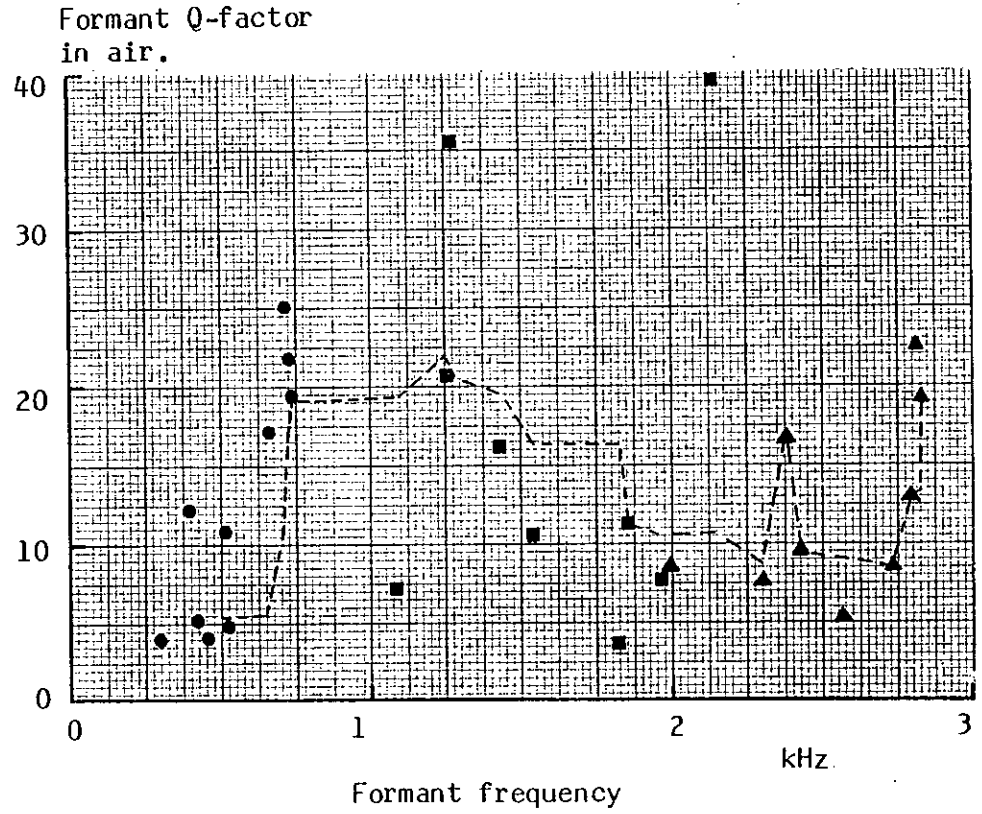
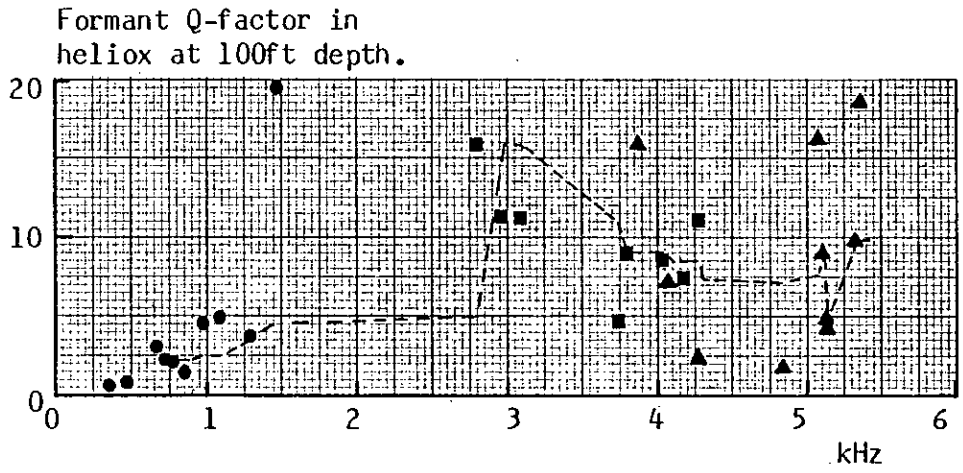
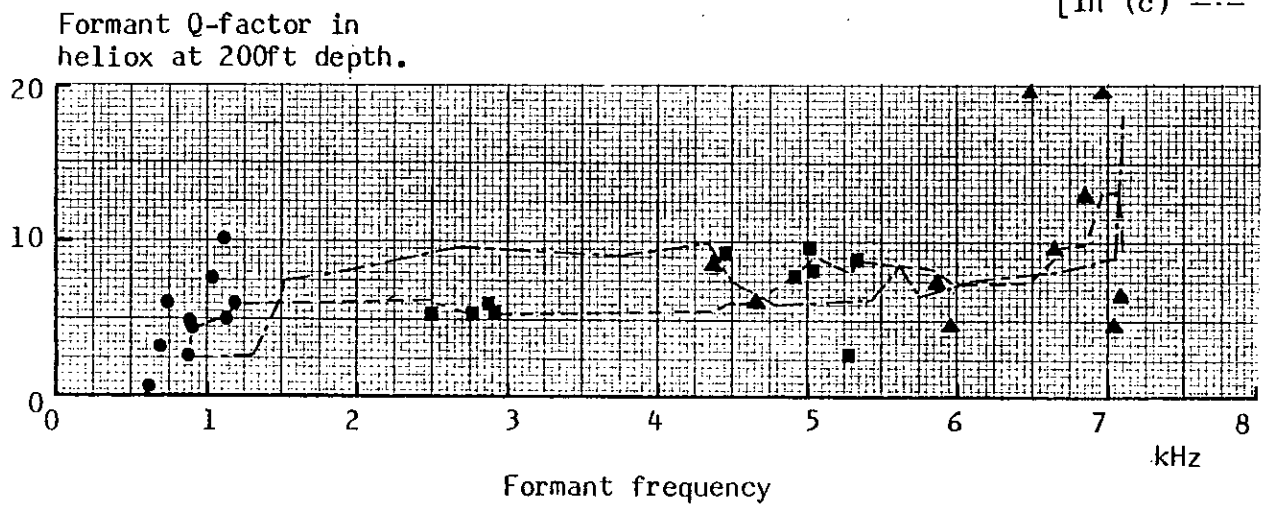


Fig. 3.17(a) Formant frequency v. Q-factor in air.
=F1, =F2, =F3.
--- Running median (5 point)



(b)



(c)

Fig. 3.17(b-c) Formant frequency v. Q-factor at (b) 100ft and (c) 200ft.
●=F1, ■=F2, ▲=F3.
--- Running median (5 point)
[In (c) —·— Warped median curve from Fig. 3.17(a).]

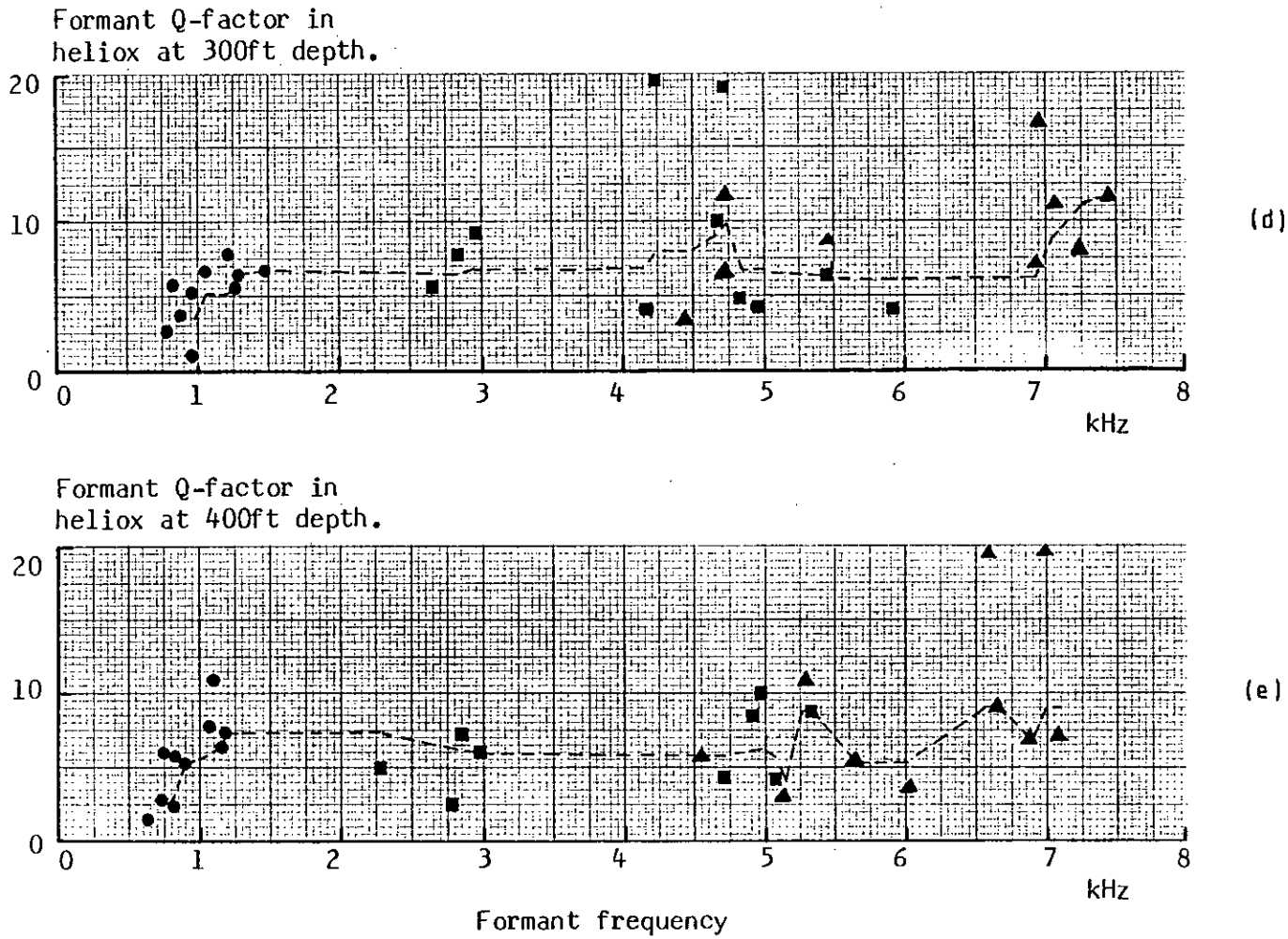


Fig. 3.17(d-e) Formant frequency v. Q-factor at (d) 300ft and (e) 400ft.
●=F1, ■=F2, ▲=F3.
--- Running median (5 point)

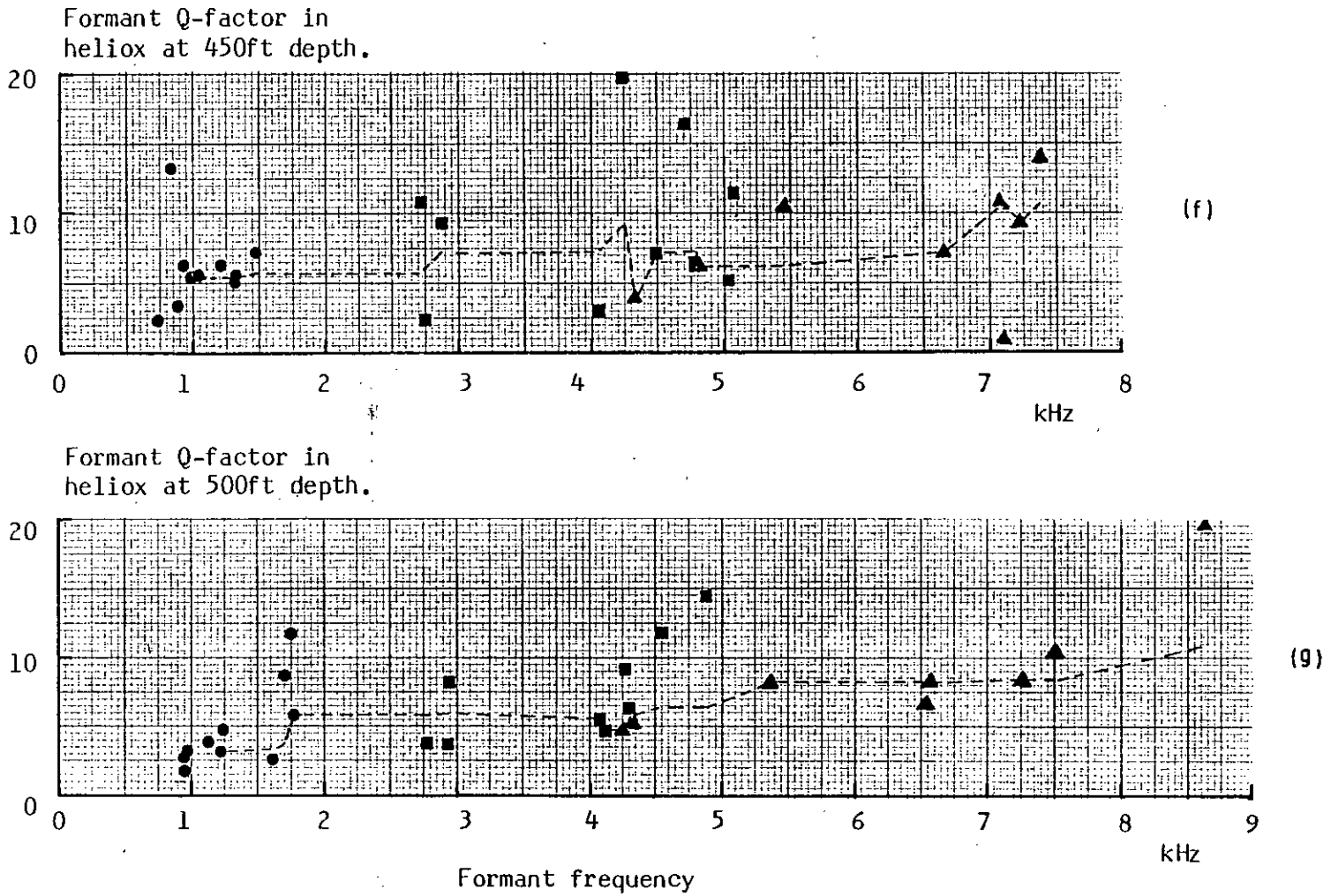


Fig. 3.17(f-g) Formant frequency v. Q-factor at (f) 450ft and (g) 500ft.
●=F1, ■=F2, ▲=F3.
--- Running median (5 point)

Q-factor for third formant data, which correlates well with the established bandwidth data for air speech [146,147].

A comparison of the characteristics in the various heliox mixtures of Fig. 3.17(b-g) with Fig.3.17(a) demonstrates that there are significant differences in the variation of Q-factor with relative F1, F2 and F3 formant frequencies at depth as compared to the trends in air. The variation of formant Q-factor about the median is much reduced at depth, and a comparison of the median characteristics themselves shows that the Q-factors of the F1 and low-frequency F2 formants are lowered compared to their values in air, indicating an increase in the bandwidth of low-frequency formants. Note however, that the spread and median values of the Q-factor data relating to the high-frequency F3 formant regions remains roughly similar from depth to depth.

3.5 AN ANALYSIS OF AMBIENT NOISE AND ITS EFFECTS ON THE AR SPECTRAL ESTIMATE

Listening to the recordings of helium speech used in the analysis so far, it was evident that there appeared to be an important level of noise contaminating the speech signal under analysis. The source of this noise was a carbon dioxide "scrubber" located within the diving chamber and driven through magnetic coupling by a motor external to the chamber. This device is basically a gas filter which extracts otherwise lethal carbon dioxide from the respiratory atmosphere by chemical means. The noise itself is due to cleansed heliox being forced through a narrow tube by a magnetically-driven pump.

It is important to establish the characteristics of the noise source

and their likely effects on the foregoing acoustic analyses. Although presumably an important problem in helium speech, an extensive survey of the available literature has revealed no mention of this topic. Therefore although they are hopefully complete, it has not been possible to corroborate the results presented here.

Lastly, since the homomorphic pitch evaluation results have already been confirmed in section 3.2.1, the discussion of the noise characteristics and their effects is confined to the AR spectral analysis technique.

3.5.1 NOISE SOURCE CHARACTERISTICS

From each reading at each depth, it was possible to isolate and digitise some 2 to 3 seconds of continuous material consisting solely of ambient noise from within the diving chamber. The starting point here has been to determine the statistical distribution of the noise source, sampled at 20kHz, at each depth. The same characteristic distribution was found for each depth, and an example is shown in Fig. 3.18, where the x-axis represents the digitised signal level (with possible values from -2047 to +2047) and the y-axis being the number of occurrences of each level over the analysis period, which in this case is of the order of 2 seconds. The distribution is well-described by a Gaussian approximation as shown by the solid curve through the data, calculated from a knowledge of the standard deviation and assuming the mean $\mu = 0$.

This is not sufficient in itself to ensure a "white" noise spectrum from instant to instant, since the noise may itself be "coloured" over short periods of time. To explore this further, the AR process was

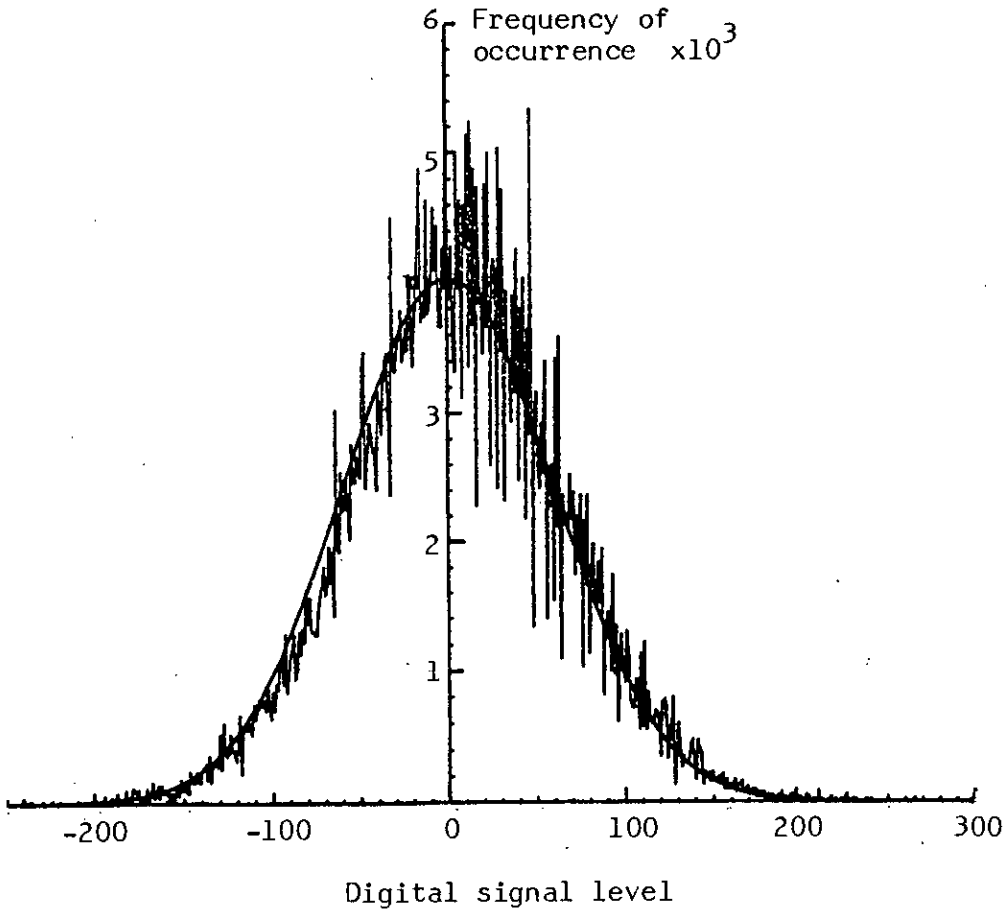


Fig. 3.18 Statistical distribution of signal level of 2 seconds of digitised scrubber noise at 400ft.
—— Gaussian distribution assuming mean = 0.

applied to the noise source, over a period of 1 second, with the same analysis conditions as those used in the voiced speech analysis, that is, N=256, 64-point roll-along, 14-pole fit and FFT resolution =20Hz. This produced a series of some 300 or so spectral frames, and, when aligned together, it was found that there were no consistent areas of spectral energy maxima which endured over more than three to four consecutive frames, and there appeared to be no long-term persistent features whatsoever. This presents a fair justification for assuming

that, over the total analysis interval of 38.4mS for voiced speech in heliox at depth (see section 3.4.1), the noise source can indeed be assumed to be Gaussian in nature with a white spectrum, thereby facilitating a means of estimating the true signal to noise ratio (S/N) of each vowel segment analysed, as follows.

Consider the energy $\mathcal{E}(n)$ of a composite waveform, consisting of a signal S(n) with additive Gaussian noise N(n), at some instant n:

$$\mathcal{E}(n) = (S(n) + N(n))^2 \qquad (3\text{-}22)$$

and the power P over a sample of m points is:

$$P = \sum_m \mathcal{E}(n) = \sum_m (S(n) + N(n))^2 \qquad (3\text{-}23a)$$

$$= \sum_m S^2(n) + \sum_m N^2(n) + 2\sum_m S(n)N(n) \qquad (3\text{-}23b)$$

$$= S + N \text{ , where S=signal power, N=noise power.} \qquad (3\text{-}23c)$$

That is, for a true Gaussian noise distribution, then the last term of equ.(3-23b) sums approximately to zero. Thus, if samples of signal-plus-noise and noise-only are available at each depth, the signal to noise ratio can be estimated:

$$\text{let} \quad x = P/N = (S + N)/N,$$

$$\text{then} \quad S/N = x-1 \qquad (3\text{-}24)$$

Using this simple relationship, the S/N ratios for each vowel section analysed at each depth have been estimated, and are tabulated in Fig. 3.19.

This method also permits an estimation of variation of noise power with time. A short segment of some 38mS of noise at each depth was extracted from the available recordings and the noise power over this period was calculated. A window the same length as the extracted segment was then scanned across the entire noise waveform and the running ratio of windowed noise power to extracted segment power was calculated. For

S/N ratios (dB) of vowel sections analysed

Item	heed	hood	hid	head	had	hard	hud	hod	who'd	herd
Depth (ft)										
0	15.2	9.3	12.6	13.0	-0.6	9.8	14.9	8.9	3.6	-0.4
100	12.5	9.4	11.0	3.3	6.9	10.1	10.2	11.0	10.3	8.3
200	13.2	10.7	9.2	11.4	12.8	10.3	10.6	8.9	13.3	12.7
300	17.1	17.3	17.5	17.7	17.9	18.1	18.3	18.5	18.9	19.1
400	11.1	9.5	9.9	9.4	10.0	9.4	10.6	8.9	12.5	11.0
450	10.5	12.9	9.4	9.3	9.7	9.7	9.0	9.5	10.0	8.0
500	3.5	3.9	3.3	1.4	4.5	3.5	0.0	2.0	1.1	1.5

Fig. 3.19 Table of S/N ratios for each vowel segment analysed.

all depths, the maximum noise power variation was found to be within ±1.5dB. Moreover, this method revealed the variations to be cyclic with a period of some 20mS, which corresponds to the 50Hz cyclic variation of noise intensity, due to the pumping action of the carbon dioxide filter, which can be heard on recording A3.

3.5.2 EFFECT OF NOISE ON THE AR SPECTRAL ESTIMATE

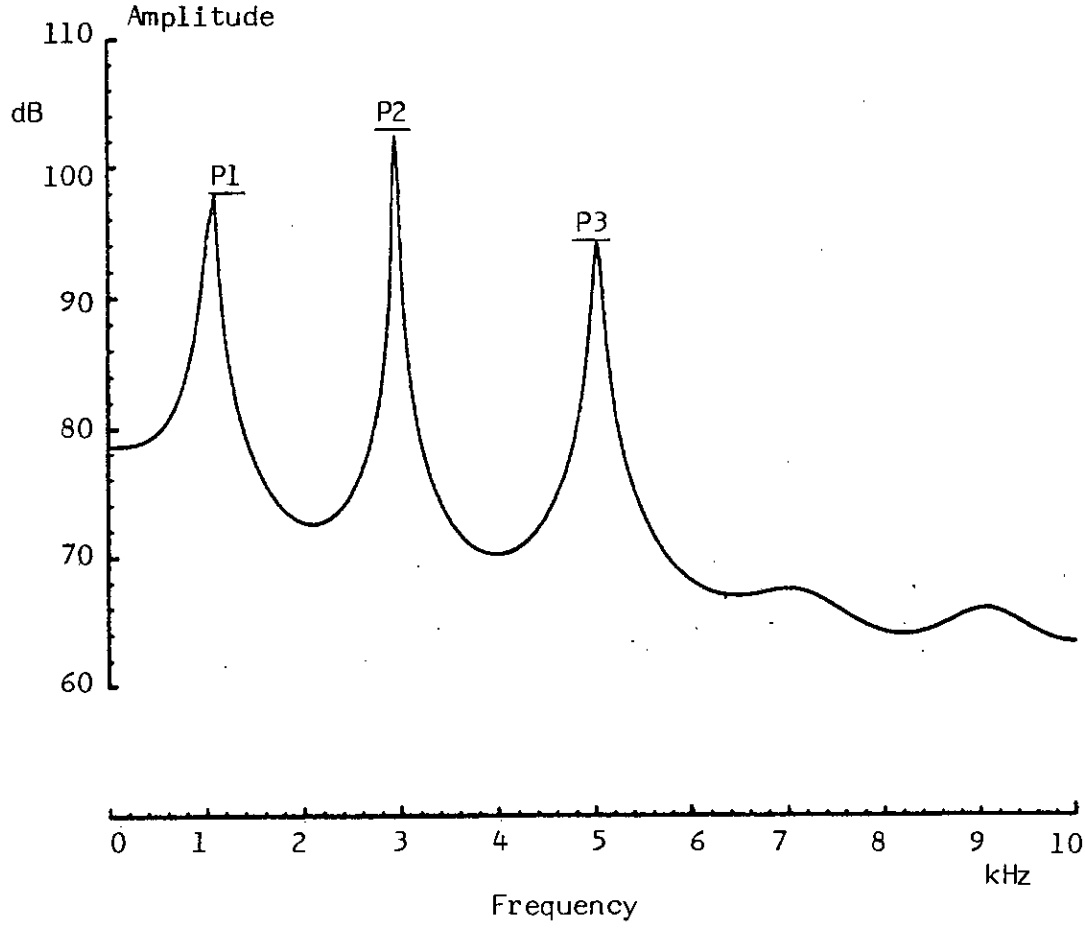
The first approach was to consider the effects of additive Gaussian noise on a signal with well-defined characteristics, and a signal consisting of 3 sinusoids of 1kHz, 3kHz and 5kHz sampled at 20kHz was chosen. Note that these frequency components have theoretically zero bandwidth. In this experiment, noise extracted from the recordings on the surface in air was added to the signal to give S/N ratios between 15dB and -3dB, and AR processing was applied to the resulting new signal-plus-noise commensurate with the parameters outlined in section 3.4.1, with the exception of preemphasis, which was not employed here.

The peak-finding algorithm was, in this case, applied to each of the 9 contiguous spectral frames produced for any given S/N ratio, and values of the spectral attributes of peak centre frequency, amplitude and bandwidth were derived and tracked from frame to frame. It was found, to a first approximation, that the standard deviation, calculated over the 9 frames of each analysis, of all three attributes roughly doubled for every 3dB fall in S/N ratios inferior to 10dB.

Whilst it was possible to visually track each spectral peak from frame to frame and hence provide fair estimates for each attribute by averaging over the 9 frames of each analysis, this was not considered an efficient solution given the amount of speech data to be analysed. A satisfactory method of providing data on each spectral attribute was derived by averaging the 9 spectra to produce one representative spectrum only, and moreover is capable of inclusion in the processing as an automatic function. Indeed, for precisely the same reasoning, this method is also useful in averaging the momentary fluctuations in formant attributes which necessarily occur in human speech, see section 3.4.1.

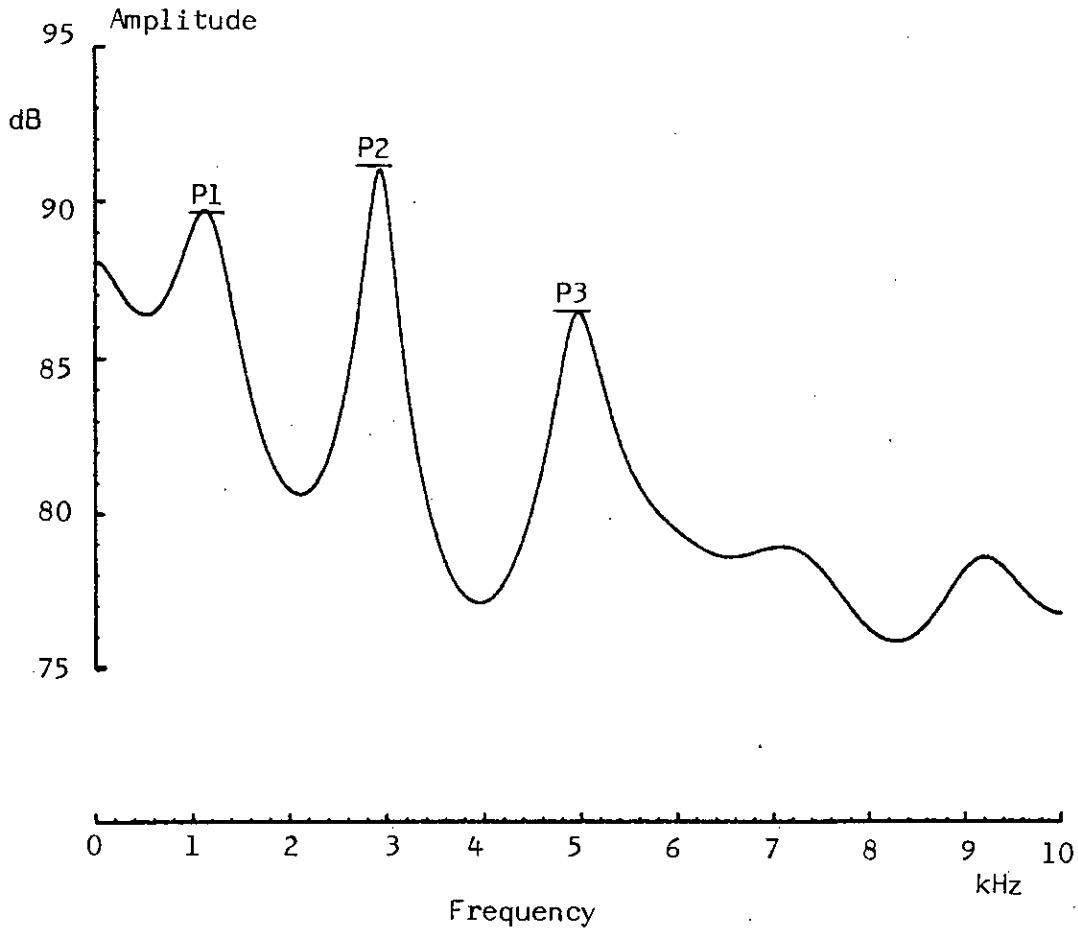
Having established this spectral frame averaging technique, the data relating to the sinusoids-plus-noise was reexamined. From comparisons of the data from each average spectrum generated, it was observed that, below a S/N ratio of 10dB (a) peak frequencies increased at a rate of roughly 20-30Hz per 3dB fall in S/N ratio; (b) all peak amplitudes fell at a rate of roughly -4dB per 3dB fall in S/N ratio; (c) there was an overall increase in peak bandwidth but not monotonically with increasing S/N ratio. The corresponding spectra and data for S/N=10dB and S/N=0dB are shown in Fig. 3.20.

In order to endorse the above findings in relation to the voiced



Peak	P1	P2	P3
Frequency (Hz)	1090	2951	5023
Amplitude (dB)	98.1	102.4	94.1
Bandwidth (Hz)	51	47	67

Fig. 3.20(a) AR spectrum and resonance data for combined 1kHz, 3kHz and 5kHz signals contaminated with scrubber noise, S/N = 10dB, no preemphasis.



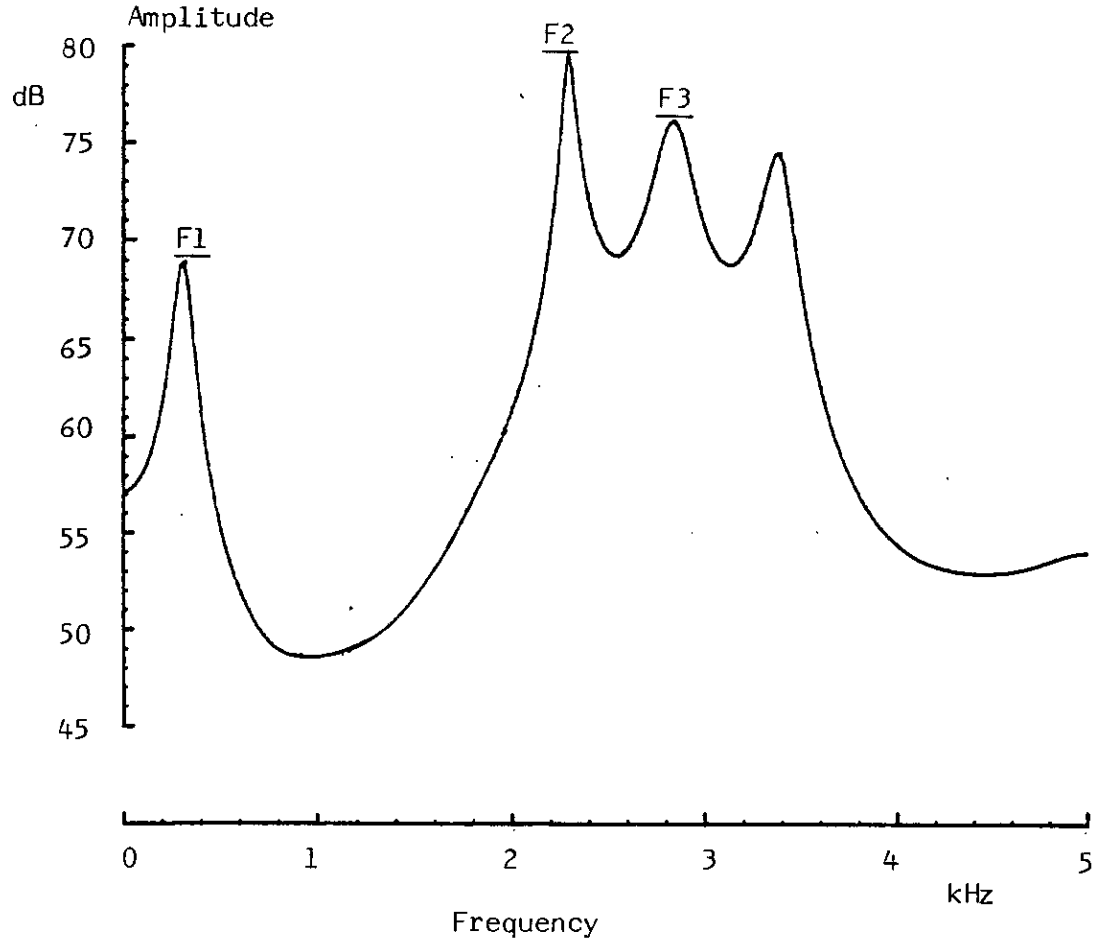
Peak	P1	P2	P3
Frequency (Hz)	1124	2927	4966
Amplitude (dB)	89.7	91.0	86.4
Bandwidth (Hz)	431	207	335

Fig. 3.20(b) AR spectrum and resonance data for combined 1kHz, 3kHz and 5kHz signals contaminated with scrubber noise, S/N = 0dB, no preemphasis.

speech analysis, the "ee" vowel section spoken in air at the surface (S/N ratio > 15dB, see Fig. 3.19) was taken and contaminated with progressively more intense noise in a manner similar to the preceding investigation, down to a minimum of S/N=0dB. Note that preemphasis was, in this case, applied to the data prior to processing. The following trends were observed: (a) formant frequency increased by some 20-30Hz per 3dB fall in S/N ratio; (b) formant amplitude remained relatively unaffected over the range of S/N ratios; (c) there was an overall increase in formant bandwidth but not monotonically with decreasing S/N ratio; second and third formant bandwidths increased by a factor of some 2-3x more than for the first formant. It was also found that at S/N ratios inferior to 6dB, the upper frequency spectrum in the range beyond 2kHz became progressively more deformed by the appearance of new 'formant' peaks, making selection of true formant peaks very difficult, which can be considered to be the limiting factor for visual identification of individual formants. The corresponding spectra and data are shown in Fig. 3.21.

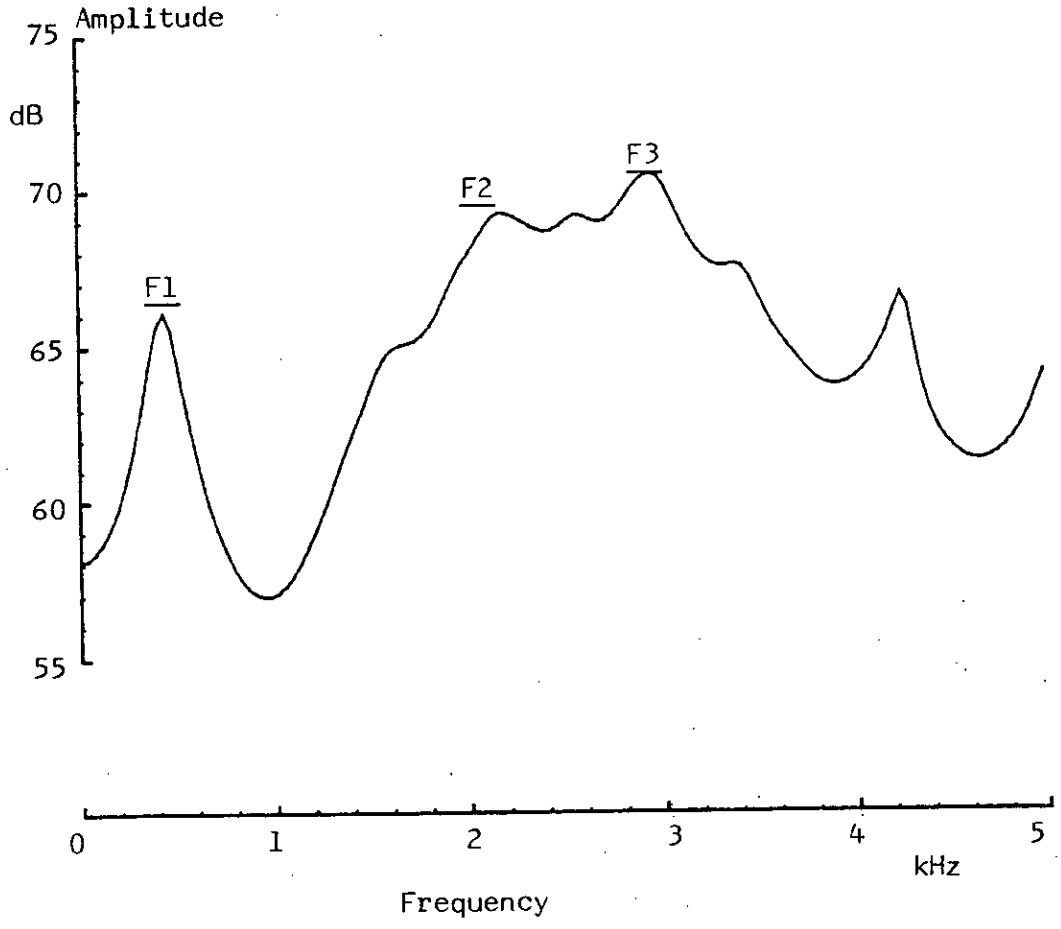
The differences regarding the observation of formant amplitude can be explained in terms of the proximity of the sinusoid z-plane poles to the unit circle, $|z|=1$, which promotes a sensitivity to any variation in the estimate of pole location due to the presence of noise. Poles which are damped, as in the case of speech, are already located well inside the z-plane unit circle, and consequently the frequency response is less sensitive to variations in pole location.

As regards the larger variation with frequency of the high frequency formants in speech contaminated with noise, this can be correlated with the effects of local S/N ratio relating to the formant amplitude slope characteristic and spectral properties of white noise.



Formant	F1	F2	F3
Frequency (Hz)	306	2288	2829
Amplitude (dB)	69	79.7	76.2
Bandwidth (Hz)	73	54	148

Fig. 3.21(a) AR spectrum and formant data for the vowel "ee" in "heed". Original vowel segment as recorded, S/N> 15dB. (Preemphasis factor μ = 0.986).



Formant	F1	F2	F3
Frequency (Hz)	429	2167	2932
Amplitude (dB)	66.2	69.2	70.5
Bandwidth (Hz)	161	371	350

Fig. 3.21(b) AR spectrum and formant data for the vowel "ee" in "heed", contaminated with additional scrubber noise, S/N = 0dB. (Preemphasis factor μ = 0.986).

In section 3.4.5 it was shown that formant amplitude decays with increasing frequency at the rate of -6dB/octave. Assuming the spectral amplitude of the scrubber noise to be flat, then the effective S/N ratio of any frequency components in the speech waveform will worsen progressively with increasing frequency at approximately -6dB/octave. With reference to the experiments with sinusoids, it was seen that a worsening S/N ratio produced the effect of a wider bandwidth in the (averaged) spectral peak, and therefore a wider formant bandwidth with increasing frequency is to be expected in speech contaminated with additive white noise. This is an important point which will be reiterated in the conclusions which follow. Finally, note that the application of preemphasis is not expected to worsen the local S/N ratio, since both signal and noise powers are elevated by the same amount.

3.6 CONCLUSIONS

In terms of fundamental frequency in a heliox environment, the results of section 3.2.2 demonstrate that neither percent helium content nor respiratory gas pressure appear to have any effect upon the fundamental frequency, thereby validating the use of signal processing strategies which preserve the pitch period of helium speech when processed by the unscrambler system. However, the results indicate that pitch is likely to rise due to environmental conditions not always under the control of a supervising authority, and particularly under conditions of psychological stress. This is important particularly where women divers are concerned, since their average fundamental frequency, which under normal circumstances is of the order of 200Hz

but ranging in the short term from 75-450Hz, tends to be already higher than that of men due to the smaller physical dimensions of their larynx [148]: all of the existing time-domain based unscrambler systems[111] depend upon the successful detection of the start of a pitch period for their operation; on triggering, a segment of the speech waveform, generally of the order of 3mS long, is captured from the start of the pitch period, and the remaining part of the waveform is discarded. The captured segment is subsequently read out from its storage buffer at a reduced speed compared to the read-in speed, thereby compressing spectral information in the inter-pitch waveform, but still maintaining pitch period. If however, the pitch period is less than the designated capture time of pitch segments as designed into the unscrambler device, i.e. at fundamental frequencies above 330Hz, then the pitch period itself will be subject to erratic expansion in time, thereby degrading the performance of the unscrambler device.

As regards the formant frequency shift characteristic in heliox, the results of the piecewise linear approximations as outlined in section 3.4.2 and Fig. 3.13 ostensibly support the classical nonlinear formant shift contour of equ.(2-15) from a consideration of its rate of change of slope at low frequencies. However, if the proposed relationship of equ.(2-15) were to hold true in this instance, then the formant shift ratios associated with the data in this experiment should demonstrate a preponderance of formant shift ratios for F1 which are greater than those for F2 or F3 due to the effect of vocal tract wall resonance. As can be seen from Fig. 3.15(k), however, the majority of F1 shift ratios are less than those for F2 or F3, which implies that the l.m.s curves of equ.(3-20) are in fact of more significance than the nonlinear curve shown fitted to the data. Thus, whilst concurring

with the classical assumption that, in general, the nonlinearities of the formant shift ratio are confined to the low frequencies [74], they are the antithesis of the classical theory relating to the accompanying formant shift ratios, which have been shown here to be less than those of higher frequencies. Note, however, that both slopes associated with each set of l.m.s curves are significantly less than the corresponding linear shift ratios predicted in Fig. 3.3.

The tacit assumption in this analysis is that individual formants have been faithfully identified from depth to depth. It is likewise assumed that the spectral averaging technique discussed in section 3.5.2 provides reliable estimates of formant frequency. From section 3.5.2 this has been shown to be the case for those vowels analysed with a S/N ratio of better than 6dB, which accounts for 80% of the data (see Fig. 3.19): there is also an improved confidence in the correct visual identification of true formant peaks from depth to depth, in that there is less clutter of the spectrum due to extraneous formant-like detail which appears as a consequence of impaired S/N ratio. In order to explain the low values of formant shift ratio, it was considered that perhaps there was some gross error in the constituent gas percentages (Fig. 3.3) logged during the recordings. However, it is the stated policy of the company, whose good services were offered in the making of these recordings of helium speech, to maintain a partial pressure of oxygen (PP02) of the order of 0.4bar in the respiratory environment of its diving chambers. Calculations using the data of Fig. 3.3 yield values for the PP02 at depth to within 10%, at worst, of this figure. Therefore, these results imply that the diver is probably adapting his own speech, presumably to render it more intelligible as he himself judges.

This assertion may be reinforced by the formant profiles of Fig. 3.14(a-j), which demonstrate that each vowel is shifted differently from another but in a consistent manner from depth to depth. As discussed in section 3.4.3, these profiles may be the result of (a) articulatory-dependent adaptation to the spectral shift of the speech spectrum in heliox and (b) adaptation of speech to minimise effects due to the heliox atmosphere on the auditory feedback mechanism. The two cases imply different processing strategies for a helium speech unscrambler system. In the case of (a), correct unscrambling of each phoneme would require speech recognition in order to identify each phoneme and warp its frequency content individually so as to improve intelligibility, whereas case (b) implies that it may be possible to linearise the formant frequency shift prior to processing. Since the speed of sound in heliox is greater than that in air, then the speech waveform propagating through heliox will arrive at the ear in advance of the same waveform were it travelling through an air medium, and may induce the diver to modify his speech to compensate for the effect of this early arrival upon his own perceptual system. If this is so, then simply delaying the arrival time of the signal at the ear may remove phoneme-specific formant profiles and hence simplify subsequent processing.

Evidence has been presented, in section 3.4.4, which demonstrates that new formants are introduced into the helium speech spectrum at depth. The translation of these unidentifiable formants back to their values in air is tabulated in Fig. 3.15(ℓ), and suggests that they may be produced as a result of actual nasal resonance. Those formants which translate back to frequencies in air of around 1kHz and 2kHz correspond well to formant locations in normal air speech known to be directly

due to acoustic coupling of the nasal cavity [37,38], although those formants around 1.5kHz cannot be so readily explained. The mechanism for this nasal coupling may be through the tissue of the soft palate as has already been suggested [80]. Another possibility, may be that the velum (Fig. 2.1) is unconsciously not held close enough, by muscular action, to the pharynx wall, thereby permitting nasal resonance.

As explained in section 2.1.2, nasal resonance depends normally on the ratio of the nasal port area to the velo-pharyngeal opening. The high pressure and density of the respiratory atmosphere is known to create respiratory difficulties such as dyspnea (see section 1.2), and it may be that the subject is adopting unusual velar postures in order to aid breathing, which will affect the area of the velopharyngeal opening in the vocal tract and therefore may promote nasal resonance.

Once again, these results imply an unscrambler processing strategy based on speech recognition to identify those vowels in which unidentified formants occur, since if they are the result of nasal coupling, the frequency location and amplitude of these formants, together with the phonemes within which they exist, will vary both with respiratory mixture and from talker to talker, depending on the physical dimensions of the nasal chambers and other physiological factors [10,145,149].

Whilst the above-described phenomena of phoneme-specific formant shift profiles and unidentifiable formants may contribute to the overall degradation of helium speech intelligibility in respect of the listener, they have been shown to be specific only to certain phonemes and can be expected to be manifest in different ways in different talkers. Since the results presented here relating to these phenomena are unique both in the sense that they have not been explored and therefore corroborated by either past or contemporary research and

that they relate to one subject only, then they will not be cited further in this thesis as regards the design of an eventual unscrambler system, except to say that they support the implementation of a system based on frequency domain processing.

The results relating to Fig. 3.16(a-g) show that formant amplitude will be attenuated in a heliox environment. This can be deduced since the characteristic amplitude roll-off in both air and heliox is approximately -6dB/octave, implying that glottal excitation source characteristics can be assumed invariant, therefore the attenuation of formants is due solely to their shift along this decay slope to higher frequency locations in heliox. Thus, spectral amplitude correction must necessarily be included in helium speech unscrambler systems.

Results from section 3.4.6 indicate that certain formant Q-factors vary with increasing depth. Specifically, in a heliox environment, although for the high-frequency formants their bandwidth increasesby the same proportion as formant frequency, that is, Q-factor is relatively unaffected, the bandwidth of the low-frequency formants, on the other hand, increases by an amount greater than the corresponding formant frequency shift ratio, thereby reducing Q-factor. Results of this nature have already been observed [91], and the empirically derived relationship from the observed data suggested that the increase in bandwidth was of the order of R^2 at low frequencies decreasing to R at high frequencies, where R is the velocity of sound ratio as defined by equ.(2-8). As an example, the median characteristic of Fig. 3.17(a) relating to formant data in air has been warped to accomodate the above postulate and is displayed in Fig. 3.17(c) as the chained line, allowing comparison against the actual median characteristic for 200ft in Fig. 3.17(c).

Let the lowest and highest formant frequencies of the median curve of Fig. 3.17(a) be $fm\ell$ and fmh respectively, with corresponding intermediate frequencies denoted by fmi, and let the lowest and highest formant frequencies of the actual median curve of Fig. 3.17(c) at 200ft be $fh\ell$ and fhh respectively. In order to produce a warp in frequency to fit the median of Fig. 3.17(a) into the frequency space of Fig. 3.17(c), then the frequency fhi in Fig. 3.17(c) which will correspond to fmi of Fig. 3.17(a) is given by:

$$fhi = fh\ell + Xr.(fhh - fh\ell) \qquad (3\text{-}25)$$

where $$Xr = (fmi - fm\ell)/(fmh - fm\ell) \qquad fm\ell \leqslant fmi \leqslant fmh \qquad (3\text{-}26)$$

Let Qhi be the required warped Q-factor at frequency fhi, and let Qmi be the Q-factor at frequency fmi in air. In order to produce a warped Q-factor characteristic corresponding to a translation in bandwidth which varies from R^2 at low frequencies to R at high frequency, then Qhi is given by:

$$Qhi = \frac{R}{(1-Xr)R^2 + Xr.R} \times Qmi \qquad (3\text{-}27)$$

A comparison of the warped and actual median characteristics of Fig. 3.17(c) shows that there is a rough agreement in the trends of both curves, producing lower Q-factors for first and second formants comparedwith Q-factor in air. Similar results are obtained at other depths, thereby verifying that low frequency formant bandwidths increase more than high frequency formant bandwidths in a heliox atmosphere. However, from the analysis of the effects of noise presented in section 3.5.2, it is seen that the effect of progressively decreasing S/N ratio, due to the formant amplitude decay with frequency of voiced speech, is to increase the bandwidth estimate, and in particular for high frequency formants. It is therefore probable that the Q-factors of formant data

in Fig. 3.17(a-g) are in fact higher than shown. In the sense however that the main point in question relates to a comparative study of Q-factor characteristic and is therefore less dependent on absolute values, then the results presented here relating to Q-factor variation with depth may indeed be valid. On the other hand, the effect of S/N ratio may be so detrimental to true bandwidth estimation that the comparison relates only to the effects of a worsening S/N ratio with depth, so increasing the variation of formant frequency location from analysis frame to frame, giving the impression of increased bandwidth in the resulting average spectrum, so effectively masking any underlying trends in the actual value of formant bandwidth. It is felt, however, that for those vowels analysed having a S/N ratio $\geqslant$ 10dB, which accounts for some 51% of the data, there is a fair degree of confidence in the derived Q-factors. Additionally, from a consideration of the effects of preemphasis on the spectral estimate of a signal with low S/N ratio, it was seen that high frequency bandwidth estimates were worst affected. The results here, however, demonstrate that it is low frequency bandwidths which increase most, an effect which cannot be directly related to the spectral properties of a signal with low S/N ratio, and so it is proposed that the results, when considered in terms of the median fit through the data, are a reasonable reflection of the true situation.

In summary, the results taken as a whole imply that the provision of a nonlinear helium speech processing system is indispensable. It has been demonstrated that, although there is pitch variation in a heliox atmosphere, this can be attributed to psycho-acoustic effects of the environment and is not directly dependent on the physical effects of the respiratory gas upon the vocal tract, therefore any unscrambler system must conserve the fundamental frequency of helium speech. The

results relating to formant frequency indicate that, whilst there is a case for supposing that vocal tract wall vibration is evidenced in a heliox environment, the overall formant shift criterion can best be approximated by piecewise-linear frequency correction as opposed to a nonlinear correction characteristic of the form of equ.(2-15). Additionally, results relating to formant Q-factor show that there is a requirement to adjust formant bandwidth such that low frequency formant bandwidth is reduced by a factor =R x greater than the correction for high frequency formant data. Lastly, note that any unscrambler system must include provisions for the correction of formant amplitude attenuation, which must be carried out before any (nonlinear) formant frequency correction, since the amplitude decay curve in heliox is essentially linear and of the order of -6dB/octave.

CHAPTER IV

INVESTIGATION OF UNSCRAMBLER SYSTEM ARCHITECTURES.

A consideration of the acoustic properties of helium speech, as detailed in Chapter III, has emphasised that the translation of the speech signal from air to a pressured heliox environment is a nonlinear process in which (1) voiced fundamental frequency is conserved together with the spectral characteristics of the glottal excitation source; (2) the speech spectrum is shifted nonlinearly, with high frequencies above 1kHz being shifted by a ratio approaching R, the ratio of the velocity of sound in heliox to that in air, and low frequencies being shifted by a ratio $<R$; (3) there is an attenuation of voiced formant amplitude directly due to the translation in frequency of formants along the glottal source characteristic; (4) formant bandwidth increases by a factor of approximately xR for high frequencies and xR^2 for low frequencies. However, most commercially available helium speech unscrambler devices are constrained to applying linear spectral corrections to helium speech in order to facilitate real time processing, see section 2.3. The most prevalent system of this type in use today, based on time-domain processing [150], is investigated in detail in the first part of this chapter, both in order to identify functional singularities which are likely to affect the intelligibility of the unscrambled speech and also to furnish a basis for a subjective comparison of performance in the simulation of more complex unscrambler systems presented in this and subsequent chapters.

Recent advances in distributed or array processing have allowed the

implementation in real time of a highly complex helium speech unscrambler based on the waveform coding technique of the short-time Fourier transform (STFT) [151,152,153,154,155], permitting the inclusion of nonlinear frequency, amplitude and bandwidth correction [156]. Such a system architecture is studied in detail, and it is shown that while certain features of the helium speech signal can, in theory, be corrected in a nonlinear manner in order to improve intelligibility, there are however certain aspects regarding the synthesis of the spectrally-corrected time waveform, through the inverse Fourier transform, which detract from the enhancement in intelligibility achieved by nonlinear correction.

The studies of the system architectures outlined above and those presented in subsequent chapters have been achieved by software simulation carried out using a Vax 750 digital computer, with the implementation of unscrambler algorithms being in floating-point arithmetic. The signal waveforms used in the system simulations are sampled to 12-bit resolution. The sampling rates are specified for each individual study.

In order to fully utilise the advantages of computer simulation in the processing of helium speech, it has been necessary to develop special software and hardware tools, which have included a graphics package to emulate a dual-trace oscilloscope with full display manipulation, including the ability to both place and move on-screen markers for waveform feature identification. The requirement to provide a means for subjective evaluation of simulated unscrambler system performance has further necessitated the development of hardware to allow real time output of processed speech signals. For logistical reasons pertaining to the provision of both real time output and variability of data sample rates,

the digitised data is first downloaded from the Vax system to a dedicated microprocessor system employing long-term floppy-disk storage, to which is subsequently interfaced the controlling hardware for digital-to-analog conversion. A detailed account of the design of this hardware and supporting software is given in Appendix B.

4.1 HELIUM SPEECH UNSCRAMBLING BY DIRECT PROCESSING ON THE TIME-DOMAIN WAVEFORM

4.1.1 BASIC PRINCIPLES

Helium speech unscramblers in the category of time-domain processing [93,106,107,108] achieve improvements in speech intelligibility by direct time-expansion of the helium speech waveform, in synchronism with each pitch period of voiced speech. This strategy maintains voiced fundamental frequency, but the pitch waveform undergoes an overall linear bandwidth compression by the ratio R, corresponding exactly to the amount of applied time-base expansion, R.

In order to investigate the effects on the power spectrum of this type of processing, consider a stochastic signal in continuous time of the form of Fig. 4.1(a) which is the impulse response of a system characterised by a complex pole pair as shown in Fig. 4.1(b). Expansion of the waveform in time by a factor R produces the waveform of Fig. 4.1(c); with the associated movement of the system poles given in Fig. 4.1(d). The original power spectrum of Fig. 4.2(a), then, is linearly warped by the time-expansion process to produce the corrected spectrum of Fig. 4.2(b), such that formant frequencies and associated bandwidths are all compressed by the factor R, therefore formant Q-factors are

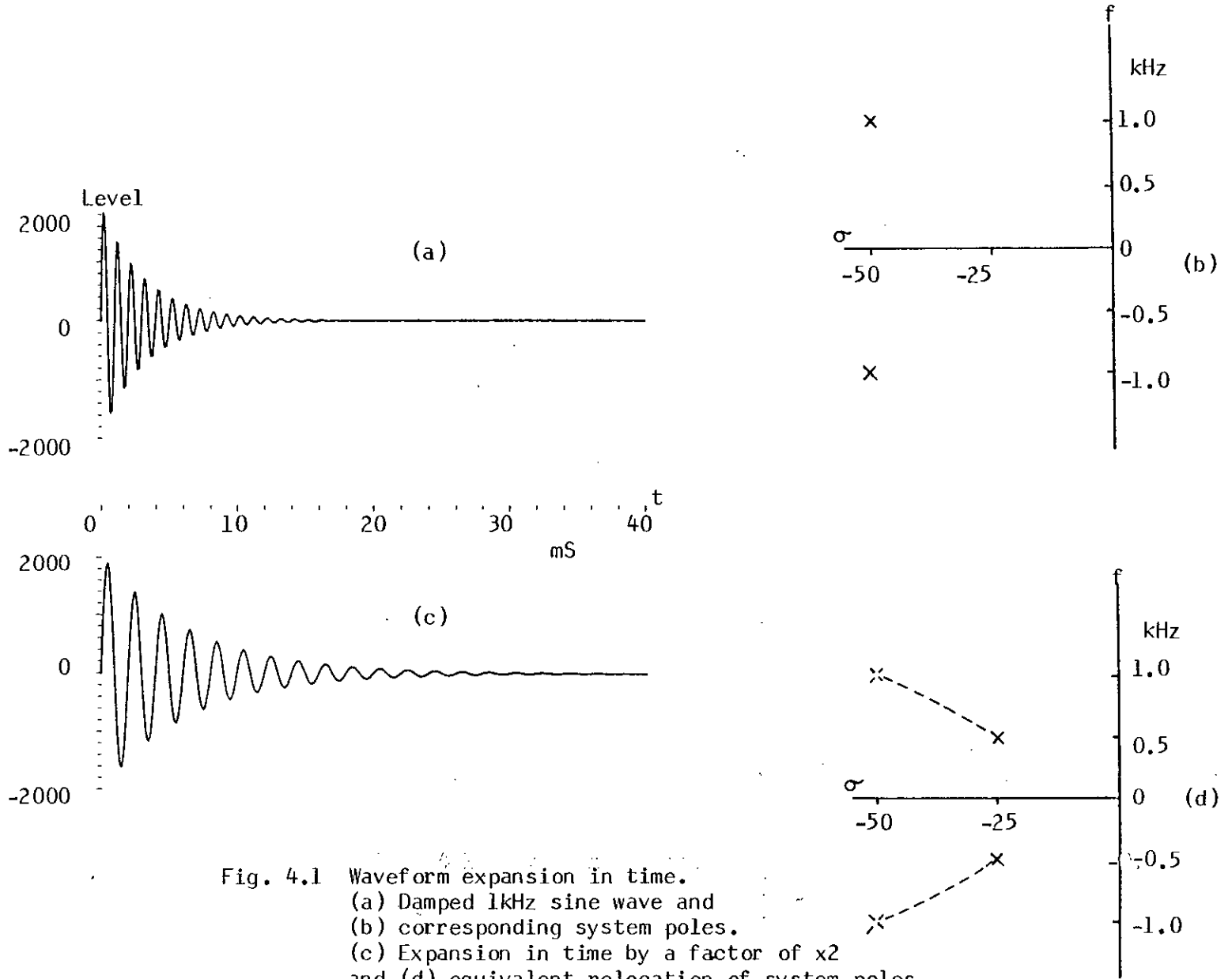


Fig. 4.1 Waveform expansion in time.
(a) Damped 1kHz sine wave and
(b) corresponding system poles.
(c) Expansion in time by a factor of x2
and (d) equivalent relocation of system poles.

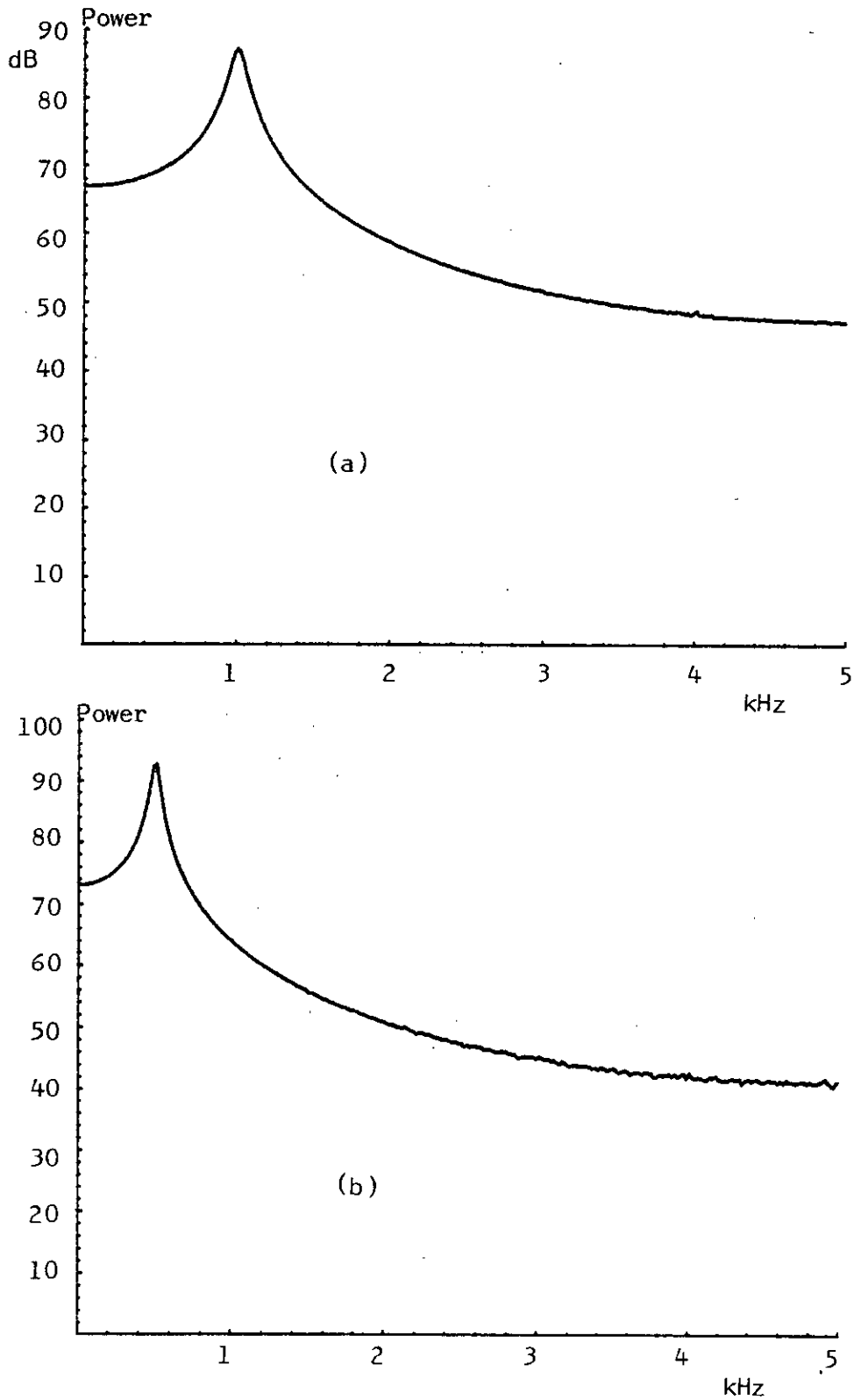


Fig. 4.2 (a) Power spectrum of waveform in Fig. 4.1(a) and (b) spectral compression by time expansion as in Fig. 4.1(c).

conserved. Another effect associated with time-expansion is that the power amplitude of each component frequency in the spectrum has been amplified, and there are two causes to which this amplification can be ascribed. Firstly, from a consideration of the conservation of total spectral power, then a spectral compression by R necessarily entails a spectral power amplification of each component frequency by R. Secondly, it can be shown that there is a further power amplification by the factor R due to expansion of the speech waveform in time but with retention of signal amplitudes, compare Figs. 4.1(a) and 4.1(c).

From the Weiner-Kinchine theorem [138], the power P in any continuous waveform f(t) over time T is given by:

$$P = \int_T f^2(t)dt = \int_F S(f)dt \qquad (4\text{-}1)$$

where S(f) is the power spectral density function.

Time expansion by a factor R but with retention of amplitude such that f(t/R) = f(t) gives a new signal power, P_n, as:

$$P_n = \int_{T.R} f^2(t/R)dt \qquad (4\text{-}2a)$$

and, with m= t/R,

$$P_n = R\int_T f^2(m)dm = R.P = R\int_F S(f)df \qquad (4\text{-}2b)$$

Thus, spectral compression (time expansion) together with retention of waveform amplitudes gives a total amplification of R^2 for corresponding spectral components between corrected and original spectra. However, since this amplification is a constant dependent only on the factor R and is independent of spectral response and waveform shape, then the original relative amplitudes of spectral formant features are conserved and no one speech waveform segment is amplified differentially from another.

4.1.2 SYSTEM ARCHITECTURE

Assume that the helium speech unscrambler system based on time expansion is arranged such that the signal is sampled digitally at some rate f_s, and that as regards time expansion itself, there is firstly a means to detect the point on the waveform from which the subsequent waveform segment is to be expanded, which will normally be from the start of the voiced pitch period. Secondly, let the signal for expansion be stored first in a storage channel and then subsequently read out at a slower rate to achieve effective time expansion of the waveform. This implies that in addition to providing a pitch detector [126,157] to provide pitch-synchronous waveform expansion, there may also be a requirement to provide a very large number, M, of storage channels since, if voiced fundamental frequency were considered capable of varying instantaneously over a large frequency range, then any number of pitch periods may occur during the time to expand any single pitch period. From the same consideration, if all the pitch waveform were to be expanded in time, then each channel would require a large amount N of variable storage space. The resulting architecture is shown in Fig. 4.3(a). However, assuming that the pitch period is approximately constant over long periods of time, then the number of storage channels required is fixed at $M = R_{max} + 1$, where R_{max} is the maximum possible integer time expansion ratio to be encountered. This is since for a constant pitch period, there may be R_{max} possible pitch periods detected in the time required to expand any existing stored waveform. In the case of helium speech, then assuming a worst-case frequency expansion of $R_{max} \doteq 3$ due to a 100% helium atmosphere, (see Fig. 3.3) then $M = 4$. However, an infinite amount of variable channel storage space N would cause the resulting time-expanded waveform to be discontinuous in places. In the

above architecture, each channel does not terminate storage of the pitch waveform until the detection of the start of each subsequent pitch period. In the extreme case of unvoiced speech occuring between voiced speech segments, this would mean that one storage channel would acquire the unvoiced speech signal until such times as a pitch pulse was detected, and there are two possible problems which this may entail: (a) if all other channels are empty, there would be a prolonged period of silence until the arrival of the first pitch pulse, at which time the unvoiced speech waveform would commence time expansion and (b) other channels would subsequently be expanding pitch waveforms contemporaneously with the unvoiced speech waveform. Thus, requirements for continuity of the speech signal demand a regular output of time-expanded speech from the unscrambler device, which therefore limits the amount of storage space N in each channel store to vary up to some maximum length. Thus implies that, for long-duration inter-pitch waveforms, only a certain maximum portion of the waveform can be stored, that is, the trailing portion of the waveform must be discarded. Furthermore, the requirements for a real time cost-effective system have produced unscrambler devices with a fixed amount of channel storage space N, which in turn limits the minimum pitch period to a value dependent upon both N and the signal sample rate, f_s.

The block structure of a helium speech unscrambler system architecture [158] based on time domain expansion, with particular reference to the system discussed in section 4.1.3, is shown in Fig. 4.3(b). Note that most systems include preamplification in the input helium speech signal both to improve dynamic range and provide some correction for the attenuation of the high-frequency spectral region in helium speech.

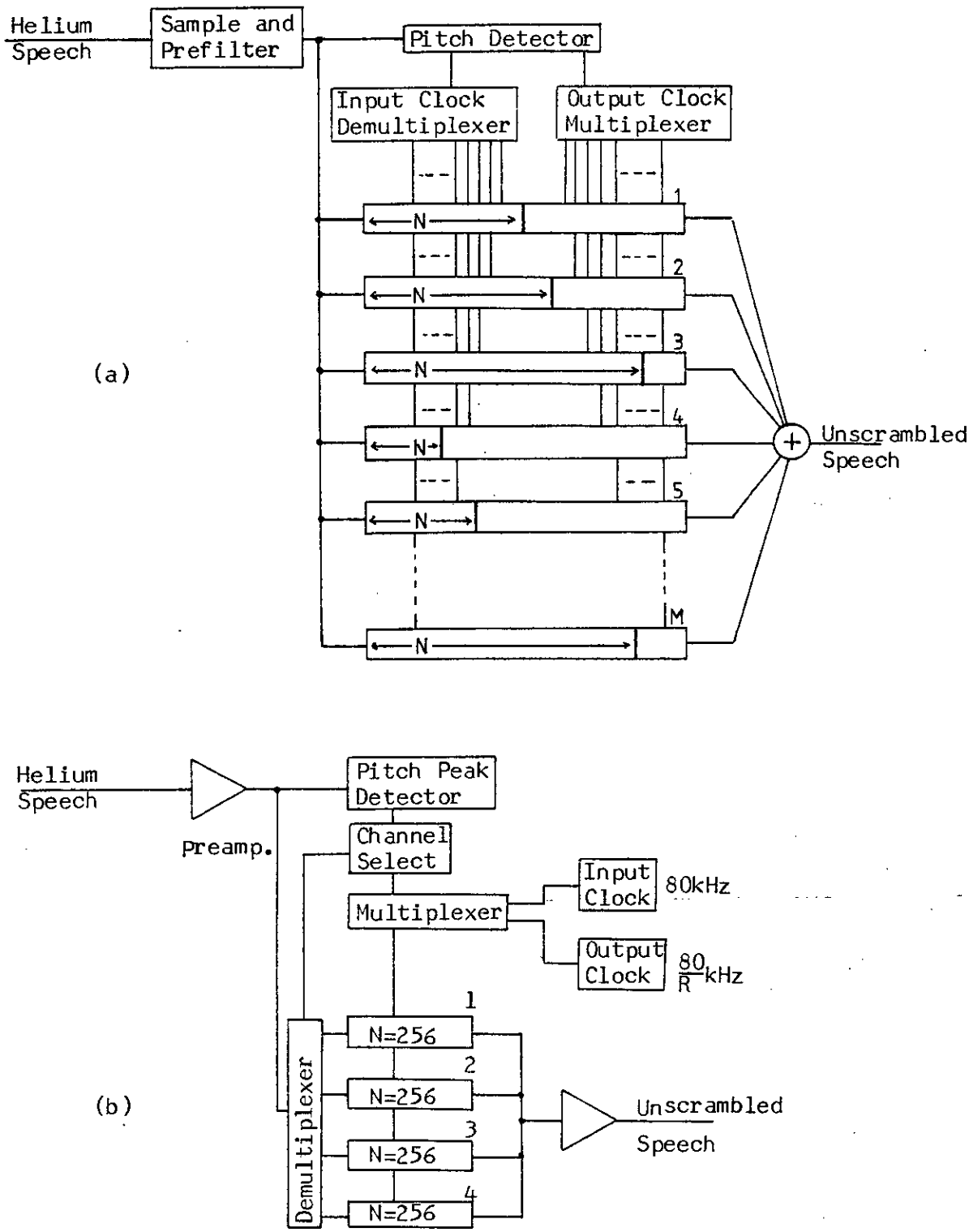


Fig. 4.3 (a) Theoretical definition of pitch-synchronous time expansion unscrambler and (b) block structure of real time system implementation.

4.1.3 SYSTEM SIMULATION

The time domain helium speech unscrambler system chosen for simulation in this case relates to a recently developed miniature unscrambler for diver-borne use [108,158], which employs the basic pitch-synchronous time expansion technique but instead of using digital waveform channel stores, employs analog charge transfer devices for waveform storage and c.m.o.s digital circuitry for control logic functions.

In the miniature system, the channel length N is 256 points long with a sample frequency of 80kHz, which gives a waveform capture time of 3.2mS, and since no other pitch period may trigger the device during this time, then the unscrambler is capable of operating with diver fundamental frequencies up to 310Hz. However, it is the design of the pitch detector circuitry which is the most crucial aspect in the operation of this system, since it not only detects the start of a pitch period, but multiplexes input and output clocks to individual storage channels.

The pitch detector in this system is based on waveform amplitude peak detection with hysteresis, and its circuit structure [159] is shown in Fig. 4.4, together with waveform timing diagrams in Fig. 4.5. The upward shift in spectral components, due to the increased speed of sound in the heliox mixture in comparison to air, produces a faster decay time-constant in the helium speech pitch waveform envelope, and therefore pitch peaks become more pronounced than in air, so that a peak detector can be successfully employed.

To explore the operation of this device and exemplify the system simulations effected in this thesis, consider firstly the steady-state condition of the system in Fig. 4.4 with no input speech.

Disregarding the full-wave rectification network around device IC1

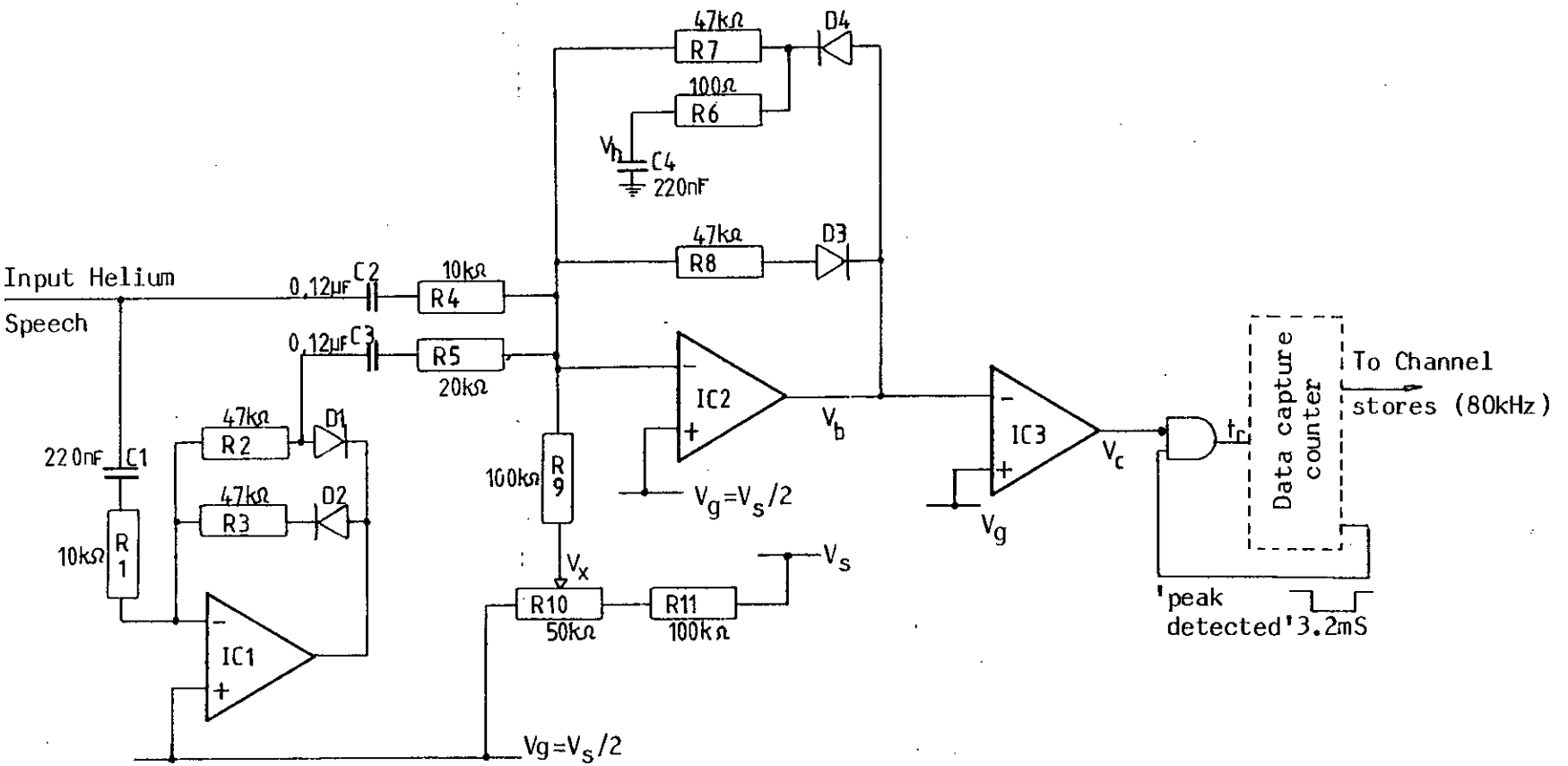


Fig. 4.4 Circuit structure of pitch detector based on waveform peak detection with hysteresis (from (159)).

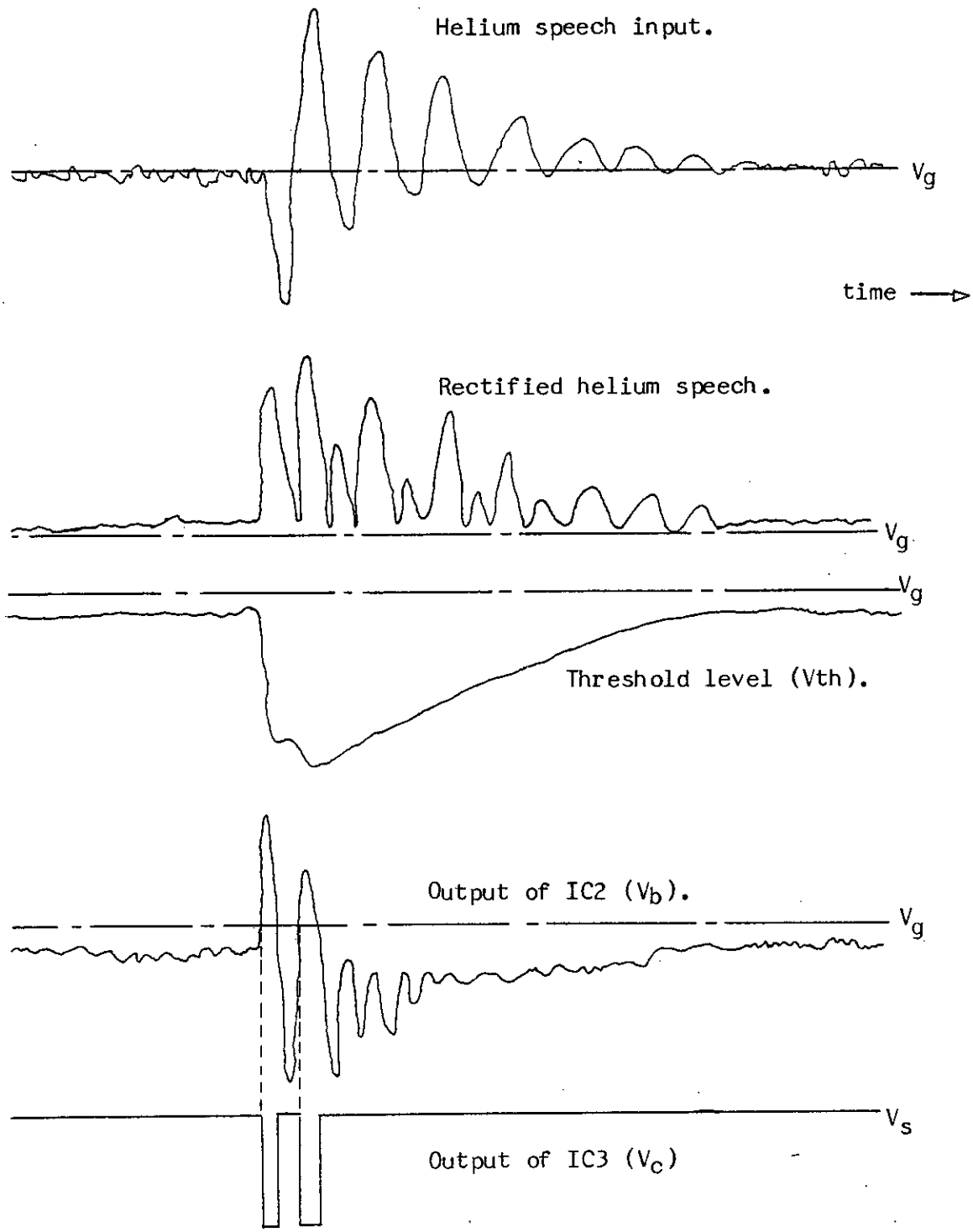


Fig. 4.5 Waveform timing diagrams for the pitch detector circuit of Fig. 4.4 (from (159)).

for the moment, then from a consideration of the net current flow into the device IC2 in the steady-state, then

$$\frac{Vs/2 - Vx}{R9} + \frac{Vs/2 - Vb}{R7} = 0 \qquad (4\text{-}3)$$

assuming the feedback path D3-R8 to be conducting with D4 reverse-biased. The circuit supply voltages are assumed to be 0V and Vs.

Solution of equ.(4-3) for Vb with substitution of the quoted values of resistance for R7 and R9 gives

$$Vb = 3Vs/4 - Vx/2 = Vth \qquad (4\text{-}4)$$

where Vth is the quiescent threshold level.

In the real system of Fig. 4.4, Vx may vary between Vs/2 and 2Vs/3, which means that in the steady-state, Vth is somewhere between Vs/2 and Vs/12, and so the 'peak detected' signal Vc = Vs in the steady-state. Neglect the hysteresis network formed by C4 and R6 for the meantime. As the input helium speech signal is full-wave rectified by the IC1 network and is added to Vth, then if Vb rises above +Vs/2 = Vg, where Vg is the peak detector trigger voltage, then Vc switches to a low voltage value below Vg, and triggers the data capture counter and the input clock multiplexing circuitry of Fig. 4.3(b).

However, it is possible that with values of Vth near to Vg, the detector will retrigger several times during any single inter-pitch waveform. To avoid this, the hysteresis network of C4 and R6 is included. Let Vh be the hysteresis voltage held on capacitor C4, which in the steady state must be Vh = Vg; the inverting action of IC2 means that Vh is then subtracted from the output voltage Vb.

The input speech signal is level-shifted to swing about Vg and is rectified about this value by the combined effect of the full-wave rectifier action of the network around IC1 and the summing action of IC2.

Let the level-shifted rectified speech signal voltage be Vr. The voltage Vb is given by:

$$Vb = Vth + Vr - Vh \qquad (4\text{-}5)$$

For the condition $Vb < Vh$, then diode D3 remains forward biased, and Vb tracks variations in the input speech waveform. If, however, $Vb > Vh$, then Vc will switch such that D3 is reverse biased and capacitor C4 charges up rapidly through D4 and R6, to the value of Vb, with a time constant Tc = 22μS. As soon as the rectified speech signal amplitude decays, then D4 is reverse biased once again, and C4 begins to discharge such that Vh decays towards Vg, although with a much longer time constant Td = 10mS. The increased value of Vh effectively lowers the value of the threshold voltage Vth by an amount $Vt\ell$:

$$Vt\ell = (Vm - Vg).e^{-t/Td} \qquad (4\text{-}6)$$

where Vm was the maximum value of Vb which occured before D4 became reverse biased.

An example of the computer program to simulate the above pitch detection circuit is shown in Fig. 4.6, and serves to illustrate the form of the programs and algorithms as applied in the simulation of the unscrambler systems which are discussed in this thesis. Each processing architecture is modelled, using floating point arithmetic, through well-defined equations which correspond to the functional properties of each device. The programming language used is the 'C' language [160], supported by the Digital Equipment Corporation's "Unix" operating system [161]. In the subroutine example "pitchdet" shown in Fig. 4.6, it can be seen that system equations (4-4)-(4-6) are implemented directly as discussed. The input speech data to the time-domain unscrambler system simulation, and to subsequent systems in this thesis, is the result of analog-to-digital conversion, to 12-bit accuracy, of the

```
#include stdio.h
#include math.h
 /* --- This program contains subroutines to provide pitch detection-*/
   float vh;  /* This declaration allows access to the value of vh
                 in the main calling routine */
#define tawc 2.2e-5  /* Charging time constant for capacitor c4 */
#define tawd 1.034e-2  /* Discharge time constant for c4 */

   int pitchdet(ts,vg,vth,vr,rst)  /* Subroutine starts here */
       float ts,vg,vth,vr;
       int rst;
       /* ts is the input data sample period. vg is the system ground
          voltage level. vth is the peak detector quiescent threshold
          level. rst is a flag indicating 'initialise' (rst=0) or
          'process' (rst=1). Note that the integer returned by this
          routine indicates 'peak detected' (vc=1) or 'no peak' (vc=0) */

       {/*1*/
       double exp(),exparg;
       float epo,rect;
       static float vb,vm,td;
       if (rst == 0) {/*2 reset detector voltages at start-up */
                  vm = vh = vg; /* vh is the simulated voltage on c4 */
                  td = 0.0; /* elapsed discharge time for c4 */
                  /* Note that vm logs the maximum voltage to which
                     c4 will charge */
                  return;
                }/*2*/
       vb = vth + vr - vh;  /* vb is the input voltage into IC3 and
                               determines whether vr is the start
                               of a pitch period */
       rect = 0.0;
       if (vr< vg) rect = 2.0 * (vg - vr); /* rectify vr about vg */
       vb = vb + vr;
       /* This is the value of vb used to determine pitch: first,
          however, calculate updated circuit voltages for next call
          to this routine. See if capacitor c4 will charge or discharge */
       if (vb> vh) {/*3*/
                  exparg = 0.0 - ts/tawc;
                  epo = (float)exp(exparg);
                  vh = vm = vh + (vb - vh) * (1.0 - epo);
                  /* Here, capacitor is charging up towards vb */
                  td = 0.0; /* reset elapsed discharge time to 0 */
            }/*3*/
        else{/*4*/
                  td = td + ts;
                  exparg = (double)(0.0 - td/tawd);
                  epo = (float)exp(exparg);
                  vh = vg + (vm - vg) * epo;
            }/*4*/
       if (vb  vg) vc = 1; /* Peak detection by threshold comparison */
        else vc = 0;
       return(vc);
   }/*1 End of routine */
```

Fig. 4.6 Simulation program for pitch detection by peak-picking with hysterisis (language = 'C').

recorded helium speech as listed in Fig. 3.1. Computer-generated test signals which are employed occasionally are also produced with 12-bit accuracy. Thus input data for processing by any simulated system is limited to integer values within the range -2047 to +2047. Note, however, that once entered into a simulation program, this data is converted to floating-point format and may, within certain processing algorithms, be manipulated to produce non-integer values. However, in order to provide a basis for subjective comparison of the performance in terms of intelligibility of the unscrambled speech from different unscrambler systems, the resultant signal from each simulation is requantised to 12-bit integer accuracy to allow real time digital-to-analog conversion (see Appendix B).

In accordance with the circuit details of the time-domain unscrambler, while the input clock rate is fixed at 80kHz, the output clock rate is continuously variable from 80kHz down to 26kHz, catering for the maximum theoretical time expansion of R = x3. Storage channel administration was such that no channel could be written into while it was still carrying out time expansion. This condition is only likely to occur for a series of pitch periods with an average period less than the signal capture time of 3.2mS. Lastly, if several channels are expanding their stored speech waveforms simultaneously, then all outputs are summed together.

Text from "The Rainbow Passage" [121] (see Fig. 3.1) as used in the present simulation can be heard spoken by the subject in a normal air atmosphere in recording A4, with the corresponding text spoken at a depth of 100ft in heliox in rec. A5. The resultant speech from the time domain unscrambler simulation can be heard in rec. A6(a), and compares favourably with the speech as unscrambled by the physical device,

rec. A6(b), which was available in this case.

4.1.4 EVALUATION

It is clear from the system simulation in section 4.1.3 that the most critical parameters in the performance of the time-domain unscrambler relate to the choice of quiescent threshold voltage Vth and hysteresis decay time constant Td. A quiescent threshold Vth which is set too close to the trigger voltage Vg may cause repeated triggering of the peak detector within any single pitch period, as indeed may a decay constant Td which is short compared to the expected time constant Te of the pitch waveform envelope decay. Conversely, the peak detector may fail to identify the start of a pitch period if Vth is set too low such that Vb, from equ.(4-5), never crosses the required trigger threshold Vg, and/or if the decay constant Td is too long compared to Te. For the given system value of Td = 10mS in this simulation, the value of quiescent threshold which was judged subjectively to produce the best continuity and intelligibility of the unscrambled speech in rec. A6(a) was found to be Vth = 0.445Vs, that is, Vx = 0.61Vs from equ.(4-4).

From recordings A6(a) and (b), the resultant unscrambled speech exhibits audible discontinuities, to the detriment of intelligibility. This effect cannot directly be related to the acoustic properties of helium speech, but is attributable to the processing mechanics of the unscrambler system. There are several possible causes for this fragmented speech output.

Assuming meantime that all voiced pitch intervals are correctly detected, then the apparent fragmentation may be due to the time expansion of only a finite portion of the pitch waveform, in that there may

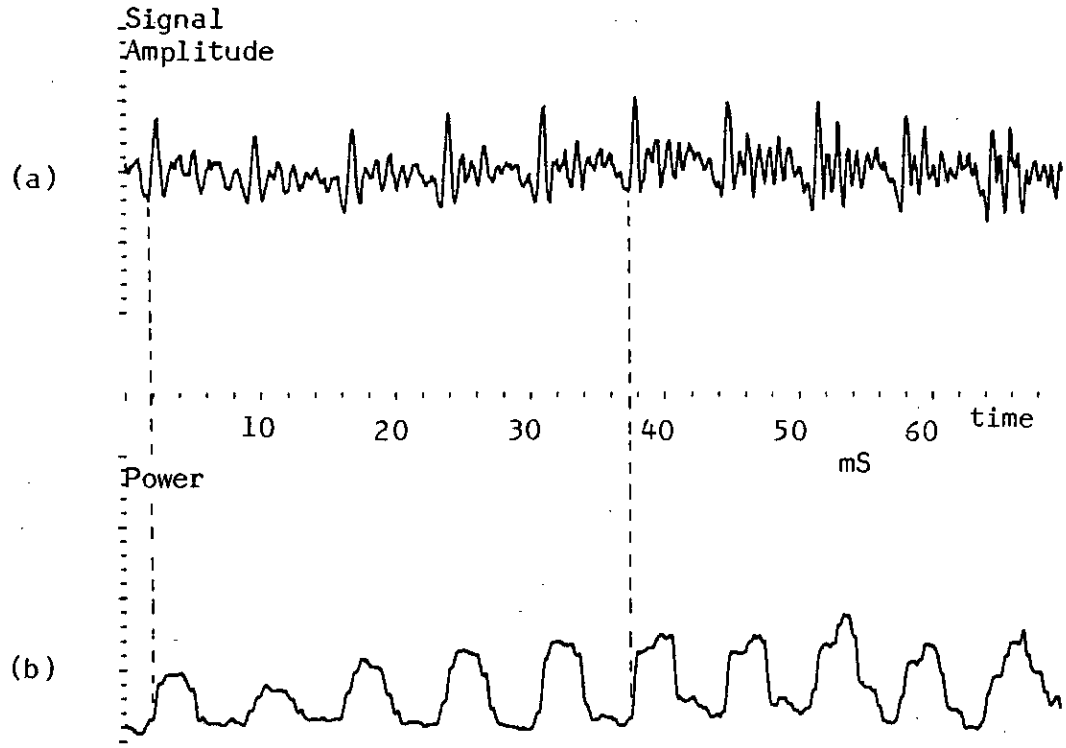


Fig. 4.7 (a) Voiced speech in air (f_s = 10kHz) and (b) running average power of speech waveform over a 3.2mS window.

be significant perceptual information residing in the discarded portion of the pitch waveform. To investigate this proposition, the running average power of the digitised speech corresponding to rec. A4 of "The Rainbow Passage" in air, was calculated by scanning a rectangular window of length Tw = 3.2mS across the entire speech waveform on a sample-by-sample basis. Since it is desirable to observe the variation of average power within individual pitch periods, this value of Tw avoids calculation of average power over several pitch periods, whose expected minimum period is of the order of 7mS from Fig. 3.7, and at the same time smooths over instantaneous variations in signal power due to individual cycles in the pitch waveform. Typical results, shown in Fig. 4.7, demonstrate that the average power of the pitch waveform for air speech falls off sharply beyond some 3-4mS after the start of each pitch period. Given that each channel store in the device retains the

first 3.2mS of the pitch waveform of helium speech, and since the inter-pitch helium speech waveform decay is much faster than in air, then it is therefore not expected that significant perceptual information is being rejected.

Another possible reason for the fragmentation of the speech may be due to the choice of quiescent threshold voltage, Vth, in that not all pitch periods are being detected. To investigate this, digitised helium speech was observed using the graphics package "scope", which simulates a single/dual trace oscilloscope, and markers were placed in the speech to identify the commencement of pitch periods as judged by manual identification. The hand-marked speech was then compared visually against the 'peak detected' signal from the simulation, and it was found that for strongly-voiced speech sounds the triggering of the detector corresponded well to the hand-marked pitch intervals, see Fig. 4.8(a-b), whereas weaker voiced pitch was less reliably detected, see Fig. 4.8(c-d), thereby accounting for part of the broken effect of the unscrambled speech. The same problem, however, relates to unvoiced speech, for which there is no defined pitch interval. The operation of the device under these conditions depends on random excursions of the speech waveform such that Vb exceeds the trigger voltage Vg, and therefore depends on the dynamic range of the helium speech coupled with the value chosen for the quiescent threshold. An auditory comparison of the original helium speech in recording A5 against the unscrambled speech in recording A6(a) demonstrates that the worst breaks in the unscrambler output, in fact, occur during periods of unvoiced speech. It has been shown that the disruption of continuity of the speech in this manner degrades intelligibility such that an increasing frequency of disruption beyond 10Hz produces a progressive impairment of intelligibility. This topic

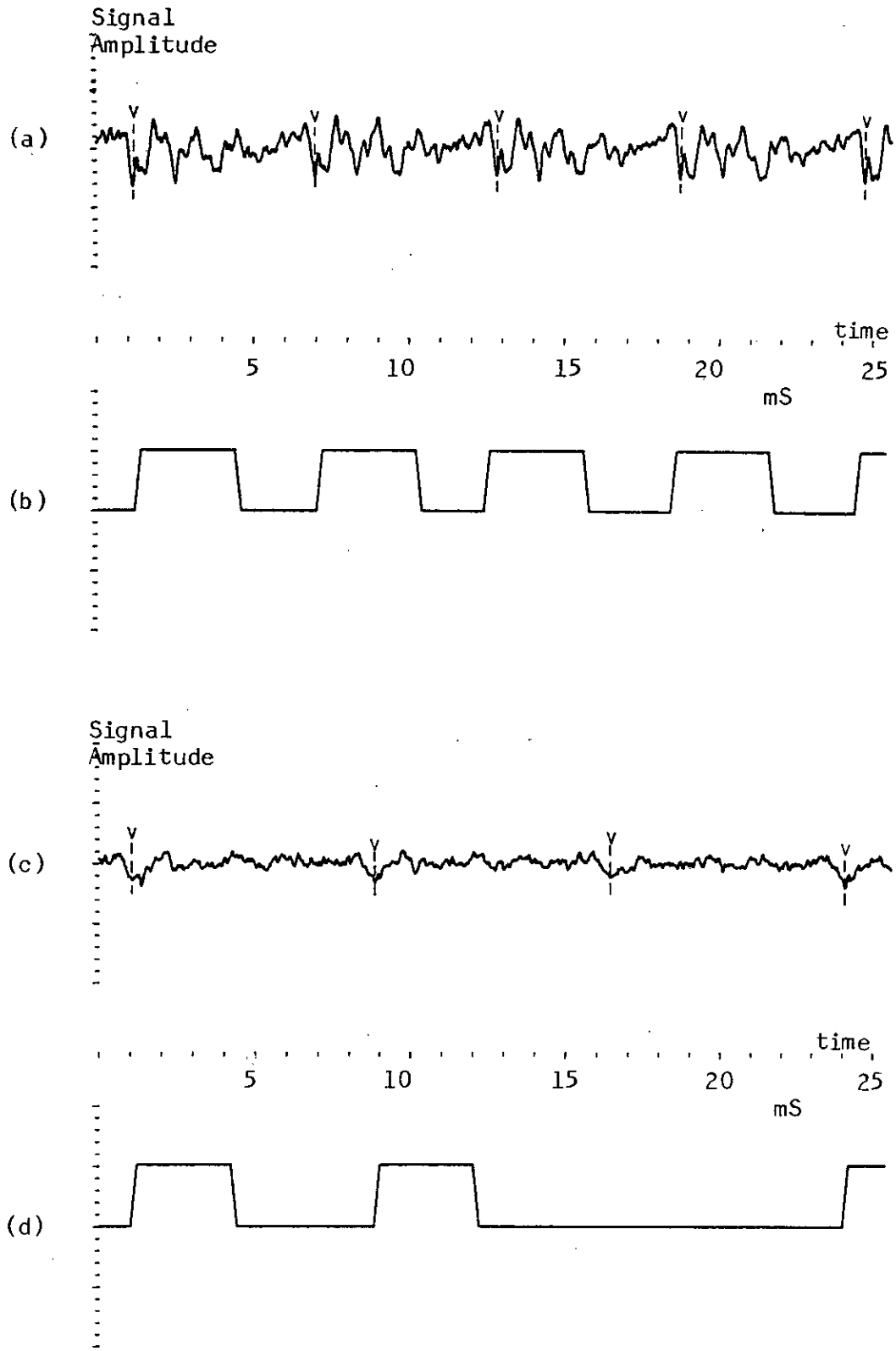


Fig. 4.8 (a) Strongly-voiced helium speech and (b) pitch detector signal from simulation with Vx=0.61Vs. (c) Weakly-voiced helium speech and (d) pitch detector output (f_s = 80kHz).

Y = manually-identified commencement of pitch period.

is explored further in section 4.3.

In addition to functional singularities due to a threshold-dependent pitch detector, the dependence of the system on processing only the time-varying speech signal itself in a linear manner renders correction of the nonlinear formant bandwidth translation characteristic, as found in section 3.4.6, impossible, and also renders difficult the correction for helium speech formant amplitude attenuation found in section 3.4.5. Since this attenuation in formant power has been shown to be of the order of -6dB per octave translation in formant frequency from air to heliox, then each area of the speech spectrum corresponding to any single formant effectively requires a different correction in amplitude dependent upon the gaseous composition of the heliox mixture, or alternatively the corresponding time expansion factor R. Thus a different preamplification filter would be required for each change in the expansion factor R.

In summary, although providing a simple and cost-effective realisable unscrambler system, the functional properties of a system employing time domain processing for the correction of helium speech are essentially incompatible with the prescribed requirements in respect of good intelligibility of the unscrambled speech.

4.2 HELIUM SPEECH UNSCRAMBLING USING THE SHORT-TIME FOURIER TRANSFORM

For the general class of linear systems, then by the superposition principle [162], if $[e_1(t), s_1(t)]$ and $[e_2(t), s_2(t)]$ are excitation-response pairs, then if the excitation were $e(t) = e_1(t) + e_2(t)$, the response $s(t) = s_1(t) + s_2(t)$. Also, if the system transfer function

is described by h(t), then s(t) is the convolution [163] of h(t) and e(t):

$$s(t) = h(t) * e(t) \qquad (4\text{-}7)$$

In the case of helium speech correction, one or all of the system waveforms h(t), e(t) or s(t) must be manipulated according to some predetermined correction algorithm. One method to achieve the required correction might be to solve for the required parameter by expressing equ.(4-7) as a series of linear differential equations. However, since the problem here relates to manipulation of the power spectrum, which is a function of frequency, then it is natural to transform the time-varying system waveforms of equ.(4-7) into the frequency domain by the convolution theorem [164]:

$$s(t) = h(t) * e(t) \rightleftharpoons S(f) = H(f) \times E(f) \qquad (4\text{-}8)$$

where X(f) is the Fourier transform of the respective time-varying signal x(t), and is given by:

$$X(f) = \int_{-\infty}^{\infty} x(t).e^{-j2\pi ft}dt \qquad (4\text{-}9a)$$

and the inverse Fourier transform is given by:

$$x(t) = \int_{-\infty}^{\infty} X(f).e^{j2\pi ft}df \qquad (4\text{-}9b)$$

where t and f represent continuously variable time and frequency respectively.

This transformation is particularly useful since the convoluted relationships of time-varying quantities transform into algebraic relationships in the frequency domain and vice-versa. This simplifies considerably the mathematical manipulation of the various component signals. Notice at this point that whereas the helium speech correction algorithms which may be nonlinear as a function of frequency, the basic assumption permitting the use of the Fourier transform representation demands that the system under consideration is linear and conforms to

the principle of superposition. This is exemplified for the speech signal by a consideration of the model of speech production presented in section 2.1 and Fig. 3.8. A further assumption relating to the use of the Fourier transform is that the speech system is time-invariant [165], that is:

$$s(t-T) = h(t-T) * \delta(t-T) \qquad (4\text{-}10)$$

where $\delta(t-T)$ is a Dirac impulse occuring at some random time t=T. This condition is required since the theoretical definition of the Fourier transform involves integration over infinite time, see equ.(4-9a). From the discussion of the speech mechanism presented in section 2.1, it was seen that the configuration of the vocal tract, corresponding to h(t-T) in equ.(4-10), could only be considered time-invariant over short periods of time, and therefore special precautions are required to ensure that the Fourier transform, in this application to speech correction, will process portions of the speech signal which can be considered as time-invariant in the short term. This involves application of time windows [166] to the speech waveform, prior to processing, which effectively force the signal to exist only for finite time. The treatise of the short-time Fourier transform unscrambler presented here assumes that the reader is familiar with the standard definitions and properties of both the continuous and discrete Fourier transforms [164,167,168] together with the spectral properties of different temporal windows [166], with particular reference to the Bartlett window [169].

4.2.1 BASIC PRINCIPLES

Consider the continuous Fourier transform E(f) of an infinite series of Dirac impulses spaced equally in time with period T:

$$E(f) = \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(t-nT) e^{-j2\pi ft} dt \qquad (4\text{-}11)$$

and since the impulse function is defined [133] by the equation:

$$\int_{-\infty}^{\infty} \delta(t-to).x(t)dt = x(to) \qquad (4\text{-}12)$$

then it can be shown [170] that E(f), from equ.(4-10), is given by:

$$E(f) = 1/T \sum_{n=-\infty}^{\infty} \delta(f-n/T) \qquad (4\text{-}13a)$$

or, alternatively, using equ.(4-12):

$$E(f) = \sum_{n=-\infty}^{\infty} e^{-j2\pi fnT} = 1 + 2\sum_{n=1}^{\infty} \cos(2\pi fnT) \qquad (4\text{-}13b)$$

That is, the Fourier spectrum E(f) of a periodic impulse train is itself periodic in frequency, and this is shown in Fig. 4.9(a) and (b). If this impulse train excites a linear time-invariant (LTI) system whose transfer function is h(t) as shown in Fig. 4.9(c) and whose spectral response is given by H(f), Fig. 4.9(d), then the resulting time signal s(t), see Fig. 4.9(e), is given by the convolution integral [171]:

$$s(t) = \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(\tau-nT).h(t-\tau)d\tau \qquad (4\text{-}14a)$$

$$= \sum_{n=-\infty}^{\infty} h(t-nT) \qquad (4\text{-}14b)$$

That is, s(t) is a periodic function in time with the impulse response h(t) repeated every T seconds. Furthermore, from the convolution theorem expressed in equ.(4-9), then the two spectra H(f) and E(f) are simply multiplied in the frequency domain, therefore the signal spectrum S(f), Fig. 4.9(f), is similar to the spectrum of E(f) in that S(f) is periodic in frequency with the same period, Fo = 1/T, but each spectral line of E(f) is now weighted according to the corresponding value of H(f). Thus, if the signal s(t) is assumed to be the result of the period impulse excitation of an LTI system, then if H(f) is known or can be estimated,

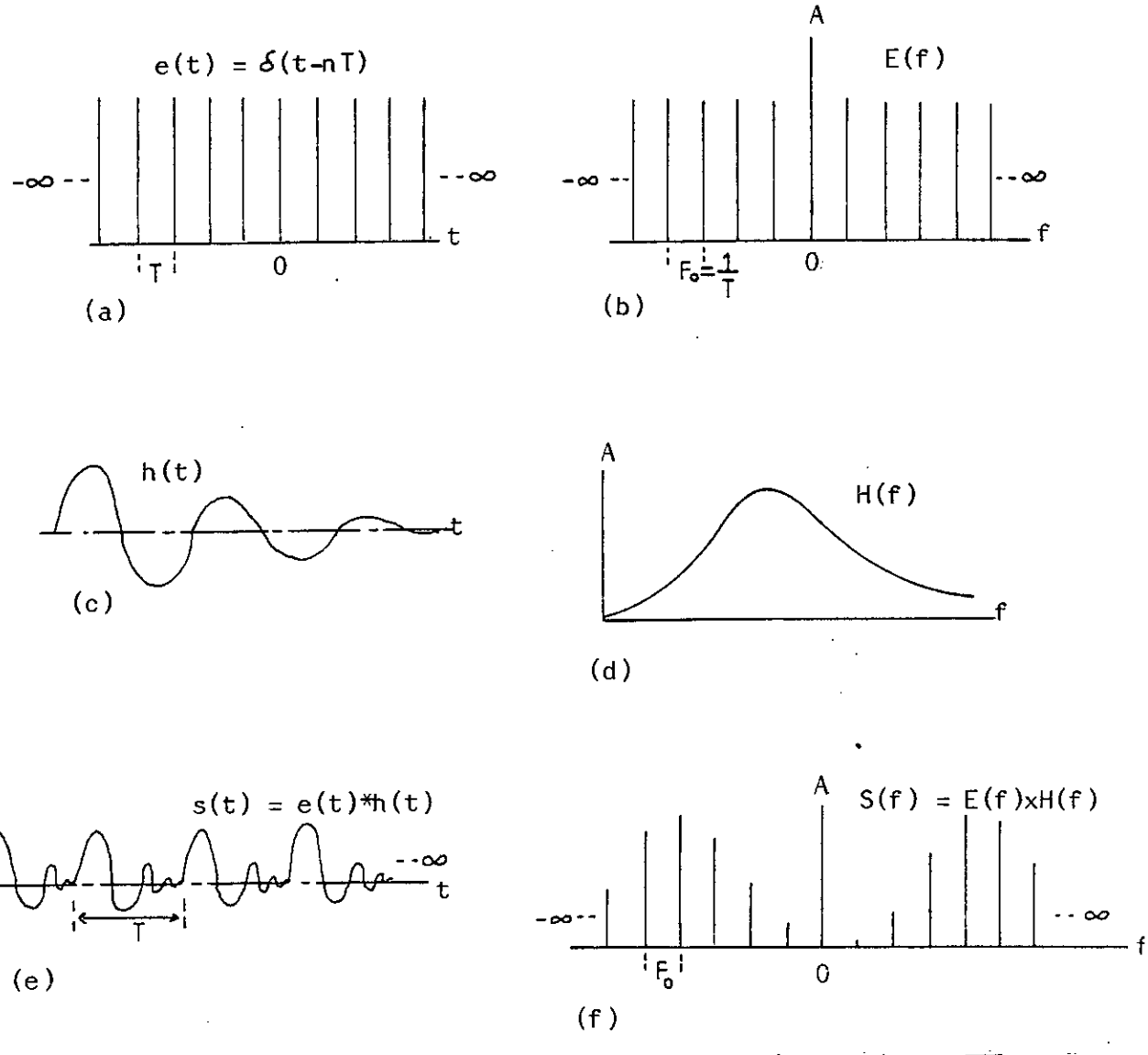


Fig. 4.9 (a) Periodic impulse train e(t) and
(b) corresponding Fourier spectrum E(f).
(c) System impulse response h(t) and
(d) spectral response H(f).
(e) Periodic signal s(t) from the convolution of e(t) * h(t) and (f) corresponding spectrum S(f) given by E(f) x H(f).

then E(f) can be recovered [92,172] from equ.(4-9):

$$E(f) = S(f)/H(f) \qquad (4\text{-}15)$$

This is an important feature in the use of the short-time Fourier transform for helium speech unscrambling. As has been confirmed in section 3.2, the fundamental frequency of voiced speech is unaltered by a pressured heliox atmosphere, and therefore the periodicity of normal and helium speech voiced waveforms is identical: if S(f) in equ.(4-15) is the spectrum of the periodic helium speech voiced waveform, and the vocal tract frequency response H(f) can be obtained, then the excitation spectrum E(f) can be derived, thereby obviating the need to maintain pitch information by pitch extraction in the time domain [173]. Correction for the helium speech distortion is then applied directly to H(f) to produce a new estimate of the vocal tract frequency response $\hat{H}(f)$, which is then remultiplied together with the excitation spectrum E(f) to produce a new spectrum Z(f) corresponding to the unscrambled speech signal z(t):

$$Z(f) = \hat{H}(f).E(f) \qquad (4\text{-}16)$$

and z(t) is obtained from the inverse Fourier transform of Z(f):

$$z(t) = \int_{-\infty}^{\infty} Z(f) e^{j2\pi ft} df \qquad (4\text{-}17)$$

The reasoning for the case of unvoiced speech is similar, and demonstrates the procurement of E(f) and H(f) in the ideal case. The excitation source is assumed to be white noise which excites the vocal tract filter h(t); E(f) is therefore flat in nature and continuous, and so any colouration in the signal spectrum S(f) is entirely due to vocal tract filtering. H(f) in this case corresponds exactly to the signal spectrum, S(f), and so can be manipulated directly to produce the corrected speech spectrum Z(f). Notice however that for the voiced case, whilst the amplitude of any spectral line at frequency f_n corresponds

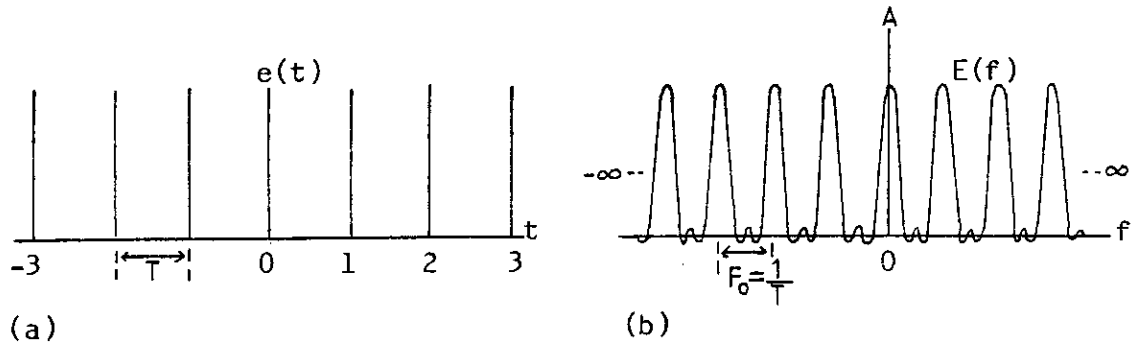


Fig. 4.10 (a)Truncated periodic impulse train and (b) Fourier spectrum.

to the transfer function response $H(f_n)$, special precautions are necessary to ensure that the spectral comb spacing Fo remains intact on manipulation of H(f).

4.2.2 SHORT-TIME SPECTRAL ENVELOPE EXTRACTION

In the case of speech, the waveform can only be considered periodic in the short term [12]. From equ.(4-13b), the (continuous frequency) Fourier transform of a truncated series of Dirac impulses, which exist in time up to $t = \pm qT$, is given by:

$$E(f) = 1 + 2\sum_{n=1}^{q} \cos(2\pi fnT) \qquad (4\text{-}18)$$

and an example of the resulting transform for q=3 is given in Fig. 4.10(a-b).

Notice that there are now sidelobes of significant amplitude located about either side of the main spectral peaks at f=n/T, where n is any integer value. The implications in terms of a truncated complex periodic signal of the form of Fig. 4.9(d) are that, not only does the frequency response of the transfer function multiply the main spectral peaks, but the sidelobes too are affected.

In the case of a periodic signal s(t) of infinite duration as in Fig. 4.9(e), the transfer function spectral envelope, H(f), could readily be obtained by employing, for example, a spectral peak-searching routine which found and logged the amplitude of each peak in the signal spectrum S(f). Such a simple method could not be used here, however, since simply calculating the amplitude of any sidelobe peak occuring at some frequency would lead to an erroneous estimation of the transfer function spectrum at that frequency. If the time function were exactly periodic within time $\pm qT$, and if q could be estimated, then it would be possible to calculate the extent of sidelobe amplitude from explicit expansion of equ.(4-18) as a function of frequency. This would in fact be beneficial, since more information about H(f) would now be known at intermediate frequencies about f=n.Fo: however, the situation is complicated further by the fact that the speech signal is quasiperiodic only in the short term, and the result of this on the spectrum, as has been discussed in section 2.1.1 (see Fig. 2.5), is to smear the energy of each spectral line according to the amount of jitter around the mean fundamental frequency [25]. Therefore the combined effects of signal truncation and quasiperiodicity render direct calculation of the vocal tract frequency response H(f), even given information regarding the statistical characteristics of the waveform jitter, very difficult and tedious, and such a solution is not propitious for real time implementation. Spectral envelope estimation algorithms [174] are therefore necessary, and the implementation of a piecewise linear curve-fitting technique [92] is discussed in section 4.2.5.

4.2.3 SAMPLING AND WINDOWING OF THE SHORT-TIME SIGNAL

Whilst the foregoing discussions have assumed functions and their transforms which are continuous both in time and frequency, practical considerations concerning implementation of the Fourier transform in real time require that the speech signal be sampled in time at some sample rate, f_s, with period T_s. From Fig. 4.9(a-b), it was seen that the spectrum of an infinite series of impulses of period T transformed into another set of impulses in the frequency domain with spacing $f = 1/T$, and this is evidently the case for a series of unit sample impulses $\sigma(nT_s)$, whose comb spacing in the frequency domain is given by $f_s = 1/T_s$, see Fig. 4.11(a-b). However, the unit sample series $\sigma(nT_s)$ is now considered to multiply the periodic continuous time signal s(t) of Fig. 4.11(c) which corresponds to the convolution in the frequency domain of the sample series spectrum $\sigma(f)$ in Fig. 4.11(b) and the continuous signal spectrum S(f) in Fig. 4.11(d) to produce the sampled signal spectrum $S_s(f)$ of Fig. 4.11(f). Notice that the original signal spectrum S(f) is now repeated about harmonics of f_s, therefore if non-overlapping spectra are required, then the signal must be low-pass filtered with a maximum cut-off frequency of $f_s/2$, which is simply a restatement of the Nyquist sampling theorem [175].

It is important to note that whilst the graphical presentation here relates to spectral amplitude, the spectral phase characteristic, although not shown here explicity, contributes equally importantly to the uniqueness of the signal spectrum [176]. Thus, while ostensibly it may appear from Fig. 4.11(f) that the spectrum is uniquely defined up to $f_s/2$, this really only relates to spectral magnitude. The phase characteristic, for most classes of signals, extends uniquely to f_s and is repeated thereafter in the aliased spectra about the harmonics of f_s.

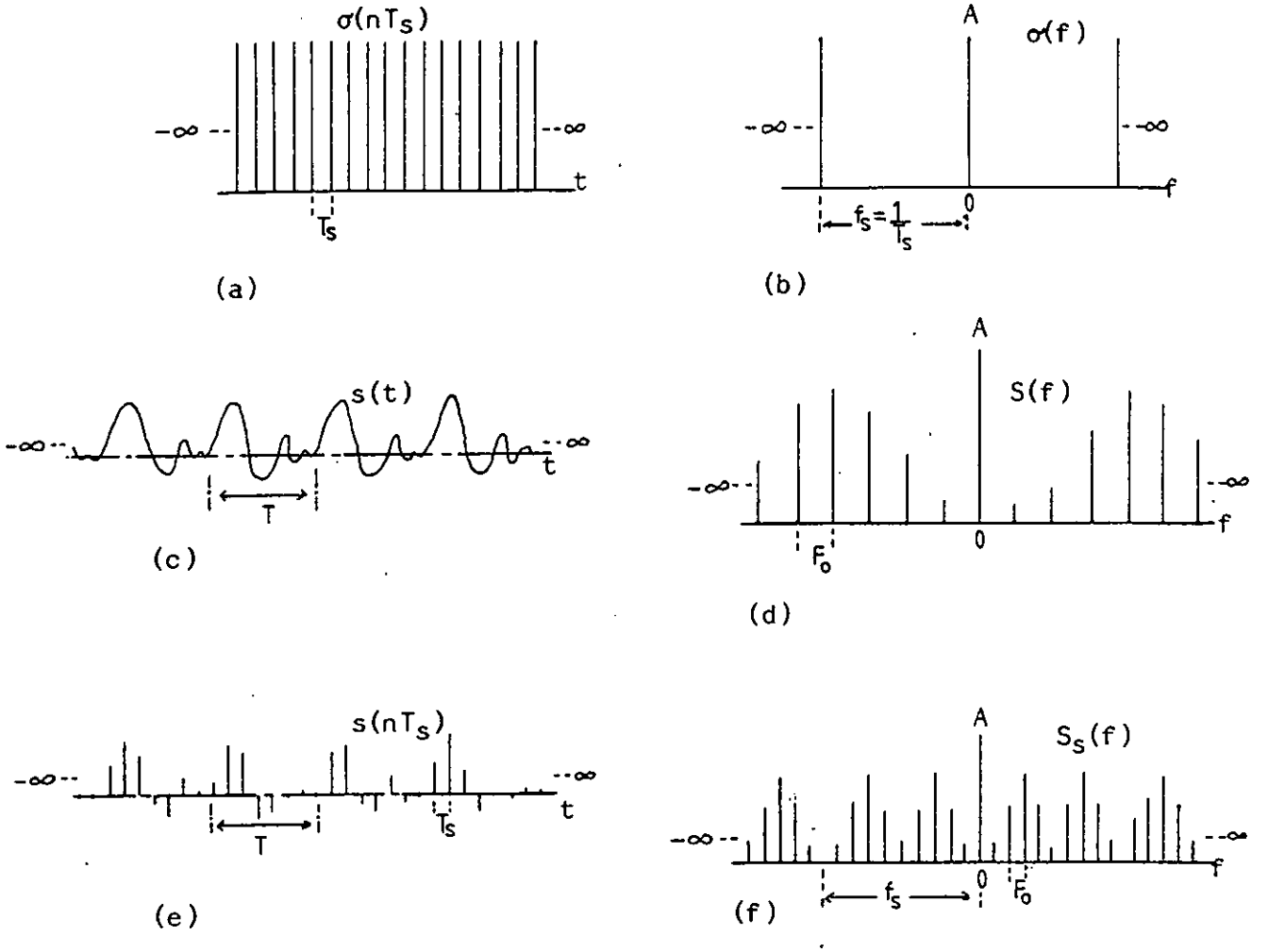


Fig. 4.11 (a) Series of unit samples $\sigma(nT_s)$ and
(b) spectrum $\sigma(f)$.
(c) Continuous periodic signal s(t) and
(d) spectrum S(f).
(e) Sampled signal $s(nT_s)$ given by $\sigma(nT_s) \times s(t)$ and (f) corresponding spectrum $S_s(f)$ from the convolution of $\sigma(f) * S(f)$.

Truncation of the sampling sequence, such that it is effectively viewed through a time window of length T_w, produces an exactly similar effect on the sample series spectrum $S_s(f)$ to that demonstrated in the truncated series and its spectrum in Fig. 4.10(a-b), except that the spectral sidelobes are in this case distributed about harmonics of f_s. Recalling that the multiplication of the sampling sequence into the continuous time signal leads to a convolution in the frequency domain, then the existence of non-zero spectral amplitude between the frequency comb harmonics of f_s means that the amplitude at any frequency f_c in the signal spectrum is no longer uniquely due to the amplitude of f_c itself, but will contain contributions from other frequencies depending on the distribution of the sidelobes and therefore the length of the time window T_w, for any given sampling frequency f_s. An example of this phenomenon, termed spectral leakage [177], is shown in Fig. 4.12(a-d), and demonstrates the spectral effects of signal discontinuities due to truncation in time. The diminution of spectral leakage entails premultiplication of the truncated sequence by a specially-tailored window function in order to render the time series continuous at the window edges in as many of its derivatives as possible [178]. In Fig. 4.12(c), the sequence is effectively windowed by a rectangular function whose value is =1, for $0 < t < T_w$, and =0 elsewhere. Premultiplication of the sequence first by, say, a triangular (Bartlett) window [169] as shown in Fig. 4.12(e) to produce Fig. 4.12(g) amounts to correlating the sampled signal spectrum $S_s(f)$ with the spectral characteristic of the window, $W(f)$, shown in Fig. 4.12(f), whose sidelobes are much reduced compared to those corresponding to the rectangular window of Fig. 4.12(c), thereby reducing the leakage effect.

Window carpentry must therefore be an integral consideration in the

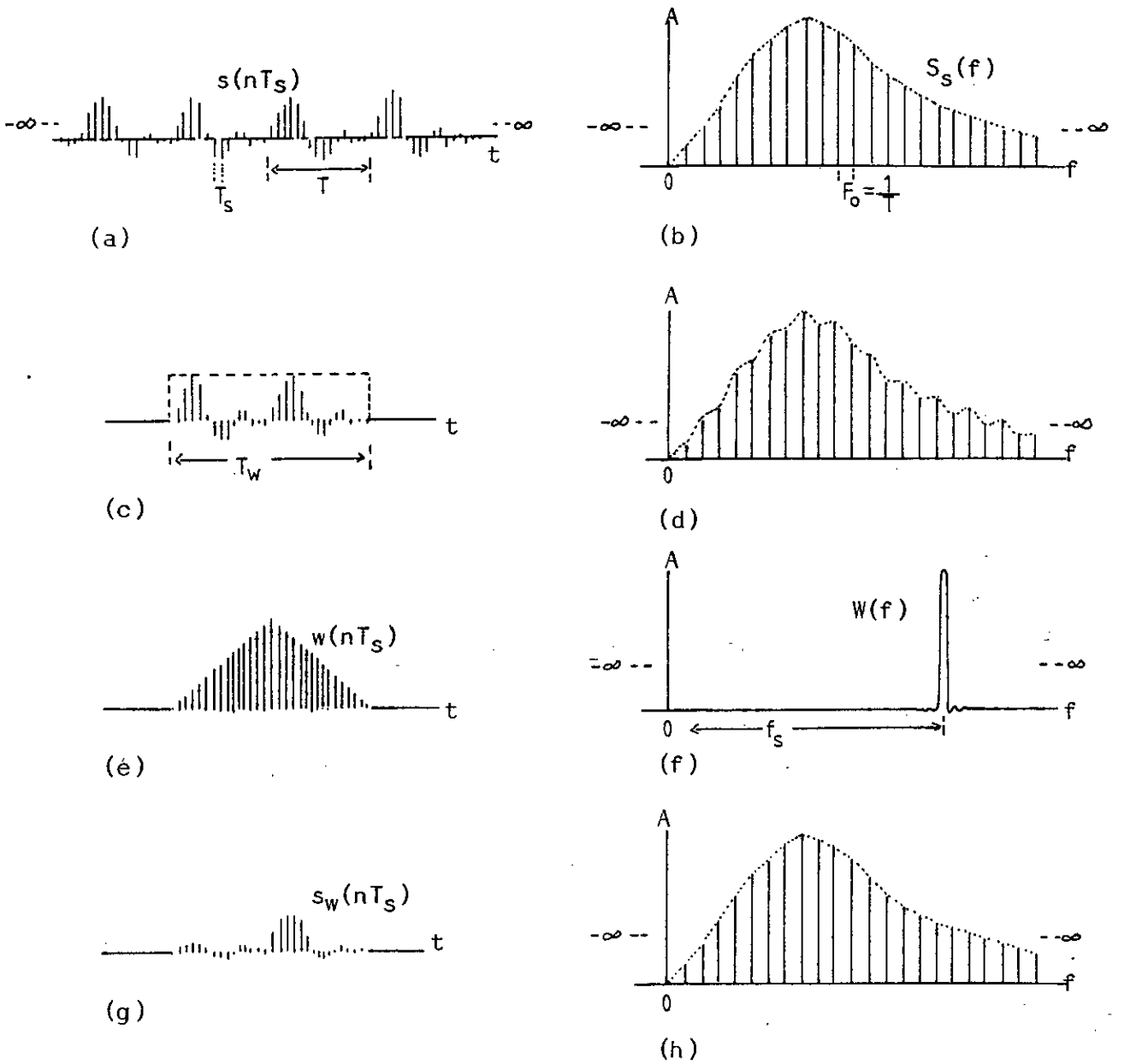


Fig. 4.12 (a) Sampled periodic signal $s(nT_S)$ of infinite duration and (b) section of corresponding spectrum $S_S(f)$. (c) Truncated signal of length T_W and (d) spectrum demonstrating spectral leakage due to truncation. (e) Triangular (Bartlett) window function $w(nT_S)$ and (f) corresponding spectrum $W(f)$. (g) Windowed signal $s_W(nT_S)$ given by $s(nT_S) \times w(nT_S)$ and (h) spectrum demonstrating reduced leakage from the convolution of $S_S(f) * W(f)$.

short-time Fourier transform if the resulting spectrum is to adequately represent the shape of short-time signal spectrum [168], see Fig. 4.12(h).

4.2.4 SPECTRAL RESOLUTION AND THE DISCRETE FOURIER TRANSFORM

The system principles as outlined in sections 4.2.1-3 have so far assumed that the short-time transform is continuous in frequency, that is, infinite frequency resolution has been assumed. In consideration of real time implementation once again, however, it is only permissible to calculate the spectral amplitude for a finite number of frequencies. Since the signal spectrum (comprising magnitude and phase characteristics) is repeated in frequency every f_sHz, then let the transform be calculated only in the range 0 - f_sHz at M discrete frequencies in the spectral domain to give a resolution of f_s/M Hz.

Since the Fourier transform now relates to a finite windowed series of N sampled data points in discrete time, then the continuous integration definition of the transform from equ.(4-9a), must now become a discrete summation over time index n, $0 \leqslant n \leqslant N-1$. Let M be the frequency index, $0 \leqslant m \leqslant M-1$.

The discrete Fourier transform (DFT) [167], which operates on the signal $x(nT_s)$ to produce the spectrum $X(m)$ at some frequency $m.f_s/M$ is then given by:

$$X(m) = \sum_{n=0}^{N-1} x(nT_s)e^{-j2\pi m.(f_s/M).nT_s} \qquad (4\text{-}19a)$$

$$= \sum_{n=0}^{N-1} x_n.e^{-j2\pi mn/M} \qquad m = 0,1,2,\ldots M-1 \qquad (4\text{-}19b)$$

where x_n is the sampled signal value at time nT_s, and the short-time windowed signal contains N sample values. There is no loss of generality in the foregoing discussions on spectral leakage and windowing in the

transition from the continuous transform to the DFT, but there are certain costraints which must be imposed upon the relative values of M and N in equs.(4-19). If x_n were a completely random complex signal, then the window length N could be very large, N>M, and yet whilst this would cause the exponential term $e^{-j2\pi mn/M}$ to cycle many times through the same values, the associated unique values of x_n with time force a unique contribution to the overall value of X(m). In the limiting case, however, where x_n is periodic (within the time window) with period = MT_s, then for any sample value x_r at time $r.T_s$, the value at time $(r + M).T_s$ is $x_{r+M} = x_r$.

From equ.(4-19b), for any frequency m, then the contributions at times n = r and n = r+M are given by:

$$x_r.e^{-j2\pi mr/M} + ... + x_{r+M}.e^{-j2\pi m(r+M)/M} \qquad (4\text{-}20a)$$

$$= x_r.e^{-j2\pi mr/M} + ... + x_r.e^{-j2\pi mr/M} \qquad (4\text{-}20b)$$

since $e^{-j2\pi mM/M} = 1$

That is, their spectral contributions are identical, therefore spectral information contributed by the time series x_n is unique in this case only up to N = M. Thus, there are N discrete frequency values for a time window of length N sample points, and equ.(4-19b) is expressed as:

$$X(m) = \sum_{n=0}^{N-1} x_n.e^{-j2\pi mn/N} \qquad m = 0,1,2,...N-1 \qquad (4\text{-}21)$$

Since the spectral resolution is given by f_s/N, then it is apparent that there is an important trade-off between time window length and spectral resolution. A long time window implies fine spectral resolution but at the expense of averaging over short-term spectral fluctuations in the signal, whereas a short time window will capture very short-time spectral variations at the expense of coarse spectral resolution.

4.2.5 SHORT-TIME FOURIER TRANSFORM (STFT) UNSCRAMBLER ARCHITECTURE

The architecture described here and subsequently simulated in section 4.2.6 is based largely on details of a STFT unscrambler system which has been implemented in real time [172], with the exception of the spectral envelope estimation algorithm, which relies on a piecewise linear curve fitting method implemented in the original simulations of the STFT unscrambler system [92].

The basic system concept is identical to that discussed in section 4.2.1, in which a short-time segment s(t) of helium speech is windowed and transformed by the STFT to produce the signal spectrum S(f); the spectral envelope H(f) is then derived and used to determine the excitation spectrum E(f) according to equ.(4-15); H(f) is then warped to remove the helium speech distortion, then remultiplied into E(f) to produce a new spectrum Z(f) corresponding to the unscrambled speech; Z(f) is then inverse transformed to produce the resulting intelligible speech waveform z(t). The sequence of operations is shown in Fig. 4.13.

In the real time system, the helium speech signal is band-limited to 7.4kHz and sampled at 14.8kHz. The samples are windowed by a symmetrical triangular (Bartlett) window and transformed with an analysis frame length N=512 points, therefore the spectral resolution is 29Hz. The fast Fourier transform (FFT) implementation [144,179,180,181] of the DFT is used in this case to aid with real time operation. Contiguous analysis frames overlap by 50% such that the data in the last half of any analysis frame A_n also constitutes the data in the first half of the subsequent analysis frame A_{n+1}. Notice that the window is structured such that addition of the frames given by $\sum_n A_n$ would restore the original signal s(t). This overlapping is required for two reasons. Firstly, the loss of information due to the deliberate

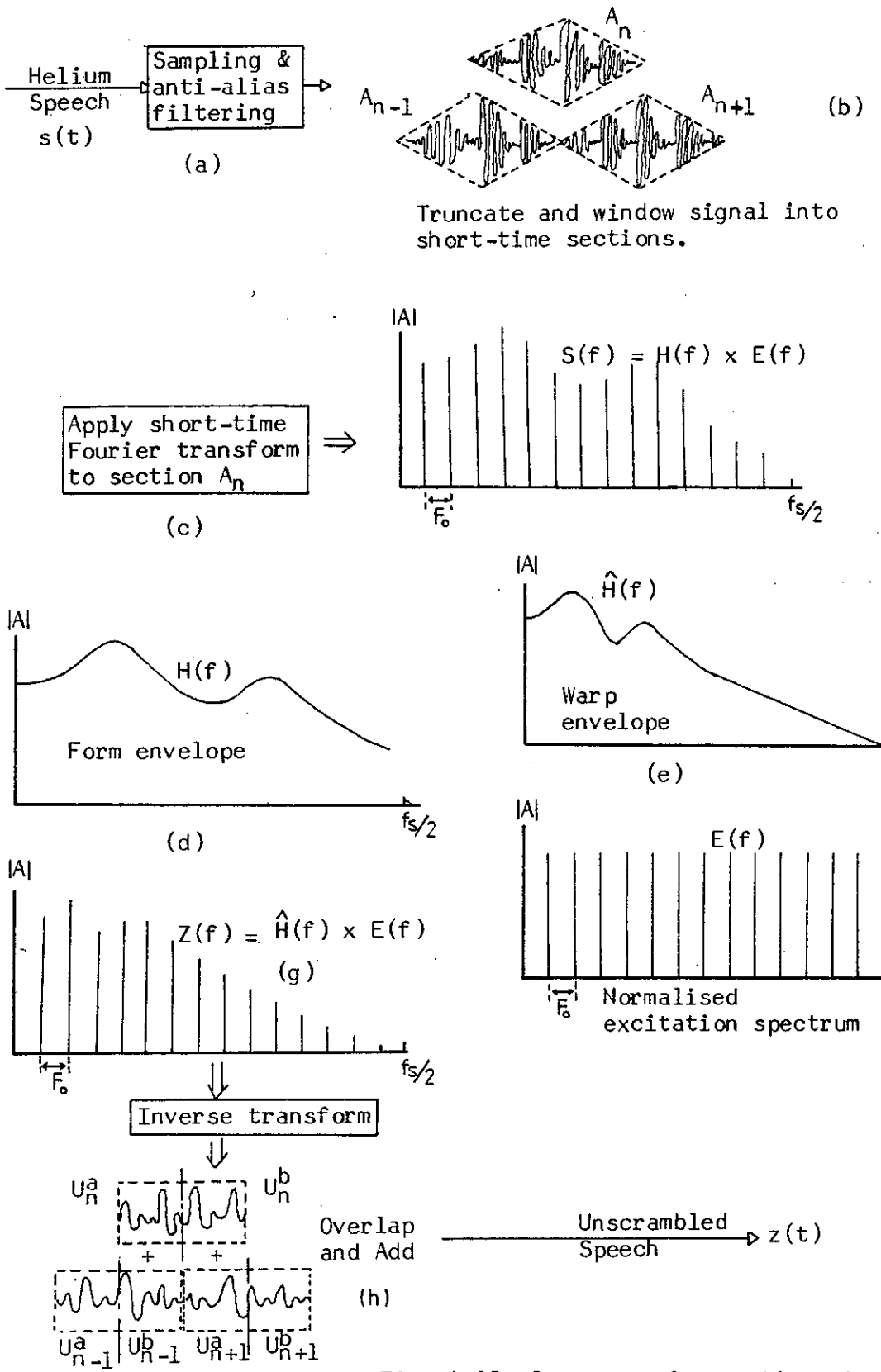


Fig. 4.13 Sequence of operations in helium speech unscrambling using the STFT.

signal deemphasis towards the edges of the window in frame A_n is restored by the emphasis at the window apex in frames A_{n-1} and A_{n+1}. Secondly, whilst the window length is 34.6mS long, new updates in the signal variations can be processed every 17.3mS, thereby improving the trade-off between spectral resolution and length of time window.

Once each frame is transformed into the frequency domain, magnitude and phase spectra are obtained from the complex frequency spectrum. In all of the STFT unscrambler system architectures reported to date, correction for the helium speech distortion is confined to the magnitude spectrum alone, with the phase spectrum being left unchanged. After magnitude spectrum correction, the new warped magnitude is recombined with the retained original phase, and the resulting complex spectrum is then inverse transformed to produce a new signal which is intended to represent the unscrambled intelligible speech. The retention of the original phase is a most important factor which will be shown in sections 4.2.6 and 4.3 to denigrate the performance of this system.

The spectral envelope is therefore derived solely from the magnitude spectrum, and the envelope estimation algorithm is based here on a piecewise-linear curve-fitting method [92]. The basic principle is very similar to that of peak detection with hysteresis in the time domain, as explained in section 4.1.3, except that the method expounded here employs a linear threshold decay characteristic whose slope is dependent on the amplitude of the last peak detected; that is, the normalised slope connecting some candidate spectral line peak to the previously detected spectral line peak (assuming a peak search from 0Hz increasing in frequency up to $f_s/2$) cannot be less than some threshold given by a minimum negative number $\mathcal{E}$.

Let $P_d = S(f_d)$ be the amplitude of the last detected spectral peak

at frequency f_d, and let $P_k = S(f_k)$ be a candidate for the subsequent spectral peak to be detected. P_k is defined to be a peak only if:

$$\frac{P_k - P_d}{f_k - f_d} > P_d \cdot \mathcal{E} \qquad (4\text{-}22)$$

This constraint is intended to restrain the spectral envelope from descending too rapidly to include sidelobe peaks as discussed in section 4.2.2. There is no limit on the maximum rising slope of the envelope in order to allow a fast recovery should a sidelobe or noise peak be detected.

The derived envelope is then divided into the original spectrum to produce an estimate of the vocal tract excitation spectrum, and this is conserved for later use; ensuing operations on the envelope involve correction for the helium speech amplitude distortion according to a predetermined equalisation characteristic dependent on depth (or percent helium), followed by (nonlinear) spectral compression to restore formant amplitudes to their corresponding frequency values in air. The warped envelope is then remultiplied by the stored excitation spectrum to preserve original pitch information and thereafter recombined with the original phase spectrum to produce a new complex spectrum, which is subsequently inverse transformed to produce the unscrambled speech signal. The resulting unscrambled speech frame is not in a suitable state for immediate output, however. With reference to Fig. 4.13, in consideration of the overlapped-frame method of analysis, then only the first half of the current frame of unscrambled speech, U_n^a, is used and added to the (stored) last half of the previous frame, U_{n-1}^b, with U_n^b being stored to await the arrival of frame U_{n+1}^a, and so the resulting unscrambled speech output is produced by a continuous overlap-and-add (OLA) process [182].

4.2.6 SIMULATION AND EVALUATION

For the purposes of the system evaluation presented here, the input signal sample rate is 16kHz. This was used as a matter of being the closest convenient value for f_s to the real time system rate of 14.8kHz, given that the digitised version of "The Rainbow Passage" [121] used for the subjective appreciation of the system simulation and performance is originally sampled at 80kHz, thereby requiring a simple decimation (inclusive of anti-alias filtering) of 5:1 of the sampled speech in this application. The spectral resolution of the FFT used in this case was 31.25Hz, giving an analysis frame length of 32mS (N=512 points) pre-multiplied by a Bartlett window [169] prior to processing, whose equation is given by:

$$w_n = 2.(n + 0.5)/N \qquad 0 \leqslant n \leqslant N/2-1 \qquad (4\text{-}23a)$$

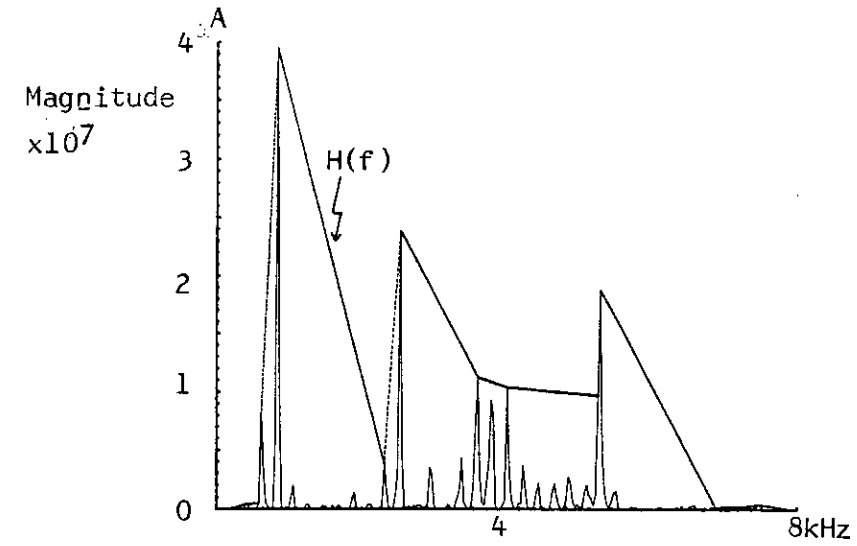
and

$$w_n = 2.(N - n - 0.5)/N \qquad N/2 \leqslant n \leqslant N-1 \qquad (4\text{-}23b)$$

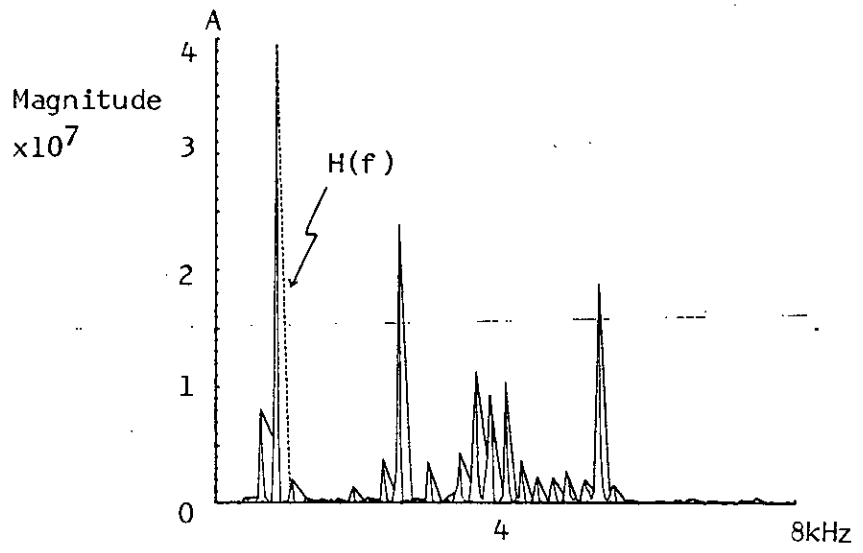
where w_n is the window value multiplying the signal sample s_n within each analysis frame.

With reference to the spectral envelope estimation algorithm of equ.(4-22), the value of the normalised slope threshold $\mathcal{E}$ which was judged to produce the best smoothed fit was $\mathcal{E} = -0.005$. For smaller negative values than this, the algorithm was found to produce a very slow envelope decay with frequency, which effectively missed several likely pitch candidates: on the other hand, large negative values for $\mathcal{E}$ were found to produce inadequate smoothing of the spectrum. An example of the envelope estimation dependence on $\mathcal{E}$ is illustrated in Fig. 4.14(a-c), and the resulting excitation spectrum estimation corresponding to the optimum value of $\mathcal{E}$ is given in Fig. 4.14(d).

Derivation of the new spectral envelope, which has been warped to correct for the helium speech distortion, was carried out by an index



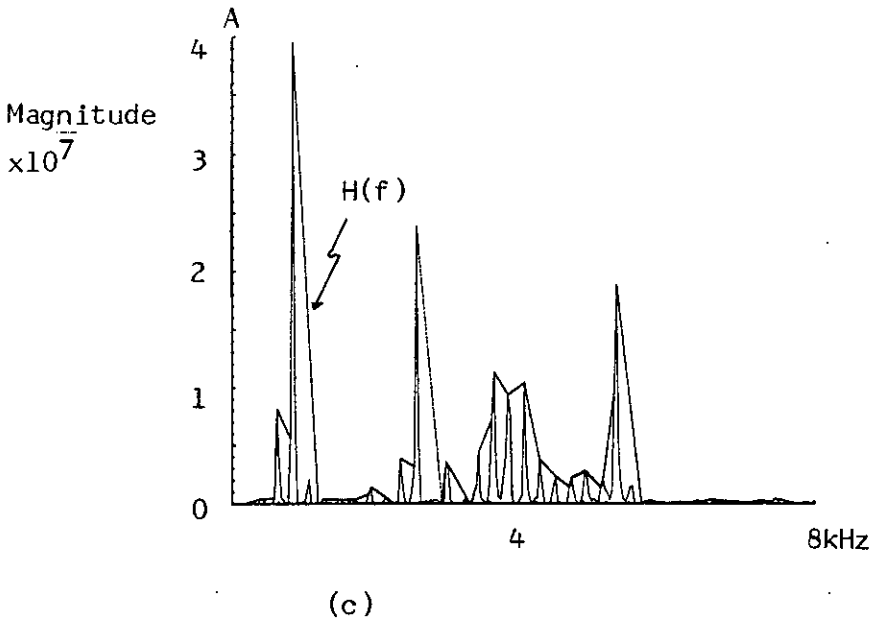
(a)



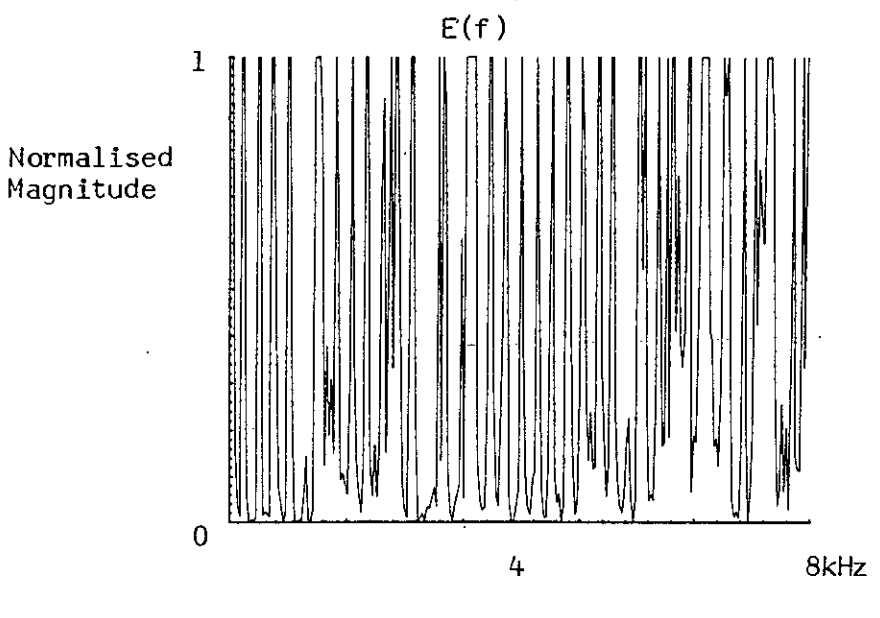
(b)

Fig. 4.14

Continue



(c)



(d)

Fig. 4.14 Voiced speech spectrum illustrating dependence on normalised slope threshold $\mathcal{E}$. (a) excessive smoothing for $\mathcal{E}$= -0.001; (b) inadequate smoothing for $\mathcal{E}$= -0.01; (c) optimum envelope estimation for $\mathcal{E}$= -0.005 and (d) derived excitation spectrum.

mapping and interpolation procedure. Let the heliox spectral envelope frequency index n, up to the aliasing frequency $f_s/2$, be such that $0 \leqslant n \leqslant N/2-1$, and let the new index k corresponding to the compressed envelope be such that $0 \leqslant k \leqslant K$.

Assuming that each frequency index k does indeed correspond to the spectral envelope of normal air speech, then each value of k will have an associated multiplication factor m_k which translates it into a corresponding frequency in the heliox spectrum: if a linear helium frequency translation is assumed, m_k is a constant for all k, whereas if a nonlinear frequency shift, such as that of equ.(2-15) is assumed, m_k will vary with index k; note too that m_k is not necessarily an integer value, but will depend on the percentage of helium and the assumed shift characteristic. Let the heliox frequency index n corresponding to air frequency index k be $\eta = m_k.k$, where n may be non-integer. Since there are only integer heliox index values due to using the DFT, the nearest available heliox index to η is $n = (\text{entier})\eta$, and this is the basis of the index mapping procedure. To allow for the possible non-integer value of η, however, the amplitude envelope for air index k, H(k), is given by linear interpolation between the values of H(n) and H(n+1) in heliox, viz:

$$H(k) = H(n) + \frac{H(n+1) - H(n)}{\eta - n} \qquad (4\text{-}24)$$

Assuming that the same sampling frequency f_s is conserved at the system output, then there is now an undefined region of the spectrum between $k.f_s/N \rightarrow f_s/2$ and its 'image' from $f_s/2 \rightarrow (N-K).f_s/N$. Simply setting this spectral region to zero amplitude is reported to produce spurious artifacts in the resulting time waveform [92], and to avoid this effect, the spectrum is tapered linearly to zero across this region from the last known amplitude value at H(K).

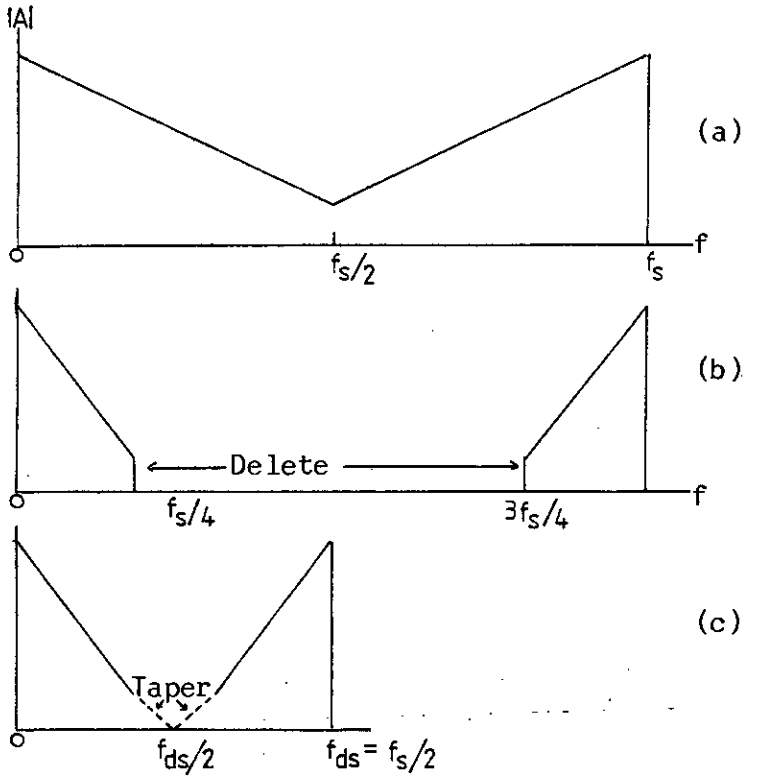


Fig. 4.15 (a) Original spectrum of speech signal sampled at f_sHz; (b) spectrum corrected for helium speech distortion; (c) redefinition of frequency space to permit output of unscrambled speech downsampled to f_{ds}Hz.
Note linear tapering over undefined frequency space.

In cases where the overall spectral compression is such that the maximum frequency index in air, $K > (N/2)/2$, then the original input sampling rate f_s is maintained at the output. In the case where $K < (N/2)/2$, then it is possible to automatically inverse tranform the spectrum to produce a signal whose output sample rate is now $f_{ds} = f_s/2$. This can be done by basically redefining the frequency space as being 0Hz to $f_{ds} = f_s/2$Hz; the remaining points in the original frequency space between $f_s/4$ and $3f_s/4$ are deleted [92]; here, spectral tapering of the undefined region is effected from $K.f_s/N$ to $f_{ds}/2 = f_s/4$, and similarly on the magnitude spectral 'image', and this procedure is shown in Fig. 4.15. Note that since the processing applies exclusively here to

real (i.e. non-complex valued) data, then it is possible to implement a modified form of the FFT which effectively obviates a consideration of the 'image' half of the spectrum, thereby saving storage space in a digitally-implemented system: the fundamentals of this algorithm are outlined in section 5.2.1. For convenience, examples of spectra shown here are confined to the low-frequency spectrum between 0Hz and $f_s/2$.

In order to obtain a subjective comparison between the STFT technique and that of the time domain unscrambler, a linear spectral compression, assuming a heliox: air velocity ratio R = 1.84 at 100ft, has been applied to helium-speech from "The Rainbow Passage" [121] and may be heard in recording A7. A comparison of this recording against rec. A6(a-b) for the time domain device demonstrates a significant improvement in intelligibility and ease of listening, in that there are less breaks in the speech output. Recording A8 demonstrates unscrambled speech output from a depth of 300ft, again using only linear frequency compression, with an assumed velocity ratio R of 1.89, being the average of the piecewise linear formant shift curves obtained in

al analysis of voiced speech presented in section 3.4.2, see

It can be appreciated from rec. A8 that the unscrambled

of poor quality. Since it is difficult to assess whether this

n in quality is due to the apparent high level of background

hether there may be contributions from the operations invol-

signal processing itself, further evaluation of the STFT

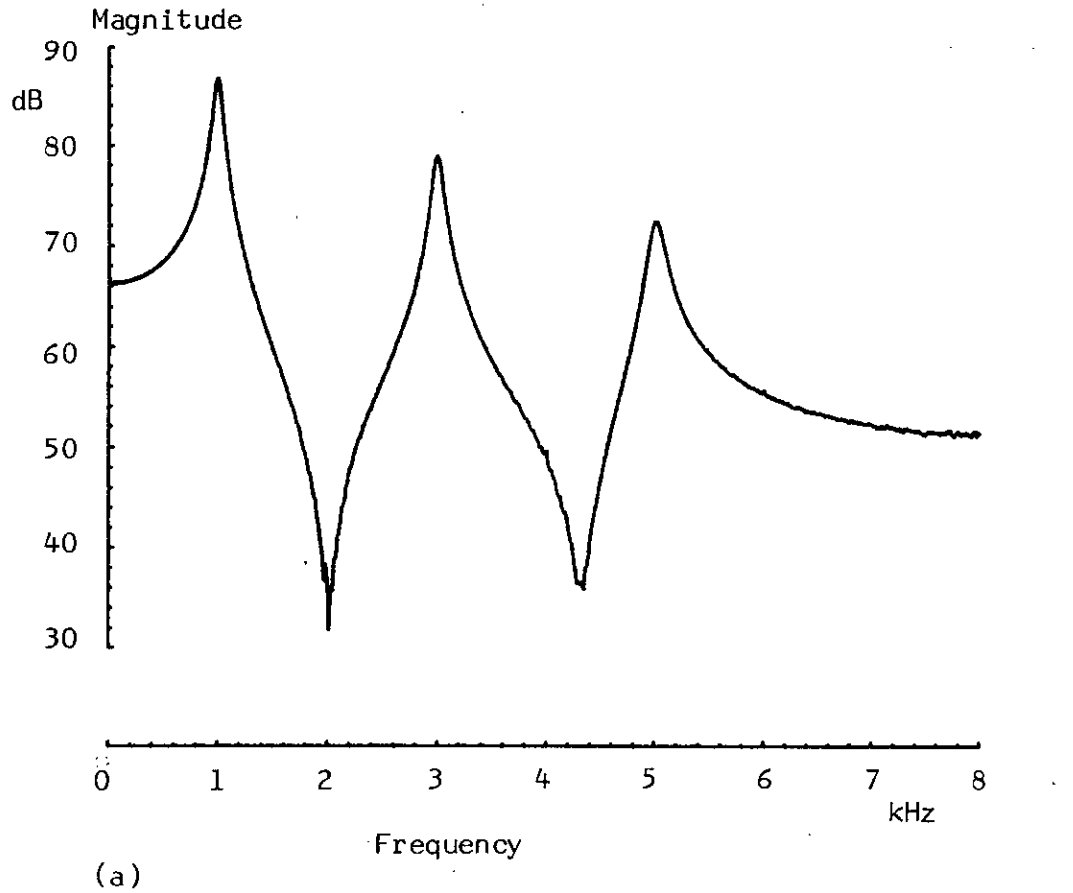
requires the use of a well-defined signal set. In this case,

used was generated by a pulse-excited all-pole (simulated)

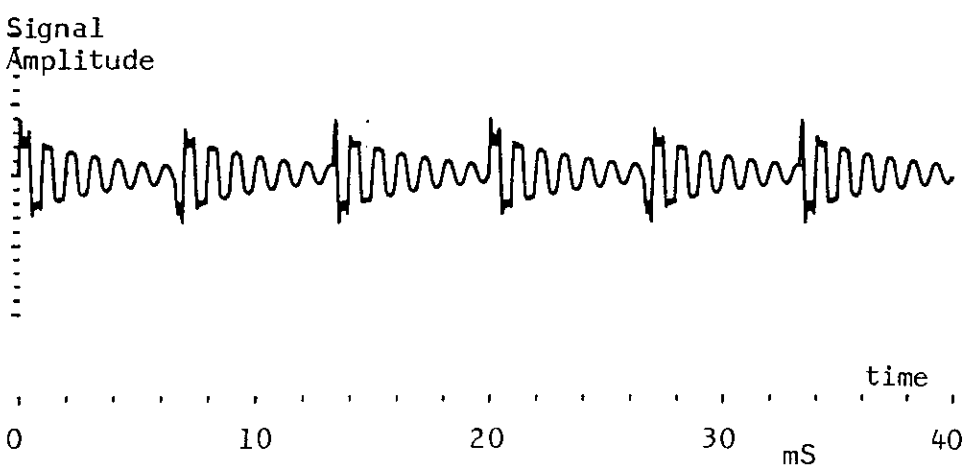
ter consisting of 3 complex-conjugate pole pairs, whose

ing frequency response or spectrum H(f) of the filter transfer

s shown in Fig. 4.16(a), with the waveform produced by



(a)

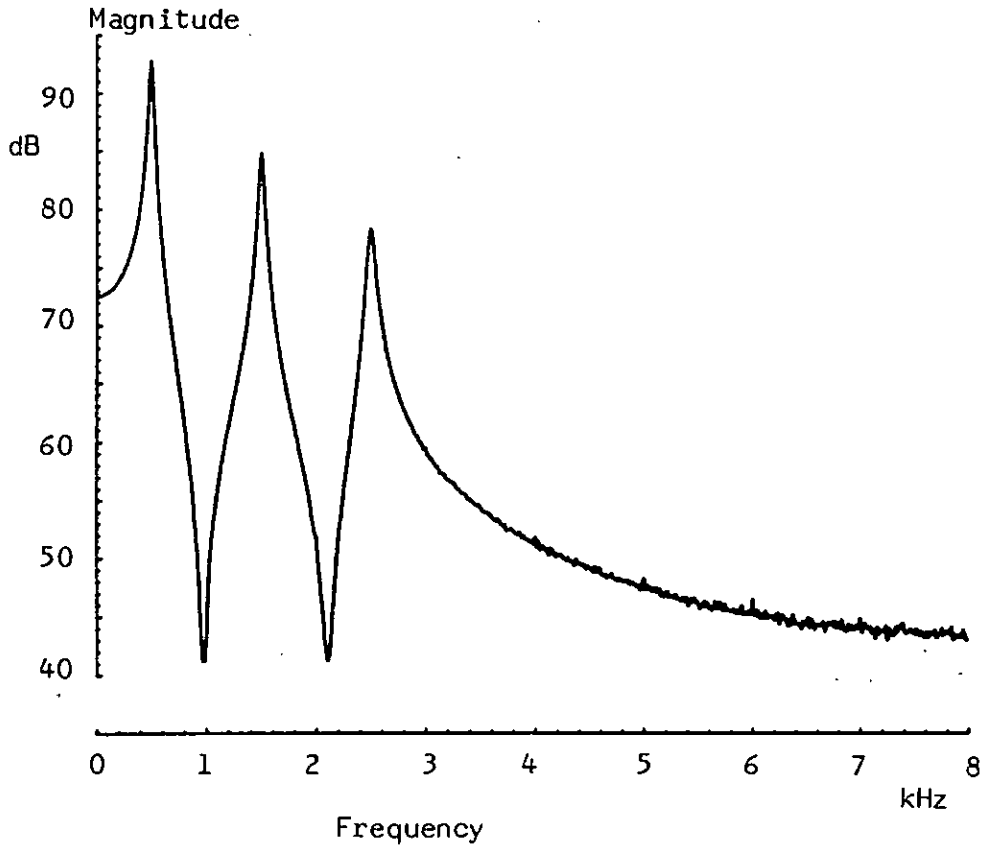


(b)

Fig. 4.16 (a) Frequency response of an LTI filter system consisting of 3 complex-conjugate pole pairs, giving resonances at 1kHz, 3kHz and 5kHz.
(b) Pulse-excited filter output (Fo = 150Hz).

periodic pulse excitation at a frequency Fo = 150Hz shown in Fig. 4.16(b). Note that the magnitude spectrum of H(f) contains formants at 1kHz, 3kHz and 5kHz, with relative amplitudes of nominally 0dB, -6dB and -9dB, and bandwidths of 80Hz, 100Hz and 150Hz respectively. The output of this (simulated) analog filter was sampled at 16kHz commensurate with the STFT simulated system input, and a recording of some 5 seconds of this signal can be heard in rec. A9. For the purposes of the present evaluation, consider only a linear shift of the spectral envelope by a compression ratio R = 2.0, assuming commensurate formant bandwidth compression with retention of relative amplitude ratios. In terms of the linear time-invariant (LTI) analog filter system, this amounts to relocating each pole pair in the system to produce the frequency response of Fig. 4.17(a). Assuming that the fundamental frequency of excitation Fo remains constant, then the resulting signal, downsampled to 8kHz, is shown in Fig. 4.17(b) and can be heard in rec. A10.

Consider now the application of STFT unscrambler processing to achieve the same frequency compression of x2. The spectrum of the signal in Fig.4.16(b) is obtained and the spectral envelope estimated as in Fig. 4.18(a). The envelope is extracted and used to produce the excitation spectrum of Fig. 4.18(b) and then compressed as in Fig. 4.18(c). The resulting spectrum on recombination of envelope and excitation spectra is shown in Fig. 4.18(d), and the final step before the inverse transformation involves remultiplication into the original phase spectrum, which is shown in Fig. 4.18(e). The resulting time waveform, as produced by the overlap-and-add (OLA) method [182], is shown in Fig. 4.19(a), and is seen visually to be very different in time structure to the ideal waveform produced by system pole relocation and retention of the pulse excitation frequency as shown in Figs. 4.17(b)



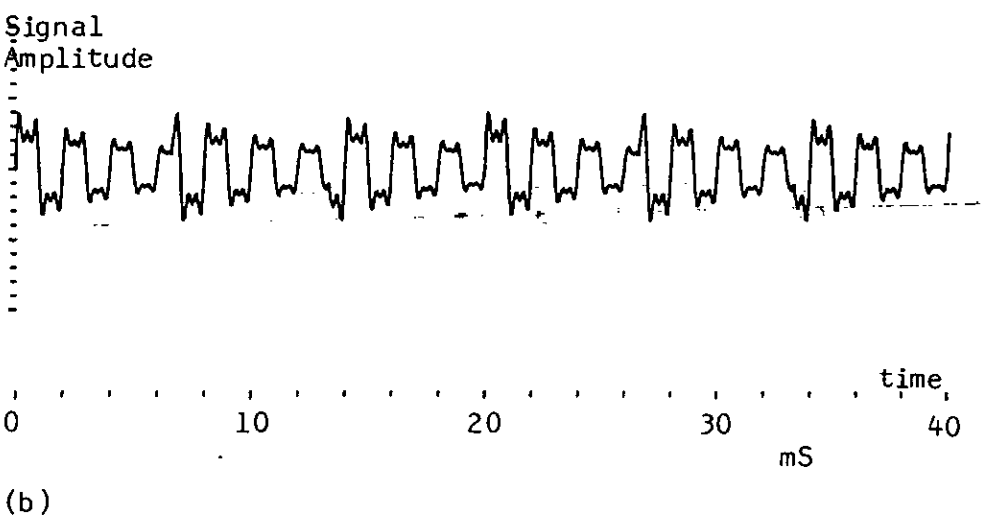
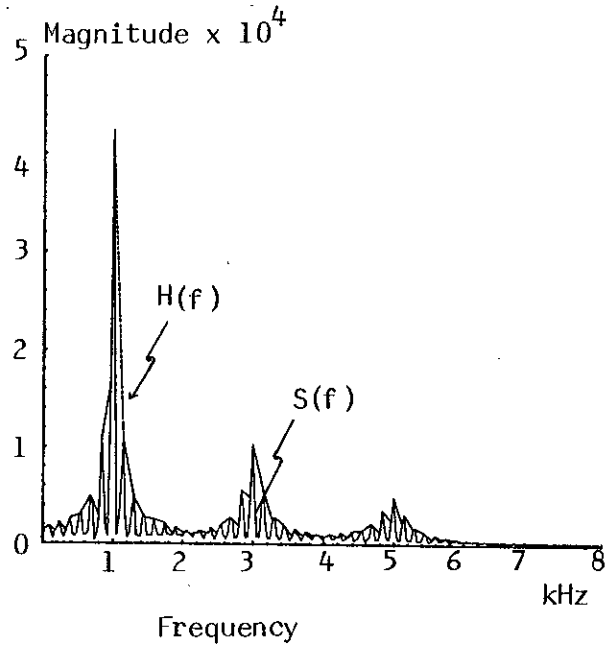
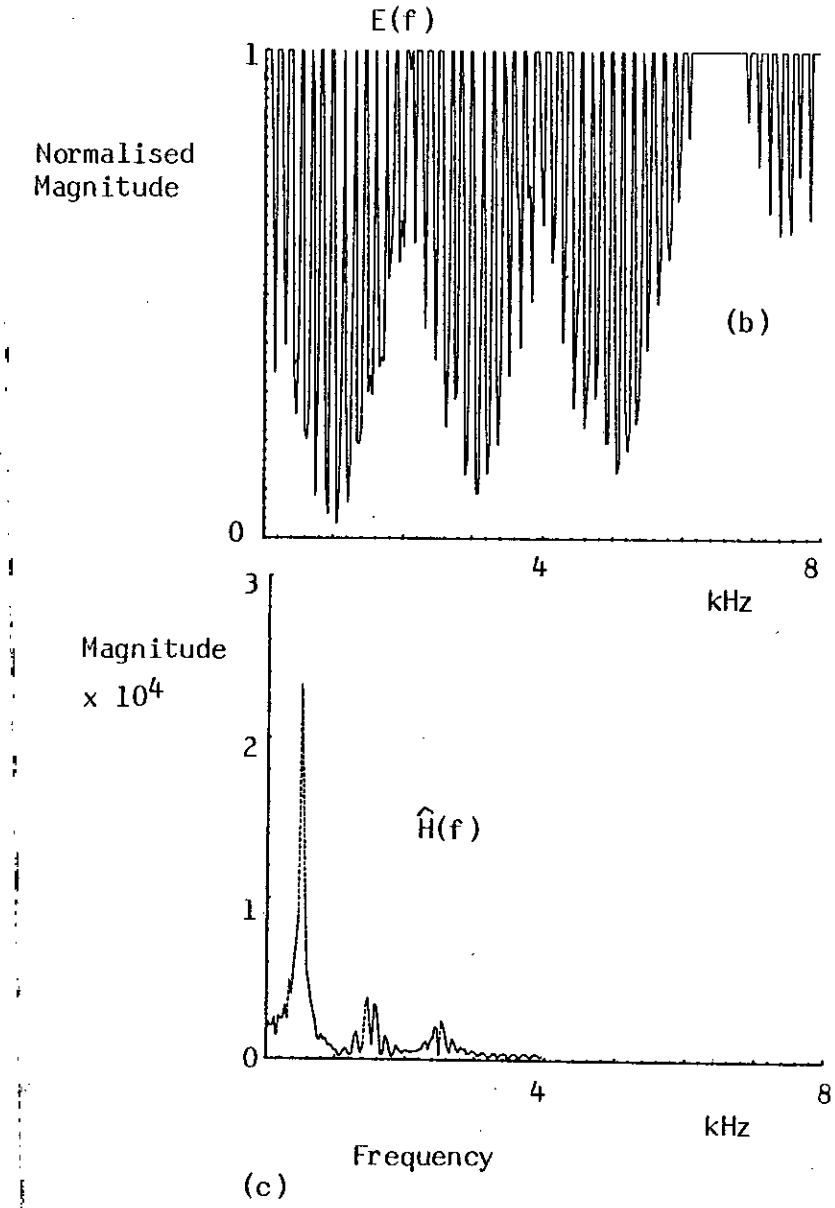


Fig. 4.17 (a) Frequency response of an LTI filter system consisting of 3 complex-conjugate pole pairs, giving resonances at 0.5kHz, 1.5kHz and 2.5kHz. (b) Pulse-excited filter output (Fo = 150Hz).


Magnitude x 10^4
5
4
3
2
1
0
H(f)
S(f)
1
2
3
4
5
6
7
8
kHz
Frequency

(a)

Fig. 4.18


E(f)
Normalised
Magnitude
1
0
4
8
kHz
(b)
Magnitude
x 10^4
3
2
1
0
$\hat{H}(f)$
4
8
kHz
Frequency

(c)

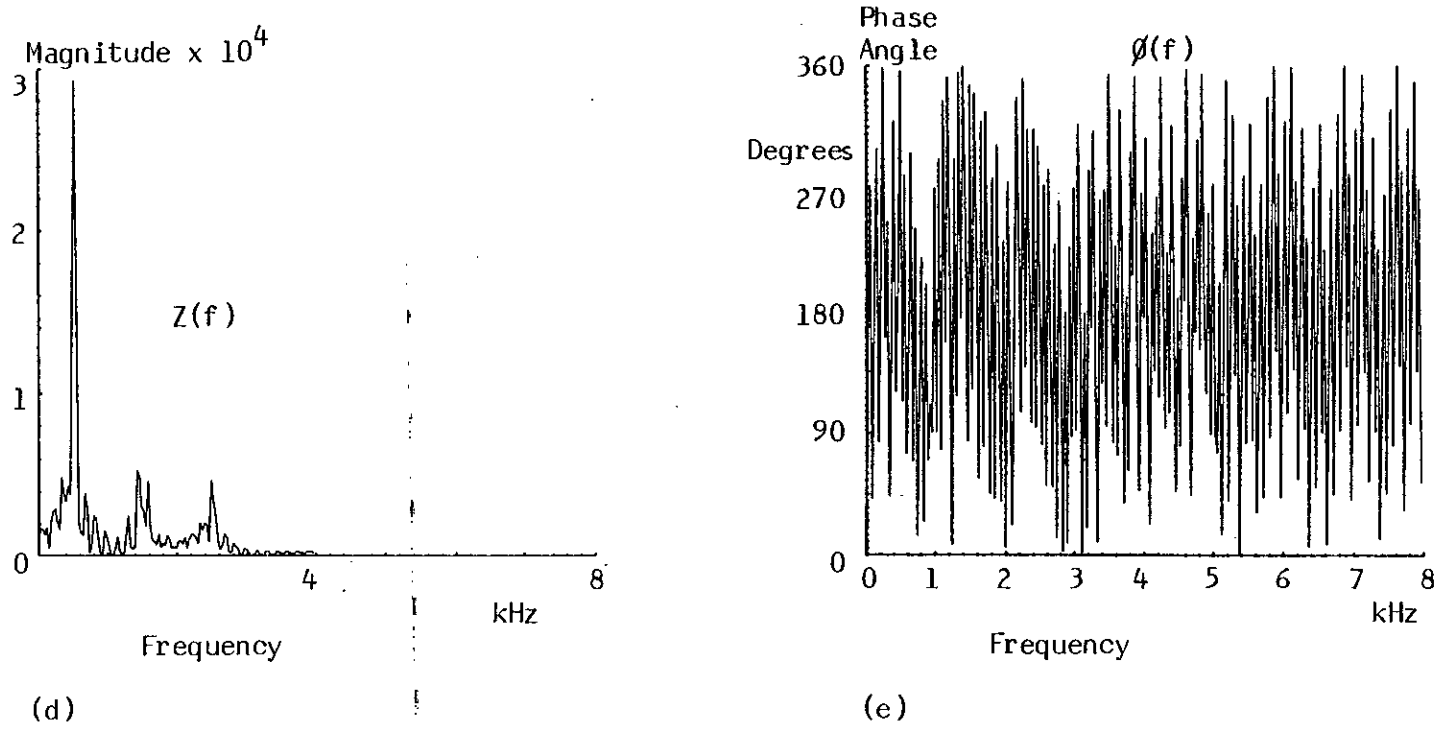


Fig. 4.18 (a) Estimation of spectral envelope H(f) from magnitude spectrum S(f). (b) Derived excitation spectrum E(f). (c) Compressed spectral envelope $\hat{H}(f)$. (d) Resulting magnitude spectrum Z(f) = $\hat{H}(f)$ x E(f). (e) Phase spectrum Ø(f).

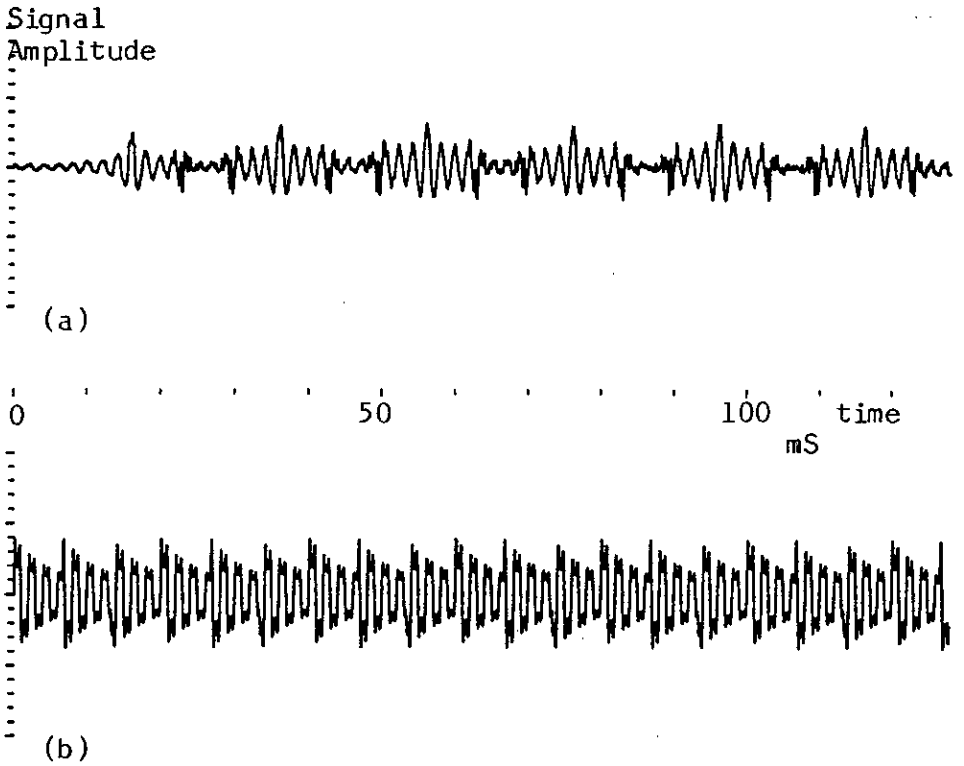


Fig. 4.19 (a) Signal resulting from spectral compression and resynthesis by the STFT of the system waveform of Fig. 4.16(b).
(b) Ideal waveform produced by system pole relocation and retention of pulse excitation frequency Fo.

and 4.19(b). Note that there is no tapering of the spectrum here, since the spectrum is compressed identically by x2, and therefore can be resampled at 8kHz by redefining the frequency space as outlined above. The resulting signal can be heard in rec. A11, and a comparison between recs. A10 and A11 demonstrates that the two waveforms sound subjectively very different, with rec. A11 sounding distorted compared to the ideal case of rec. A10 which corresponds to the pulse-excited LTI system. Since this distortion is expected to exist in the unscrambled speech from the STFT system, then it is important to establish its root cause with a view to suggesting improvements to eliminate its mechanism from the unscrambler system.

The only apparent feature of the signal spectrum which has been deliberately modified has been the spectral envelope H(f), and so it is reasonable to assume that there are deficiencies in the initial envelope estimation as produced by the piecewise-linear method detailed in section 4.2.5. To explore this proposition, autoregressive (AR) spectral analysis [183] was employed to furnish the optimum estimation of the spectral envelope (see section 3.3) using a 10th order model approximation (no applied preemphasis) and an autocorrelation window of 512 points. The resulting time waveform, using this hybrid technique involving AR envelope estimation and STFT spectral correction and waveform synthesis, was virtually identical to that of Fig. 4.19(a), and therefore estimation of the spectral envelope can be waived as a likely cause of the signal distortion.

There is, however, one important aspect of the spectrum which has so far received little attention. The phase spectrum in the STFT system is left unaltered throughout the unscrambling operation, therefore the warped spectral envelope is combined not only with the original excitation spectrum, but also with the original phase spectrum of the signal. The reason for permitting phase retention is that human speech perception is based mainly on magnitude spectrum information [10], and therefore as far as the phase information is concerned, "since the ear is known to be relatively insensitive to moderate phase distortion, the simple approach of doing nothing has been taken" [92]. There is an important dichotemy of principles here, however. Whilst the speech perceptual system within the human brain does indeed appear to depend mainly on spectral magnitude [10], transduction of the acoustic waveform of the ear [18], however, depends on the signal phase characteristic [184]. Phase retention is the most important aspect of the STFT

system affecting signal intelligibility, and this topic is investigated at length in the discussions in section 4.3 which follow.

4.3 DISCUSSION AND CONCLUSIONS

The principal advantages of the time domain helium speech unscrambler system relate to its cost-effectiveness and realisability as a real time device. It is also readily amenable to miniaturisation [108], and therefore its resulting small physical size and low power consumption have pioneered the inception of a diver-borne device to permit unscrambling in situ, thereby reducing the signal bandwidth required for through-water speech transmission. It is, however, limited in performance in that it can only provide linear frequency compression of the speech spectrum, with a token correction of spectral amplitude provided by analog filtering and preamplification at the input to the device, whereas it has been shown in section 3.4.2 that, at the very least, a piecewise linear spectral compression of helium speech is necessary, and from section 3.4.5 it was shown that spectral amplitude correction is a function of both frequency and depth (or percent helium).

The principle disadvantage of the time domain device, however, relates to its dependence on pitch detection, by peak threshold triggering, which is directly responsible for the discontinuity of the unscrambled speech, to the detriment of intelligibility. To demonstrate this effect, the (simulated) technique was modified such that if there was no speech awaiting time expansion within the device, that is, if a pitch peak had not been detected in the time taken to clear all expansion buffers, then the input signal was routed to bypass the

expansion circuitry (although still presented to the pitch detector) and so pass directly to the system output, being simultaneously down-sampled commensurate with the actual output sampling frequency. The results of this modification are demonstrated in recordings A12(a-b), where rec. A12(a) is the original (simulated) device output for helium speech unscrambled at 100ft, and rec. A12(b) is the system output inclusive of the above modification. The resulting speech is clearly easier to interpret and therefore by extension more intelligible. However, application of the modified (simulated) system to helium speech at 300ft depth may be heard in rec. A12(c), and suffers a severe degradation in intelligibility compared to rec. A12(b). This is considered to be due to the apparently lower S/N ratio of the speech, in that whilst audibility of the signal is aided by continuity, intelligibility is degraded by the tonal change in ambient noise, whose frequency components are compressed on the one hand during active time expansion, but left unchanged on the other hand during idling of the expansion operation, so effectively distracting the human perceptual mechanism and thereby masking the speech with intermittent noise. Figure 4.20 demonstrates the effect on intelligibility of alternating intervals of equal length of speech and white noise [185]. The intelligibility is measured in terms of the percentage of words heard correctly, and is plotted against frequency of interruption, which ranged from 0.1Hz to 10kHz, and the signal-to-(interruption)noise ratio was varied from +9dB to -18dB. Shown also is the response, marked 'Quiet', where the intervals between speech were silent, with no added noise.

It can be seen from Fig. 4.20 that for a noise interruption rate in the range 10-560Hz, and for all S/N ratios, intelligibility is increasingly impaired. The pitch frequencies of most male and female talkers

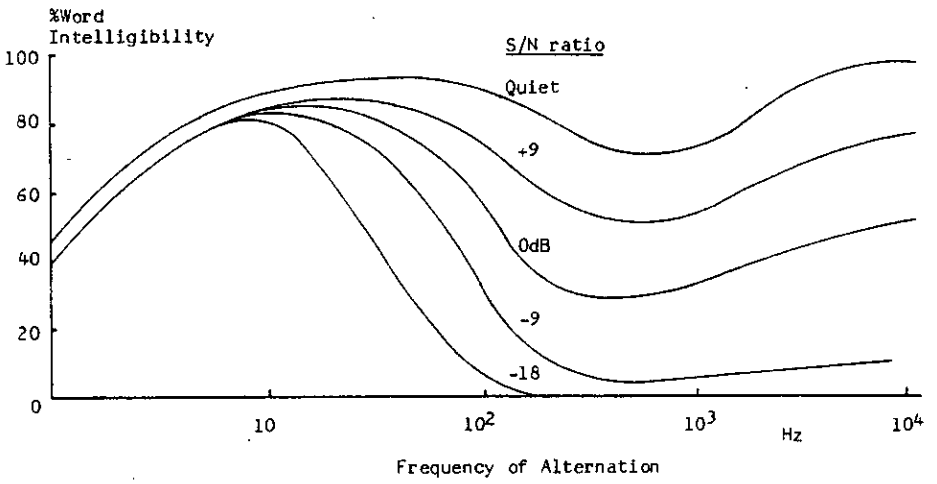


Fig. 4.20 Word intelligibility as a function of the frequency of alternation between speech and noise, with signal-to-noise ratio as the parameter (from (185)).

fall within this range [32]; hence if a succession of unvoiced pitch decisions are made in error of actual voiced speech sounds, then the voiced speech will be effectively masked in noise as the pitch detector malfunction causes either no output to be produced in the actual system or throughputs uncorrected voiced helium speech/noise directly to the output in the modified system simulated here. Thus, whilst signal continuity is an important aspect of intelligibility, care must be taken to ensure that, at the very least, each portion of the input signal is processed in exactly the same manner to avoid inadvertent masking of the unscrambled speech.

Signal continuity and homogeneity of processing are a feature of the short-time Fourier transform unscrambler system [92,172], and additionally, signal integrity is conserved with no requirement to discard portions of the input speech signal. Although the real time implementation

of this complex system excludes presently a diver-borne unscrambler system and so necessitates conservation of helium speech bandwidths for transmission, the frequency domain-based rationale of this technique permits ease of nonlinear spectral correction of both frequency and amplitude for the helium speech distortion. The voiced/unvoiced nature of the speech is intrinsically conserved with no explicit pitch detection required whatsoever, so preserving signal continuity.

The principal limitation of this system, however, relates to the retention of the original phase spectrum and its subsequent recombination with the (co-phase) magnitude spectrum prior to execution of the inverse Fourier transform to produce the unscrambled speech. Consider (a) the phase characteristic of Fig. 4.21(a) which corresponds to the signal produced by a pulse-excited all-pole LTI filter which exhibits a 3-formant spectrum with resonances at 0.5kHz, 1.5kHz and 2.5kHz (see Fig. 4.17(b)), and compare this against (b) the phase spectrum of Fig. 4.21(b) of the signal from a similar all-pole LTI filter but with formants at 1kHz, 3kHz and 5kHz (see Fig. 4.16(b)) as outlined in the evaluation experiments of section 4.2.6. In order to produce a magnitude spectrum corresponding to the signal in case (a), the magnitude spectral envelope corresponding to the signal of case (b) would be compressed and then ultimately recombined with the original phase spectrum of Fig. 4.21(b) in an effort to produce the waveform of Fig. 4.17(b). The ideal phase spectrum of Fig. 4.21(a) is, however, evidently very different in form to that of Fig. 4.21(b) corresponding to the waveform of Fig. 4.19(a) as synthesised by the STFT unscrambler. Since the basic assumption in the use of the STFT system refers to a linear time-invariant model of the speech mechanism, then the phase spectrum of the spectrally-compressed signal of Fig. 4.19(a) should in theory correspond to

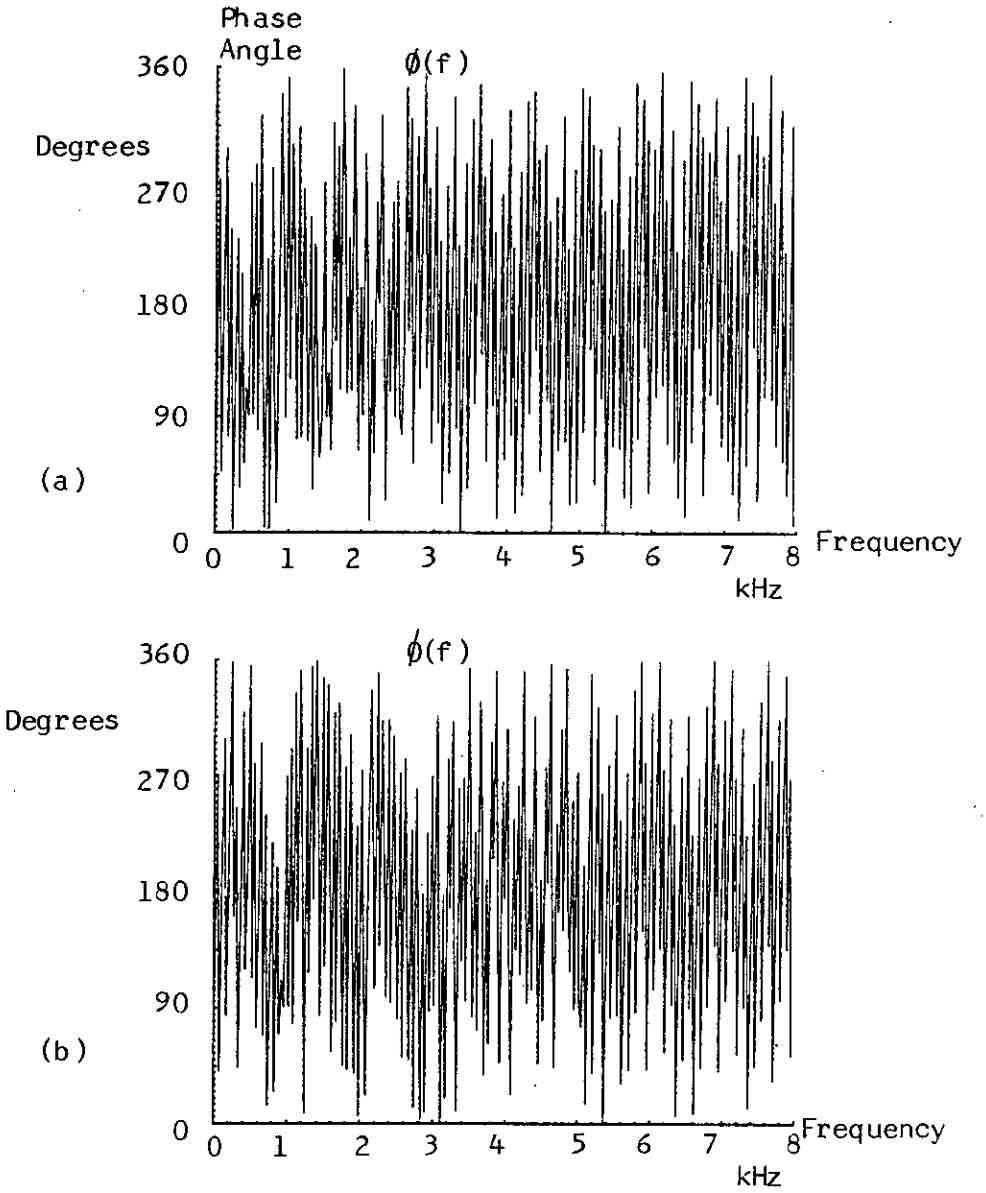


Fig. 4.21 (a) Phase spectrum corresponding to the ideal waveform produced by the pulse-excited LTI system with relocated poles, see Figs. 4.17(b) and 4.19(b).
(b) Phase spectrum corresponding to the LTI system waveform of Fig. 4.16(b) and also used to synthesise the STFT waveform of Fig. 4.19(a).

the ideal LTI phase spectrum of Fig. 4.21(a). The reason offered for retention of the original phase spectrum relates to the apparent insensitivity of the ear to spectral phase, see section 4.2.6. This maxim, however, is an oft-misquoted version of de Boer's rule [186]. The complete statement of this rule asserts that :

"the timbre (or apparent frequency content) of a sound does not change when the phases of the components are shifted by a constant amount and/or by amounts that are linearly dependent on frequency".

In the case of Fig. 4.21, it can be considered that the resulting spectrally-compressed signal of Fig. 4.19(a) may have been produced with the correct phase as in Fig. 4.21(a), but is then passed through a phase-shifting network, prior to output, whose phase characteristic is non-uniform with frequency, resulting in the phase spectrum of Fig. 4.21(b). The following extract summarises the importance of relative phase on the perceived timbre of the acoustic signal [186]:

"... Thus when the phase $\phi(f)$ of the component with frequency f is changed by an amount:

$$\phi(f) = a + \ell.f \qquad (a \text{ and } \ell \text{ being constants}) \qquad (4\text{-}25)$$

no change of the timbre will occur. The evidence for this rule comes from experiments from signals of which the components have equal spacings. Such signals, be they harmonic or not, have a steady sound quality.

The term $\ell.f$ in the formula is easily understood. In fact it represents a simple shift in time equal for all components. For a steady sound this is, of course, not noticeable. The important term is the constant a, expressing a constant phase shift of all components. The phases of the components can all be changed by the same amount without bringing about a change in timbre. The experiments, on which this statement is based, are fundamentally carried out as follows. A harmonic sound complex is set up by a modulation process. A carrier of, say, 1.6kHz is modulated by a signal that contains components of, e.g., 200, 400 and 600Hz. Note that the components of the modulating signal are to be harmonically related. The modulating signal thus has a fundamental frequency of 200Hz. If now the carrier is also a harmonic of 200Hz, the resulting signal will be purely periodic. It then consists exclusively of harmonics of 200Hz, namely 1.0, 1.2, 1.4, 1.6, 1.8, 2.0 and 2.2 kiloHertz. This signal sounds as a sharp tone of a pitch of 200Hz.

If the carrier frequency (prior to modulation) is now shifted a very small amount in frequency, all components of the resulting signal incur the same frequency shift. We can describe this by saying that, with

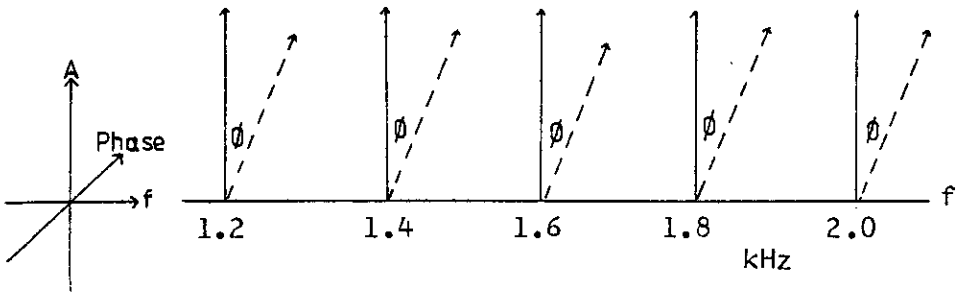


Fig. 4.22 Signal of which the components undergo a common phase shift (from (186)).

respect to their former positions, the components have acquired a common phase shift that increases proportionally with time (Fig. 4.22). Various values of this common phase shift thus alternately present themselves. If such a common phase shift produced an audible change, we would hear beats (in this case slow variations of timbre). Nothing of the kind occurs, however, and we must conclude that such a uniform phase shift is not detectable. We have studied this phenomenon for many frequency ranges, and the rule seems to hold always. The exceptions could all be traced down to artefacts giving rise to non-uniform phase shifts."

Complementary research has supported this phenomenon [187,188,189], and the consensus of opinion is that changes of phase in a signal which are nonlinear with frequency alter the envelope of the time waveform, which creates extraneous frequencies when transduced by the ear, thereby generating an 'internal' magnitude spectrum [190] in the human perceptual system which may be very different to that obtained by a straightforward machine-generated Fourier analysis of the signal.

Although approached from a different point of view, dependency of the time waveform on spectral phase is reinforced by experiments involving signal reconstruction from either phase or magnitude spectra alone [191,192,193], which implies that the phase spectrum contains significant information relating to the frequency content of the signal [194], and therefore that magnitude and phase spectra of LTI systems are in essence unique to each other [176].

In conclusion, helium speech correction based on frequency domain processing offers continuity of the unscrambled speech through intrinsic conservation of the vocal tract excitation spectrum, thereby improving the naturalness and intelligibility of the resulting speech waveform. Application of nonlinear magnitude spectrum correction algorithms can be achieved with relative ease, which again offers the potential of greatly improving the intelligibility of the unscrambled speech. However, the principal enigma concerning spectral correction techniques regards the inadequacy of information in respect of the correct spectral phase characteristic of the unscrambled speech signal, commensurate with the assumption of an LTI model of the speech mechanism. The design of a helium speech unscrambler system which will provide an enhanced fidelity of the unscrambled speech beyond that of the STFT system must therefore seek to resolve the relationship between spectral magnitude and phase without a sacrifice of the desirable system features detailed above, and this challenge is pursued in Chapter V.

CHAPTER V

CORRECTION OF THE HELIUM SPEECH EFFECT BY AUTOREGRESSIVE SIGNAL PROCESSING.

The desirable properties of a helium speech unscrambler system in terms of its processing methodology have been enucleated in Chapter IV. Continuity of speech output from the system has been identified as an important criterion which can be fulfilled by the use of frequency domain-based unscrambling techniques coupled with implicit conservation of the vocal tract excitation characteristic. Helium speech correction involving the manipulation of frequency domain parameters via the short-time Fourier transform (STFT), inherently assumes that the human speech waveform is produced by a vocal mechanism which may vary in the long term, but can be essentially considered as a linear time-invariant (LTI) filter system in the short term over some 10-30mS. It is this assumption which simplifies the transformation from the time waveform to its equivalent frequency domain spectrum representation. Nonlinear correction algorithms can then be applied with relative ease to the resulting spectrum to remove the helium speech distortion, with resynthesis of the unscrambled speech waveform via the inverse transformation [92,195].

Nevertheless, use of the STFT in itself as the direct processing vehicle in the accomplishment of helium speech correction has been shown to incur certain inadequacies: since the vocal tract excitation is considered invariant from air to heliox, spectral correction must necessarily be confined to the spectral envelope only, and the performance of the piecewise-linear method of envelope (and hence excitation spectrum) extraction was shown in section 4.2.6 to be dependent on the

value chosen for the normalised slope factor. A further problem was identified from the recombination of the corrected STFT magnitude spectrum with the original phase spectrum of the helium speech in order to synthesise the time waveform corresponding to the unscrambled speech. Phase correction was not effected since it was proposed that the ear is insensitive to phase, with the human speech perceptual mechanism depending mainly on the spectral magnitude characteristic. Unconditional auditory phase insensitivity, however, has been shown to be a false premise [186,190], with the consequence that original phase retention in this application may affect the quality of the unscrambled speech.

In this chapter, a new residually-excited linear predictive coding (RELPC) unscrambling technique [196] is developed based on the consideration of the speech mechanism as an LTI filter system which is invariant in the short term. The processing architecture of the system detailed here involves both autoregressive (AR) modelling [197] and the Fourier transform [179], and therefore the reader is recommended to refer to sections 3.3, 4.2.1 and 4.2.3-4 if unfamiliar with these signal processing techniques.

In the RELPC unscrambler system, firstly an inverse or prediction error filter (p.e.f.) is formed from both short term stochastic and spectral characteristics of the speech waveform, and this inverse filter is then used to derive the vocal tract excitation waveform (or residual signal) without recourse to direct pitch extraction techniques. The residual is subsequently conserved while new p.e.f. parameters are calculated to produce a synthesis or autoregressive filter (a.r.f.), corresponding to a vocal tract transfer function which has been corrected for the helium speech distortion. The residual signal is then used to excite the a.r.f. to produce the time waveform corresponding to normal

air speech. The philosophy of this approach is that ease of nonlinear correction of the helium speech effect is maintained through parameterisation of the speech signal. Moreover, the phase characteristic of the unscrambled signal is automatically regulated by a treatment of the unscrambling task as a problem involving short term linear time-invariant filters, and in particular by production of the unscrambled speech signal by direct excitation of the autoregressive filter.

5.1 BASIC PRINCIPLES OF THE RELPC SYSTEM

5.1.1 DERIVATION OF THE RESIDUAL EXCITATION FUNCTION

The first step in helium speech unscrambling using the RELPC technique involves derivation of the vocal tract excitation function, or residual, by inverse filtering of the helium speech waveform. The residual signal is stored and conserved for later use in the synthesis section of the unscrambler system.

Inverse filtering is achieved here by autoregressive modelling of short term segments of the helium speech signal. As has been detailed in section 3.3, the assumption in the construction of the p.e.f. is that the speech mechanism is considered to be a LTI filter system in the short term, excited by either a train of Dirac impulses for voiced speech or by Gaussian noise for unvoiced speech. Therefore, the inverse filter coefficients, derived from stochastic properties of the speech signal, must reflect information about both the vocal tract transfer function and source spectrum characteristics, but as one convolved entity. The residual signal contains information regarding the rate of vocal tract excitation only (assuming perfect inverse filtering): that is the fundamental frequency (if any) and type of speech (voiced/unvoiced) [198].

Mathematically, the problem is expressed by equs.(3-1) and (3-2), which are restated here to emphasise their importance.

Given the (sampled) p.e.f. impulse response h(nT) with frequency response H(mF) and input helium speech signal s(nT) with spectrum S(mF), then the residual signal e(nT) with spectrum E(mF) is given, from convolution theory, by:

$$s(nT) * h(nT) = e(nT) \rightleftharpoons S(mF) \times H(mF) = E(mF) \qquad (5\text{-}1)$$

and, from the right hand side of equ.(5-1),

$$S(mF) = E(mF)/H(mF) \qquad (5\text{-}2)$$

Calculation of the p.e.f. coefficients associated with h(nT) is performed through the properties of linear prediction [199] and Levinson recursion [137] precisely as set forth in section 3.3.2.

In terms of the RELPC unscrambler system, a most important attribute of AR signal analysis concerns the spectral characteristic of the residual, This was shown, in section 3.3.3, to be flat (white) and equal to some constant K for every frequency component existing in the residual spectrum. It can therefore be appreciated, from equ.(5-2), that manipulation of the p.e.f. coefficients is equivalent spectrally to manipulating H(mF) to produce a new frequency response $\hat{H}(mF)$. Therefore if E(mF) = K is left unchanged, then a new speech signal spectrum $\hat{S}(mF)$ corresponding to the unscrambled helium speech $\hat{s}(nT)$ is given by:

$$\hat{S}(mF) = K/\hat{H}(mF) = K \times B(mF) \qquad (5\text{-}3)$$

where B(mF) is the frequency response of the AR (synthesis) filter. Thus, the unscrambled speech signal $\hat{s}(nT)$ is given, in the time domain, by:

$$\hat{s}(nT) = e(nT) * b(nT) \qquad (5\text{-}4)$$

where b(nT) is the impulse response of the AR filter. The corresponding sequence of operations for the basic RELPC system is shown in Fig. 5.1.

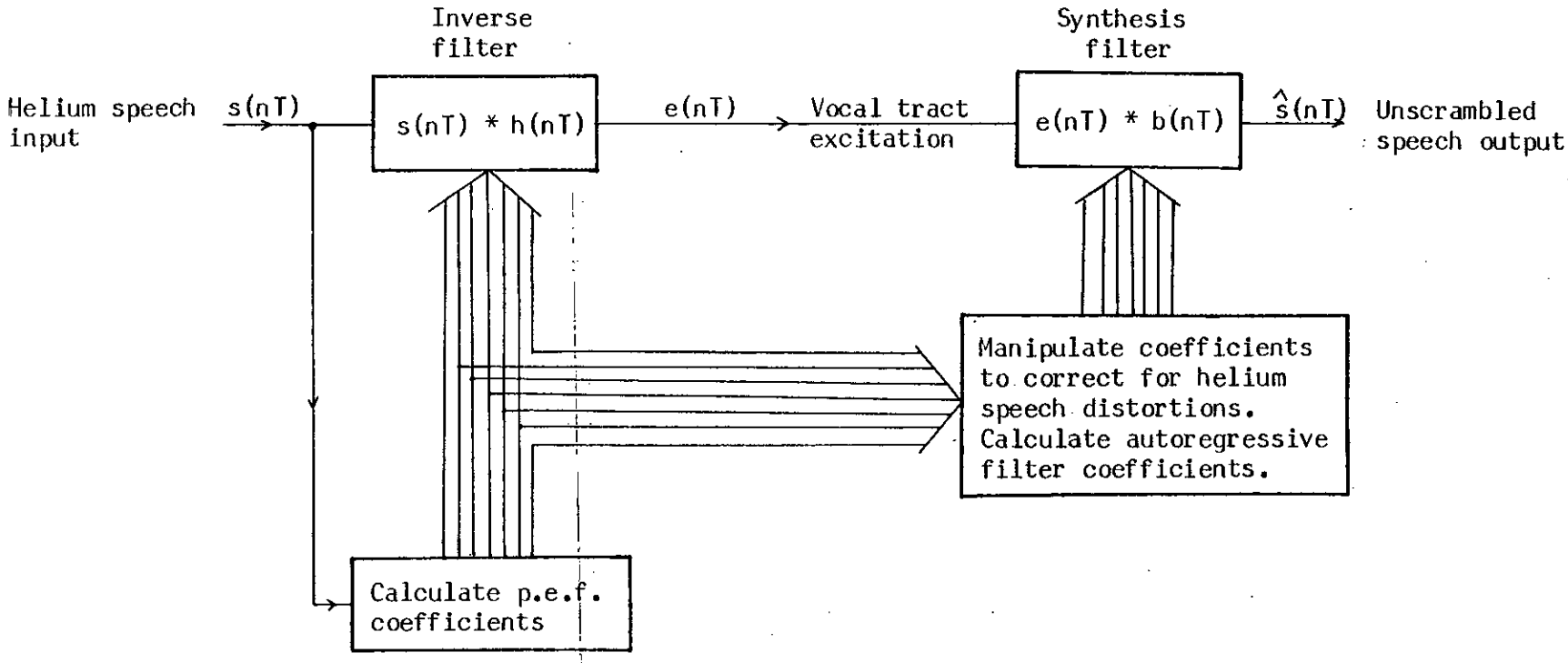


Fig. 5.1 Block diagram of the basic principles of the RELPC unscrambling technique.

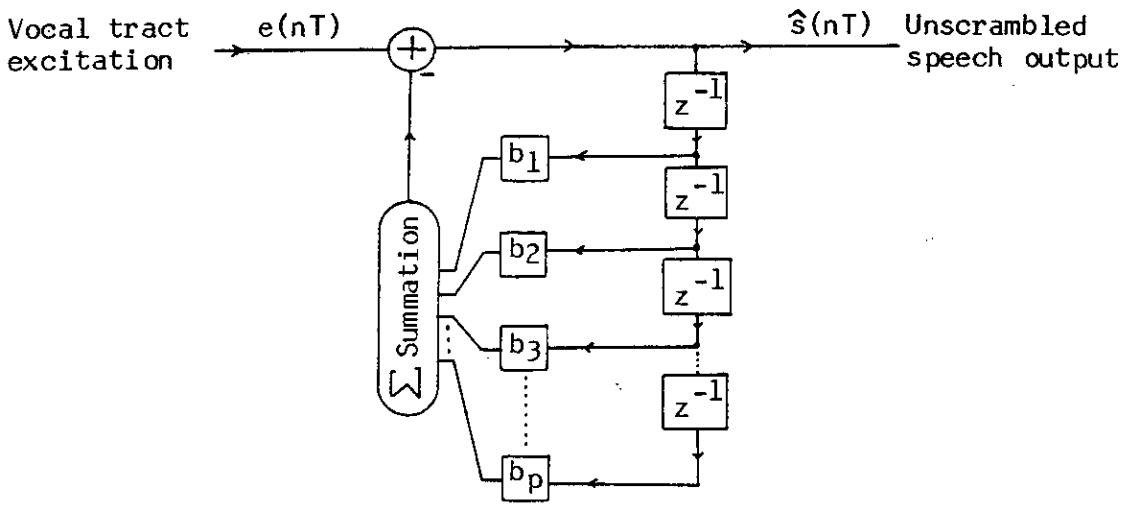


Fig. 5.2 Autoregressive filter model of the human vocal tract.

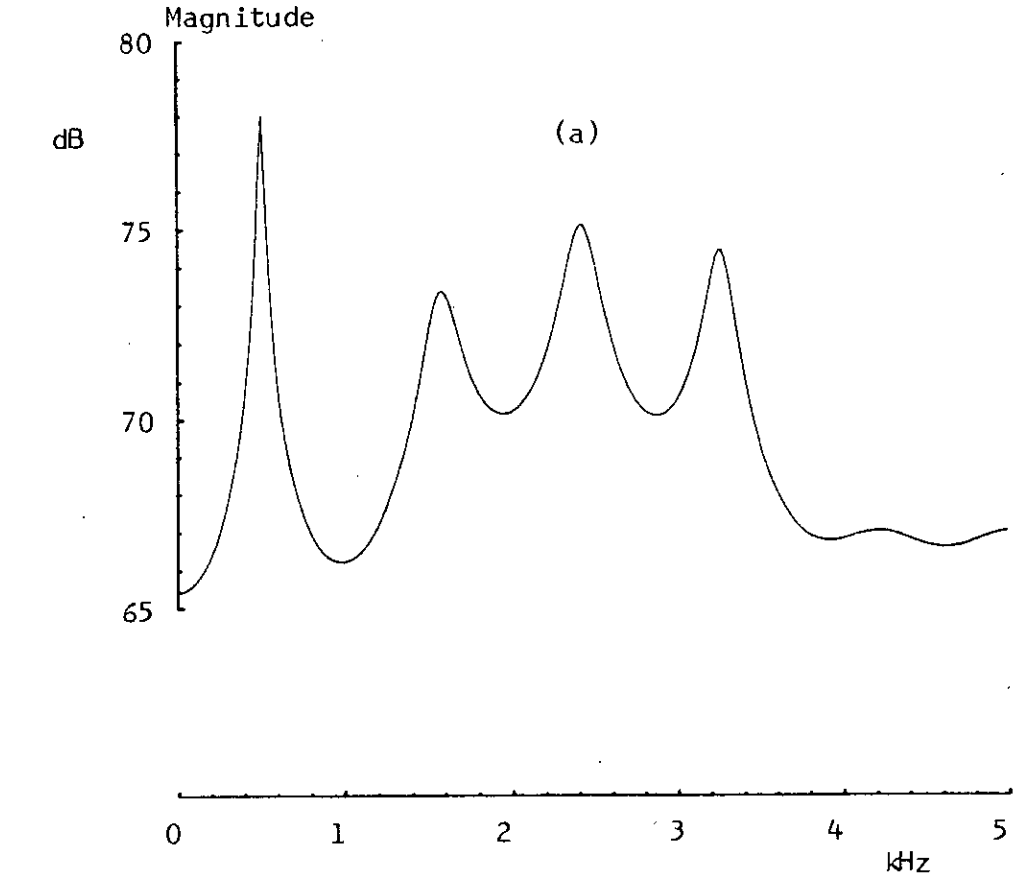
5.1.2 ESTIMATION OF THE SYNTHESIS FILTER STRUCTURE

The synthesis filter which is excited by the residual signal to produce the unscrambled speech is identical in structure to the AR filter of Fig. 3.8, and is shown here again in Fig. 5.2. The aim is to generate the AR filter coefficient series b_k, $0 < k \leq p$, ostensibly from a manipulation of the already-known p.e.f. coefficient series a_k. Two methods of achieving this have already been reported [75,200], both of which in effect solve for the locations of the zeros of the inverse filter (poles of the vocal tract transfer function). The first takes the inverse filter impulse response given by the series $a_0 + a_1z^{-1} + a_2z^{-2} + \ldots a_kz^{-k} + a_pz^{-p}$ and expands it in time, interpolating between existing coefficient values of a_k to estimate a new synthesis filter structure which has more stages compared to the corresponding p.e.f. due to the interpolating procedure [75]. This technique, however, is limited to linear scaling only of the vocal tract frequency response. Another method [200], currently under investigation, solves explicitly for the locations of the p.e.f. transfer function zeros by substitution

of matched-z transform [201] approximations in the p.e.f. impulse response equation, with subsequent manipulation of the (polar) z-transform representation of the system. This method permits generation of a synthesis filter whose coefficients have been scaled to achieve nonlinear correction of the helium speech effect. It has been intimated, however, that this technique incurs problems in the synthesis of the AR filter due to the use of the matched-z transform algorithm.

For the novel proposal in this thesis, consider the properties of the AR system related to the Fourier spectrum. In section 3.3, it was demonstrated that the vocal tract frequency response can be obtained from the inverse (smooth) spectrum of the p.e.f. impulse response. It should therefore be possible to re-estimate filter values by, say, applying the helium speech correction to the smoothed spectrum, and then applying the inverse Fourier transform to provide a new coefficient series corresponding to the corrected vocal tract response. However, the problem with this solution relates, as with the STFT system solution in section 4.2, to insufficient knowledge of how to manipulate spectral phase in a way commensurate with operations on the magnitude spectrum. Magnitude and phase for the smoothed spectrum (14-pole fit), corresponding to the vowel "*er*" in "h*er*d" spoken in air, sampled at 10kHz (see Fig. 3.11), are shown in Fig. 5.3, demonstrating that magnitude and phase spectra occur in unique pairs for a signal generated by a LTI system.

An alternative solution which would avoid having to manipulate spectral phase would be to use the smooth power spectrum derived from the p.e.f. impulse response, defined by multiplying each complex Fourier spectrum component by its complex conjugate. From equs.(3-18) and (5-2), then from the definition of the DFT, for the n^{th} frequency component:



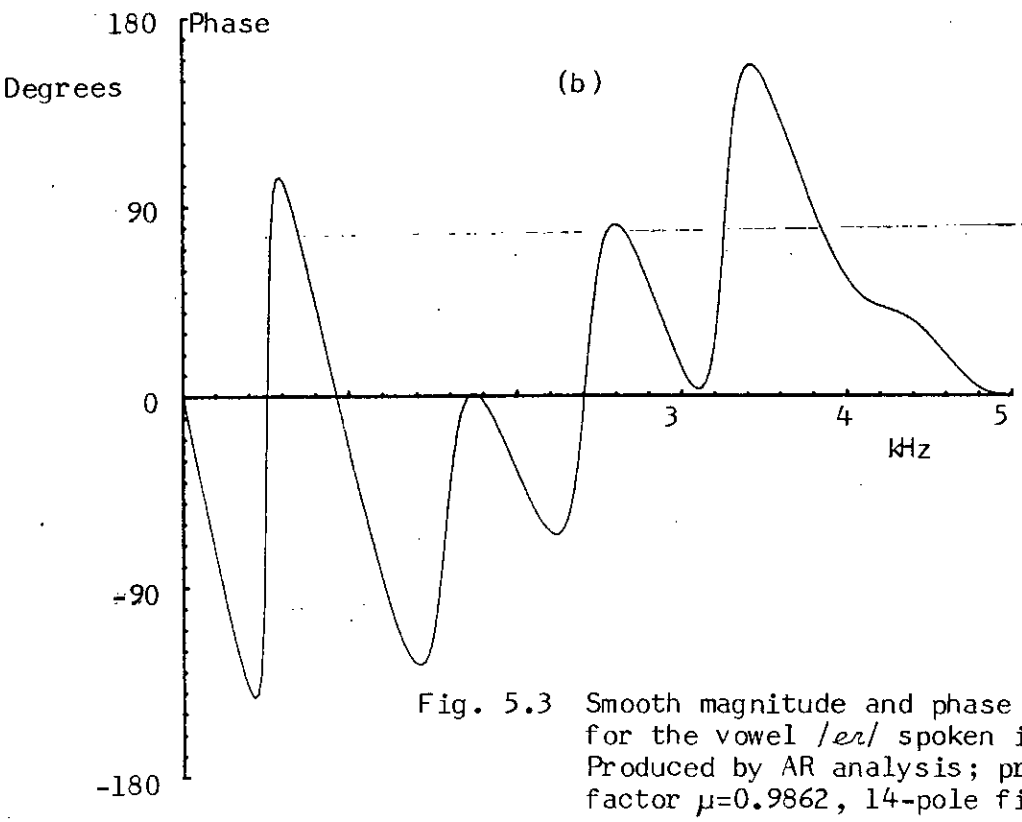


Fig. 5.3 Smooth magnitude and phase spectra for the vowel /er/ spoken in air. Produced by AR analysis; preemphasis factor μ=0.9862, 14-pole fit.

$$P(n) = \frac{K}{\sum_{k=0}^{p} W^{nk}.a^k \times \sum_{k=0}^{p} W^{-nk}.a^k} \qquad (5\text{-}5)$$

where P(n) is the smooth power spectrum amplitude for frequency component n, and W represents complex frequency, $W = e^{-2j\pi/N}$, where N is the length of the DFT.

Say that P(n) is corrected for the helium speech effect to produce a new power spectrum estimate, $\hat{P}(n)$. Let the corresponding inverse filter be given by the series b_k, $0 \leqslant k \leqslant p$, $b_o = 1$. The problem is now to estimate the values of the b_k from the N known values of $\hat{P}(n)$, $0 \leqslant n < N$, where $N < N$ and will be dependent on the amount of corrective spectral compression applied to the smoothed power spectrum. The general form of equ.(5-5) is still valid, and therefore, for any value of n,

$$\sum_{k=0}^{p} W^{nk}.b_k \times \sum_{k=0}^{p} W^{-nk}.b_k = K/\hat{P}(n) = K_n, \text{ say} \qquad (5\text{-}6)$$

Thus, there are a total on N equations of the form of equ.(5-6), each involving the p+1 -length coefficient series b_k, $0 \leqslant k \leqslant p$. Consider a simple filter coefficient series with p=2. Then

$$(b_o.W^o + b_1.W^n + b_2.W^{2n}) \times (b_o.W^o + b_1.W^{-n} + b_2.W^{-2n}) = K_n \qquad (5\text{-}7a)$$

$$b_o^2 + b_o b_1.W^{-n} + b_o b_2.W^{-2n} + b_o b_1.W^n + b_1^2 + b_1 b_2.W^{-n} + b_o b_2.W^{2n} + b_1 b_2.W^n + b_2^2 = K_n \qquad (5\text{-}7b)$$

For large values of p, then each of the N equations of the form of (5-7b) will be very unwieldy, and although arguably amenable to solution in real time, this particular approach has not been pursued further in this thesis.

5.1.3 A NOVEL METHOD FOR COMPUTING ANALYSIS AND SYNTHESIS FILTER STRUCTURES

The novel method presented in this thesis which allows calculation of both p.e.f. and a.r.f. structures with relative ease depends upon the relationship between the autocorrelation function and the power spectrum, as defined by the Weiner-Kinchine theorem (138):

$$R(\tau) = \int_{-\infty}^{\infty} P(f)\, e^{j2\pi f \tau} df \qquad (5\text{-}8a)$$

or, in discrete form, for an N-length DFT,

$$r(m) = \sum_{n=0}^{N} P(n) e^{j2\pi mn/N} \qquad (5\text{-}8b)$$

Solution for the p.e.f. coefficient series a_k, $0 \leqslant k \leqslant p$, by Levinson recursion, demands the use of the first p+1 autocorrelation lags, r(0) ... r(p), and it is proposed here that these be calculated through the power spectrum by application of equ.(5-8b) directly to short-time segments of the helium speech signal. That is, the input signal sequence is first transformed to the frequency domain by the DFT, whereupon each complex frequency component is multiplied by its own complex conjugate to produce a cophase power spectrum. The inverse DFT is then used to provide the first p+1 autocorrelation coefficients by the direct application of equ.(5-8b). Note at this stage that special precautions are required to force a linear autocorrelation from the cyclic autocorrelation afforded by the DFT, and this topic is discussed further in section 5.2.2.

From the autocorrelation series, solutions are then obtained for the coefficient series a_k of the p.e.f., and the residual signal is derived by passing the input helium speech through the inverse filter formed by the p.e.f.

Notice at this point that a frequency spectrum for the input helium

speech has already been derived. The novel aspect of the RELPC system is that operations to correct for the helium speech distortion are applied directly to the original power spectrum calculated from the short-time DFT of the helium speech signal.

After correction for the helium speech effect, the resulting power spectrum is then inverse transformed to provide a new autocorrelation sequence r(m) corresponding to the unscrambled speech, from which the coefficient series b_k corresponding to the AR synthesis filter is calculated, again by Levinson recursion. The residual signal is subsequently used to excite the AR filter to produce the unscrambled speech for which formants have been corrected, but type (voiced/unvoiced) and fundamental frequency of excitation have been conserved.

In comparison to the methods for obtaining the b_k outlined in section 5.1.2, note that here, prior to the application of spectral correction for the helium speech effect, the power spectrum is not smooth, in the sense that it still contains information regarding the vocal tract excitation function, both in terms of its overall spectral contour and the frequency spacing between spectral peaks, which is related to both the type and fundamental frequency of excitation of the vocal tract, see Fig. 2.4. Therefore, applying e.g. spectral compression to the power spectrum of voiced speech necessarily entails an effective decrease in the equivalent fundamental frequency of the resulting speech. However, this apparent imperfection is redressed here by a consideration of an implicit but rarely explicitly-expressed property of autoregressive processing, namely that although any single prediction error filter is constructed from a signal having, say, a well-defined fundamental frequency, the resulting inverse filter will produce a spectrally white residual when applied to any similar signal originating

from the same, fixed LTI system, but which possesses a different or even quasiperiodic fundamental frequency, assuming the LTI system in question is excited by a train of Dirac impulses or by white Gaussian noise.

The proof here is given in part by a theoretical treatment of AR processing as a maximum entropy process [132,197], and partly from a heuristic consideration of the problem.

The entropy, or uncertainty of information about some process consisting of N possible events, is defined by [202,203]:

$$\mathcal{H} = -K\sum_{i=1}^{N} p_i.\log(p_i) \qquad (5\text{-}9)$$

where $\mathcal{H}$ is the entropy of the process, K is some positive constant, and the p_i are the probabilities of occurence of each of the N possible events of the process.

Maximum entropy, that is minimum information known about the process, is defined to occur when each of the probabilities p_i is the same, $p_1 = p_2 = p_i = p_N = K$, where K is a constant, K=1/N.

In terms of the inverse filtering problem, then the entropy of the spectrum of the residual signal must be maximised, that is [204]:

$$\mathcal{H} = \frac{1}{4.fs}\int_{0}^{fs} \ln(P_r(f))df \qquad (5\text{-}10)$$

subject to the constraints

$$r(m) = \int_{0}^{f_s} P_s(f)e^{-j2\pi mf/f_s}df \qquad 0 \leqslant m \leqslant N \qquad (5\text{-}11a)$$

and $$\left|\ln(P_r(f))\right| < \infty \qquad (5\text{-}11b)$$

where $P_r(f)$ is the power spectrum of the residual, f_s is the sampling frequency and $P_s(f)$ is the power spectrum of the input signal. Maximisation of the entropy $\mathcal{H}$ in equ.(5-10) is similar to the case of equ.(5-9),

and occurs when $P_r(f)$ is maximally flat so that all frequency components of $P_r(f)$ have the same power amplitude. The condition which is of paramount importance in this application to helium speech correction is that there is no stipulation as to their distribution in the frequency domain, only that their power amplitudes must all be equal in order to maximise entropy.

Let the inverse filter frequency response be H(f). The convolution of the inverse filter transfer function and the time-domain signal s(nT) leads to multiplication in the frequency domain, and therefore since

$$P_r(f) = H(f) \times P_s(f) \qquad (5\text{-}12)$$

then, from equ.(5-10),

$$H = \frac{1}{4.f_s} \int_0^{f_s} \ln(H(f) \times P_s(f))df \qquad (5\text{-}13)$$

The solution for the inverse filter structure which forces the maximum entropy condition upon equ.(5-13) is given by the linear prediction (autoregressive) solution of section 3.3.2, providing that the class of signals under consideration originate from an LTI system characterised by an all-pole transfer function, as is assumed to be the case here for the speech waveform. Moreover, the conclusion, from the above consideration of AR processing as a maximum entropy problem, is that the inverse filter is specific, not to the signal waveform as a whole, but rather to the characteristic transfer function of the LTI system whose excitation has produced the signal. Thus, for any given LTI system, the application of the corresponding p.e.f. to the signal produced by random or periodic excitation, at any frequency, of the LTI system will produce a residual with the excitation frequency preserved but exhibiting maximum entropy, that is, having a maximally white spectrum.

This principle is perhaps difficult to grasp, particularly since calculation of the p.e.f. coefficients requires the use of the autocorrelation series, which evidently has a very different form depending on the fundamental frequency of any given signal.

As an illustrative example and heuristic proof of the above postulate, consider an LTI all-pole system characterised by a power spectrum which exhibits resonances at, nominally, 1kHz, 3kHz and 5kHz with bandwidths 80Hz, 100Hz and 150Hz and relative power amplitudes of 0dB, -6dB and -9dB respectively.

Fig. 5.4(a-c) demonstrates three different waveforms, sampled at 16kHz, generated from this system by impulse excitation over a 3:1 range of excitation rate, namely 150Hz, 100Hz and 50Hz respectively. Fig. 5.5(a-c) shows the corresponding autocorrelation sequences calculated explicitly using a 4096-point truncated signal sequence, that is,

$$r(i) = \sum_{n=0}^{4095} s_n \cdot s_{n-i} \quad , \qquad 0 \leqslant i < 4096 \qquad (5\text{-}14)$$

where $s_{n-i} = 0$ for $n-i < 0$. As can be seen from Fig. 5.5, the autocorrelation sequences are very different from each other when viewed as waveforms as presented here.

The p.e.f. was chosen to be an 11^{th} order filter, i.e. p=11, so requiring the first 12 autocorrelation lags (r(0) ... r(11)) commensurate with calculation of the p.e.f. structure via Levinson recursion. Fig. 5.6 shows a graphical and tabulated comparison of the corresponding autocorrelation sequences, normalised to the value of r(0) in each case, which is permissible since one of the properties of Levinson recursion is that all autocorrelation lags can be scaled by any arbitrary factor without affecting the final values of the p.e.f. coefficients. Note that whereas the autocorrelation waveforms of Fig. 5.5 appear grossly

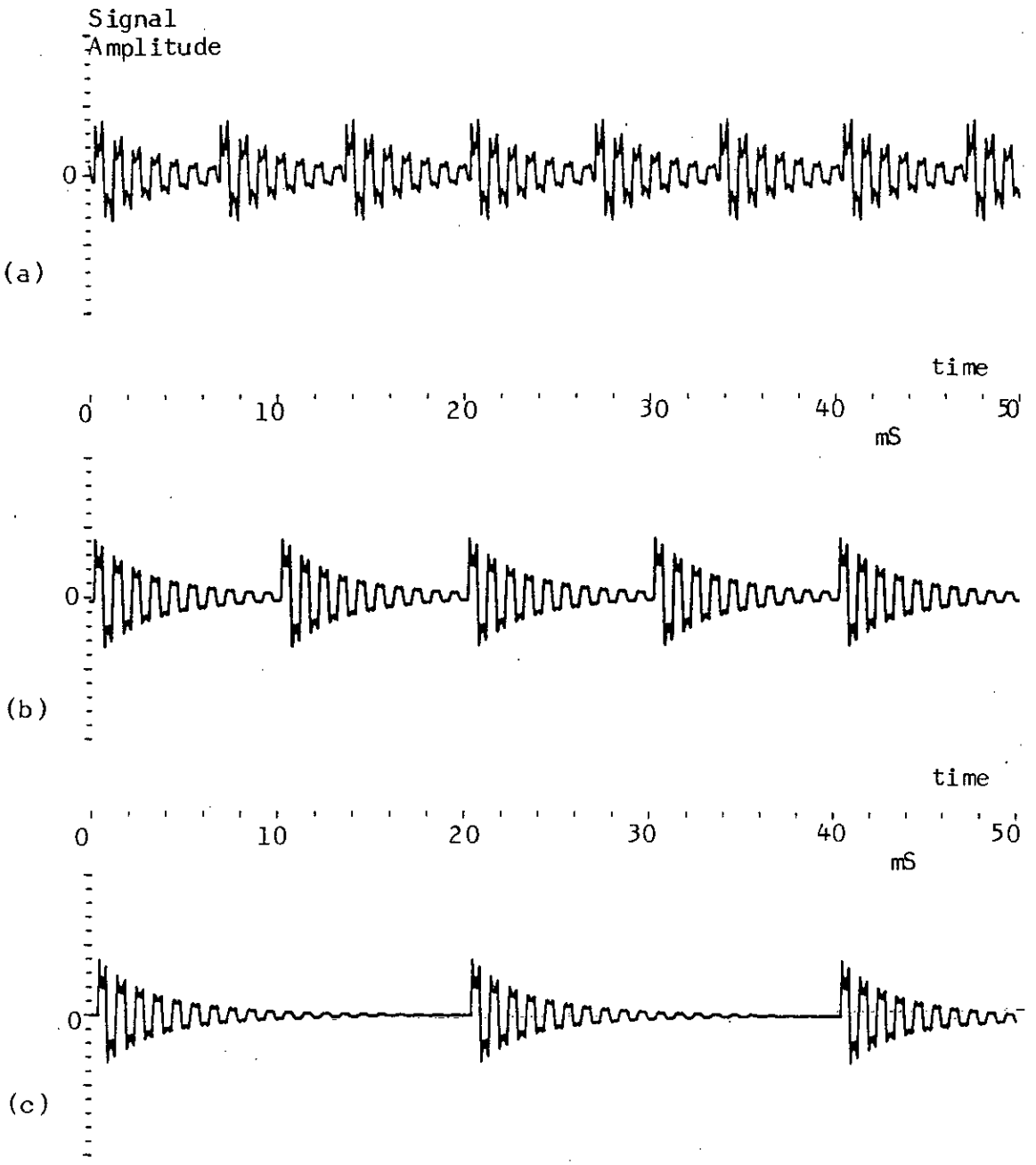


Fig. 5.4 Impulse excitation of an exact LTI filter system with frequency response having resonances nominally at 1kHz, 3kHz and 5kHz, bandwidths 80Hz, 100Hz and 150Hz, and relative peak power amplitudes of 0dB, -6dB and -9dB respectively.
Excitation at (a) F_o=150Hz (b) F_o=100Hz and (c) F_o=50Hz.

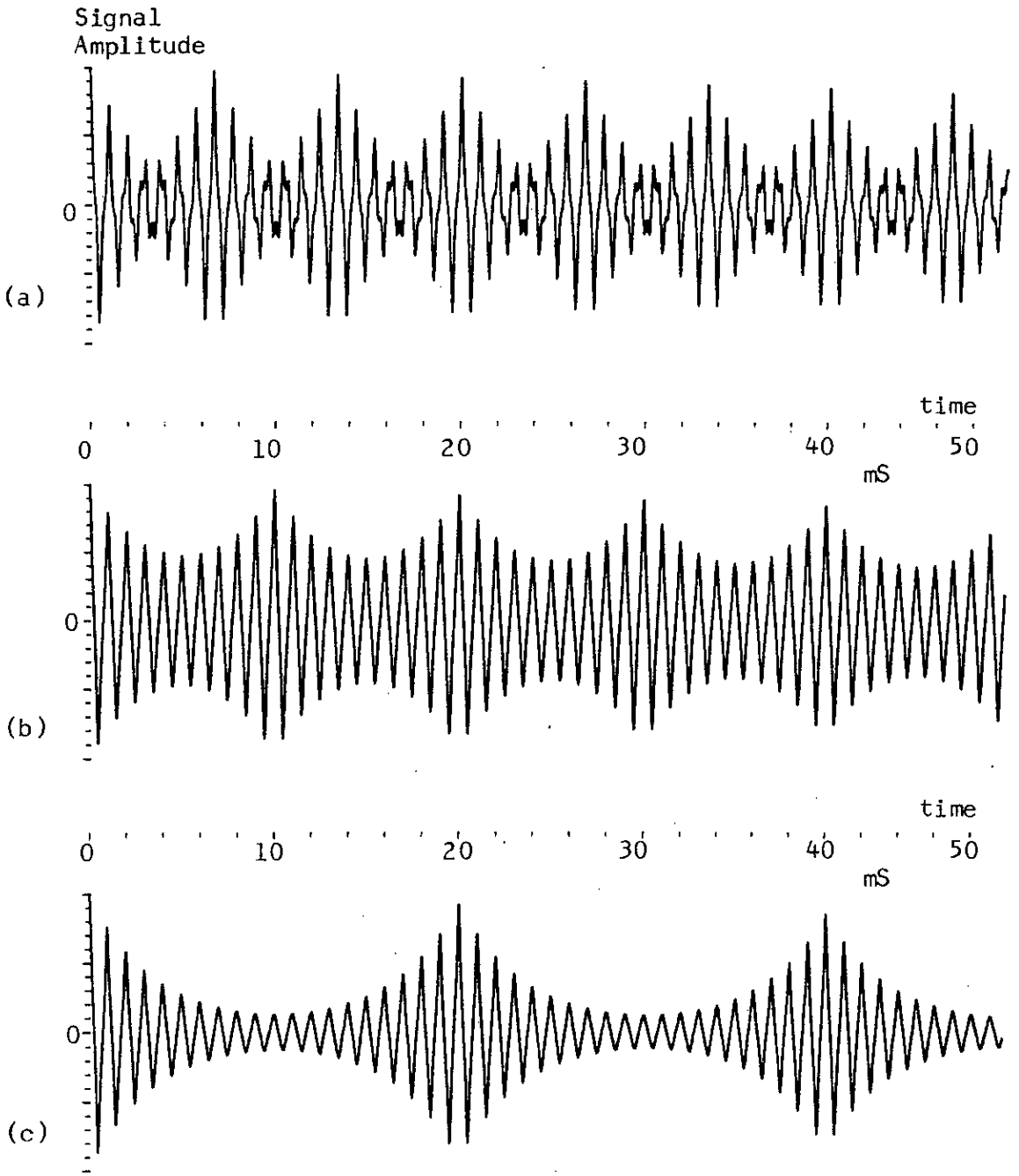
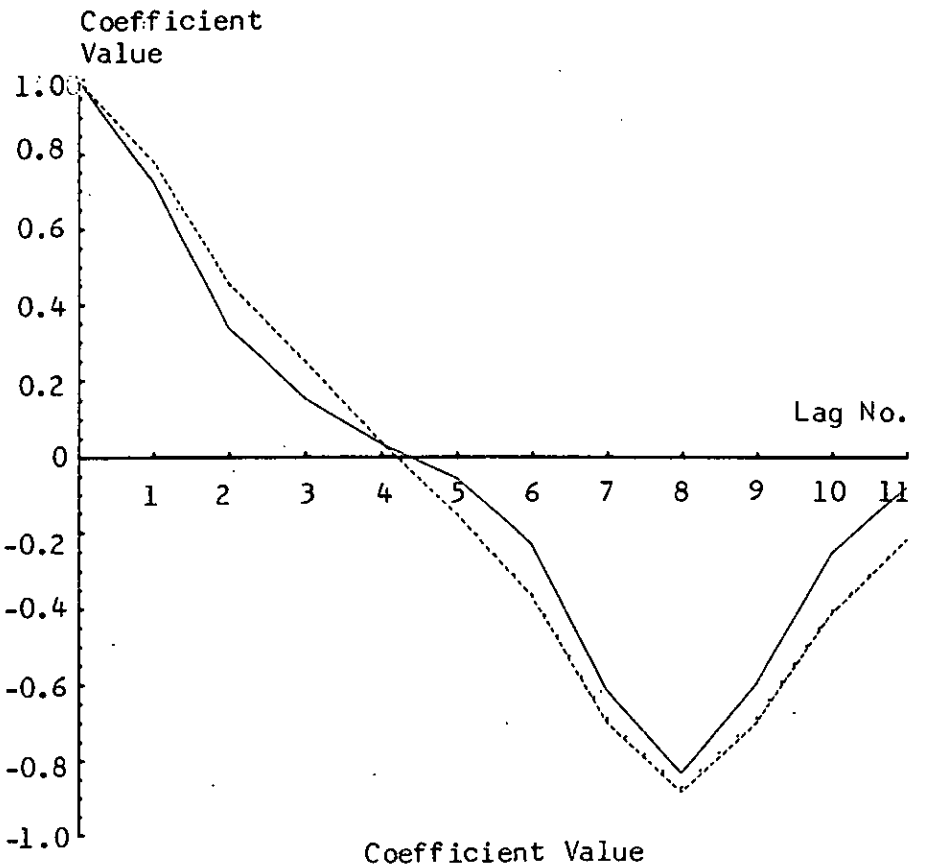


Fig. 5.5 4096-point autocorrelation waveforms for the corresponding signals shown in Fig. 5.4(a-c).



Coefficient Value

Lag No.	Fo = 150Hz (——)	Fo = 100Hz (-------)	Fo = 50Hz (· · · ·)
0	1.0	1.0	1.0
1	0.724489	0.781289	0.778316
2	0.336077	0.455021	0.451964
3	0.152874	0.246909	0.247641
4	0.035238	0.039067	0.039982
5	-0.055930	-0.154782	-0.153395
6	-0.231364	-0.367609	-0.360764
7	-0.616017	-0.700727	-0.689694
8	-0.83481	-0.885635	-0.873633
9	-0.597487	-0.699718	-0.687535
10	-0.253139	-0.412128	-0.404325
11	-0.080661	-0.219166	-0.218213

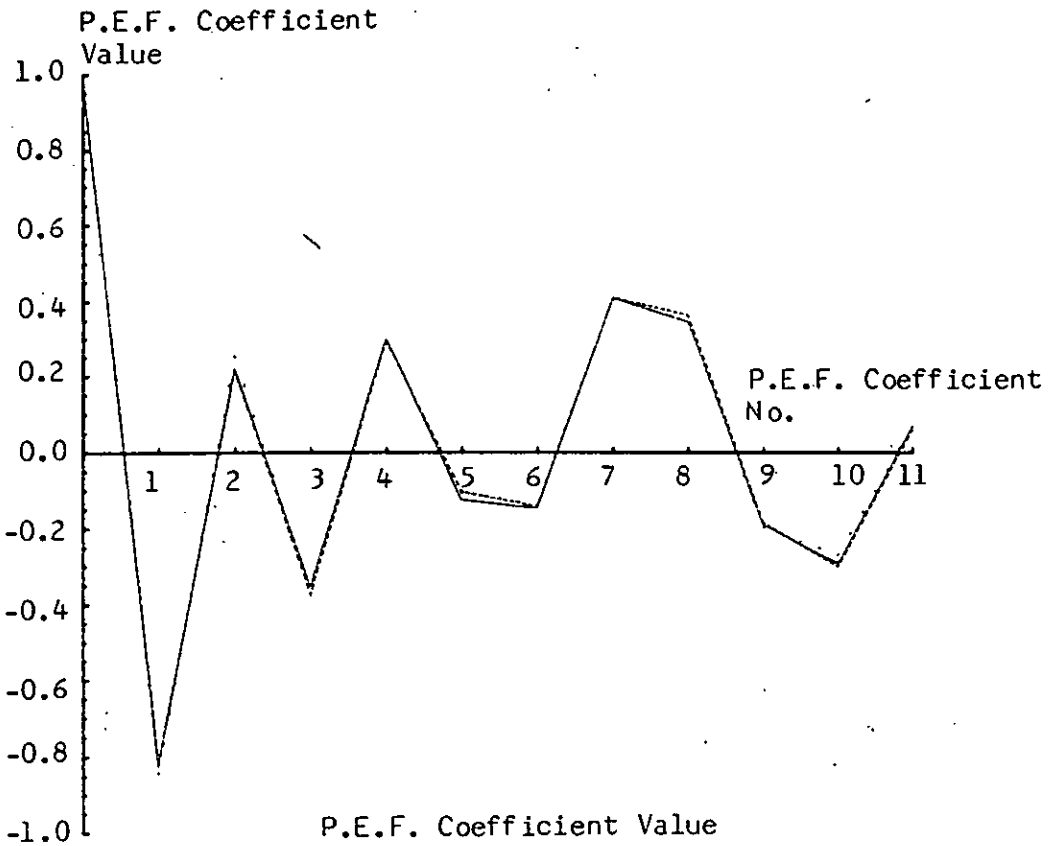
Fig. 5.6 First 12 autocorrelation lags from each waveform of Fig. 5.5(a-c) (normalised to associated values of r(0)).

dissimilar, when taken on a microscopic comparison in the region of r(0), the sequences are in fact very similar.

The resulting p.e.f. coefficient series for each of the three waveforms of Fig. 5.4(a-c) are shown in Fig. 5.7, and demonstrate that the p.e.f. structures are virtually identical for each waveform. To verify the p.e.f.'s independence of excitation frequency in producing a maximally white residual spectrum, each waveform of Fig. 5.4 was passed through the p.e.f. corresponding to the LTI system excited by the 50Hz impulse excitation (Fig. 5.7), in order to produce 3 separate residual signals. These are shown in Fig. 5.8(a-c), and it can be seen that the pulse shape of the residual is identical from waveform to waveform.

Finally, each of the 11^{th} order p.e.f.s given in Fig. 5.7(a-c) was converted to the autoregressive filter structure of Fig. 5.2 and excited by a single Dirac impulse to produce an impulse response. Since the above discussion maintains that the p.e.f. extracts all the information about the LTI system transfer function, then any corresponding a.r.f. should be the exact copy of the LTI system. This can be seen to be the case from Fig. 5.9, which compares the DFT of the impulse response of the original LTI filter system (Fig. 5.9(a)) against that of the autoregressive filter corresponding to the p.e.f. derived from the AR analysis of the 100Hz signal (Fig. 5.9(b)). Fig. 5.9(b) was found to be identical in form for each of the a.r.f. impulse responses derived from AR processing of the 50Hz and 150Hz signals.

The above discussion therefore vindicates the process of correcting the helium speech power spectrum directly, in that although the corrected spectrum has, theoretically, the wrong excitation information, this is effectively ignored in the subsequent calculation of the a.r.f. structure using the autocorrelation sequence, as derived from the inverse



Coeff. No.	Fo = 150Hz (———)	Fo = 100Hz (--------)	Fo = 50Hz (· · · ·)
0	1.0	1.0	1.0
1	-0.828695	-0.813588	-0.844892
2	0.222306	0.212968	0.255559
3	-0.354133	-0.375674	-0.374024
4	0.302959	0.291719	0.295158
5	-0.126316	-0.105072	-0.107237
6	-0.148255	-0.143753	-0.139925
7	0.40942	0.410685	0.40743
8	0.346256	0.363143	0.34725
9	-0.1929	-0.185769	-0.198677
10	-0.295119	-0.30299	-0.271773
11	0.0684669	0.0592543	0.0680021

Fig. 5.7 Prediction error filter coefficient values from applying Levinson recursion to each autocorrelation lag series of Fig. 5.6.

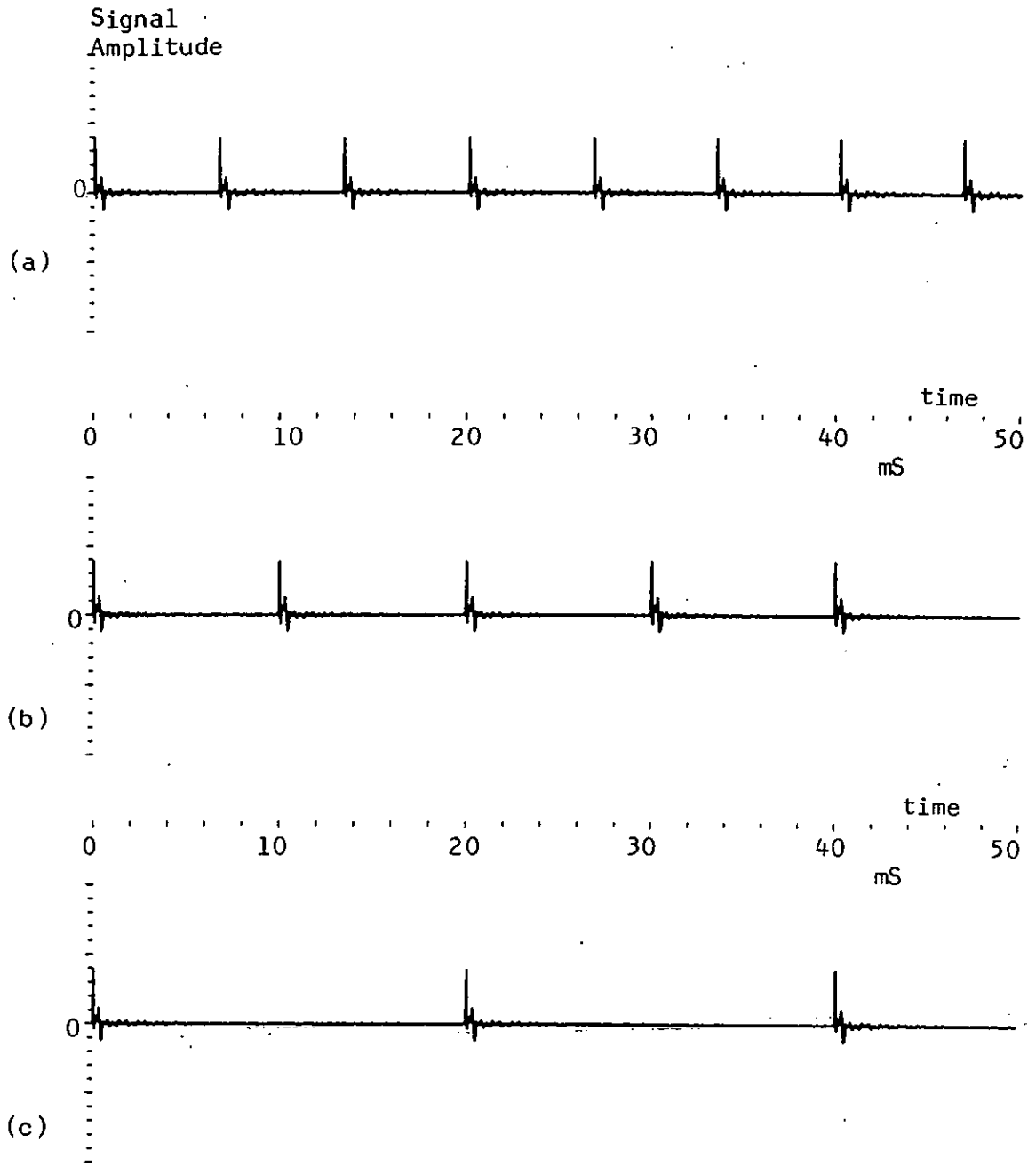


Fig. 5.8 Residual signals corresponding to each waveform of Fig. 5.4 passed through a fixed p.e.f. corresponding to the signal of Fig. 5.4(c) (F_o=50Hz).

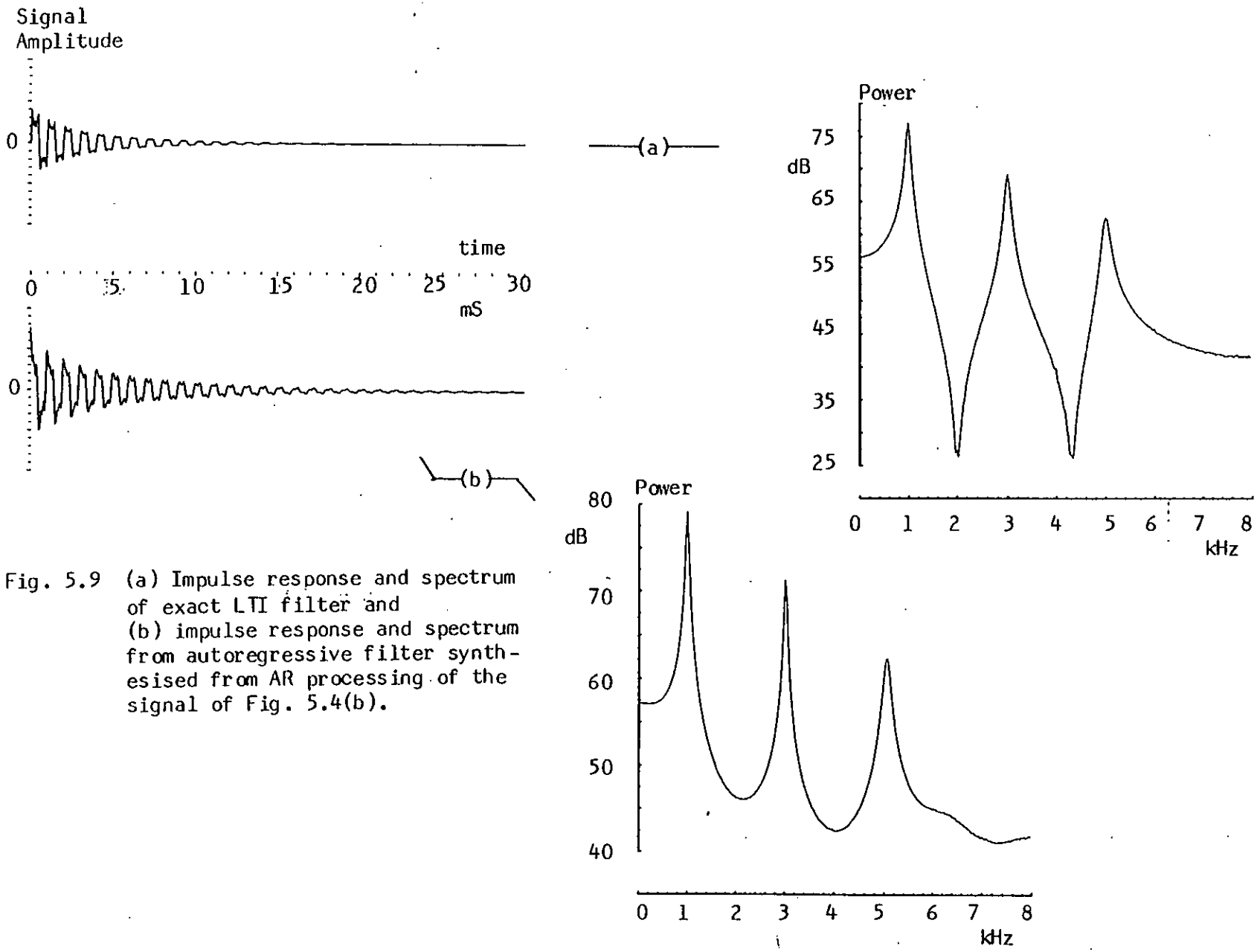


Fig. 5.9 (a) Impulse response and spectrum of exact LTI filter and (b) impulse response and spectrum from autoregressive filter synthesised from AR processing of the signal of Fig. 5.4(b).

transform of the power spectrum.

The elementary block processing specification of the RELPC system can now be identified, and is shown in Fig. 5.10.

5.2 RELPC SYSTEM DETAIL AND SIMULATION

5.2.1 THE DISCRETE FOURIER TRANSFORM OF REAL DATA

The particular implementation of the discrete Fourier transform employed in the present system simulation calculates the Fourier spectrum of a 2N-point real sequence, using an N-point complex DFT, from a consideration of the symmetry of the DFT of real-valued sequences [205]. This particular technique has in fact been employed throughout this thesis both in the AR speech analyses of Chapter III and the STFT system simulations of Chapter IV.

In its fast Fourier transform implementation, the technique offers little or no computational advantage per se over the standard 2N-point FFT, but is however useful since the resulting spectrum contains information only about N discrete frequencies from 0Hz to $f_s/2$Hz, where f_s = sampling rate. This is a desirable property since it obviates any consideration of the aliased portion of the spectrum, thereby halving the computation ultimately involved in spectral correction for the helium speech effect. The technique additionally requires only half the digital storage space of a standard 2N-point FFT operating on a 2N-point data sequence; it may therefore offer advantages in both computing speed and memory management in a microprocessor-based implementation of a real time RELPC unscrambler system. However, this version of the DFT is not in common use, and therefore a rudimentary explanation of the

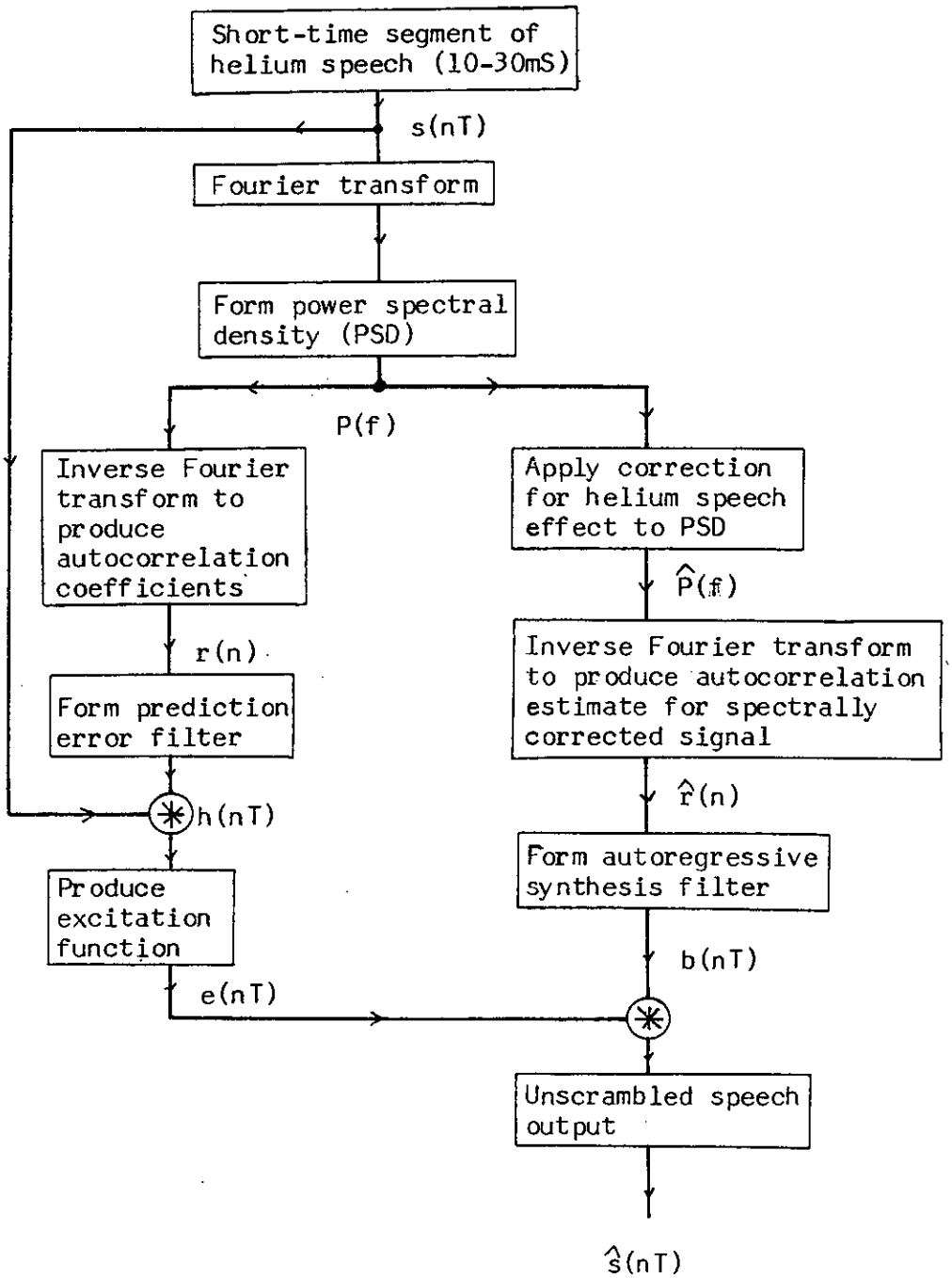


Fig. 5.10 Elementary block processing specification for the RELPC unscrambler system.

theory is given here.

Consider the N-point DFT of a N-point real sequence:

$$X_k = \sum_{n=0}^{N-1} x_n . W^{nk} \qquad 0 \leqslant k \leqslant N-1 \qquad (5\text{-}15)$$

where $W = e^{-j2\pi/N}$. Since x_n is real, then the real part of X_k, $Re\{X_k\}$ is an even function of k and the imaginary part, $Im\{X_k\}$ is an odd function of k, over the N possible values of index k, that is [206]:

$$Re\{X_k\} = Re\{X_{N-k}\} \qquad (5\text{-}16a)$$

and $$Im\{X_k\} = -Im\{X_{N-k}\} \qquad (5\text{-}16b)$$

that is $$X_k = X^*_{N-k} \qquad (5\text{-}16c)$$

where (*) denotes complex conjugation.

Consider now two independent real sequences x_n and x_n, each N-points in length. Then if:

$$Y_k = \sum_{n=0}^{N-1} (x_n + jx_n) . W^{nk} \qquad (5\text{-}17a)$$

then $$Y_k = X_k + jX_k \qquad (5\text{-}17b)$$

and, from equs.(5-16),

$$X_k = (Y_k + Y_{N-k})/2 \qquad (5\text{-}18a)$$

and $$X_k = (Y_k - Y_{N-k})/2j \qquad (5\text{-}18b)$$

where X_k and X_k are the complex spectra corresponding to the series x_n and x_n respectively. Equations (5-17)-(5-18) together form the basis for calculating the N-point complex DFT of a 2N-point real sequence. Say the input sequence s_r, $0 \leqslant r \leqslant 2N-1$, is shuffled into a complex N-point data array of the form $x_n + jy_n$ such that the even-indexed data of the original sequence fills the real array, $x_n = s_{2n}$ and the imaginary array is given by $y_n = s_{2n+1}$, $0 \leqslant n \leqslant N-1$.

The DFT of the complex data sequence $x_n + jy_n$ is then calculated from equ.(5-17a):

$$Y_k = \sum_{n=0}^{N-1} (x_n + jy_n) . W^{nk} \qquad 0 \leqslant k \leqslant N-1 \qquad (5\text{-}19)$$

To clarify the situation, let $w = W^{\frac{1}{2}} = e^{-j2n/2N}$. Rewriting equ.(5-19):

$$Y_k = \sum_{n=0}^{N-1} (s_{2n} + js_{2n+1}).w^{2nk} \qquad 0 \leqslant k \leqslant N-1 \qquad (5\text{-}20a)$$

$$= \sum_{n=0}^{N-1} s_{2n}.w^{2nk} + j \sum_{n=0}^{N-1} s_{2n+1}.w^{2nk} \qquad (5\text{-}20b)$$

$$= A_k + jB_k \qquad (5\text{-}20c)$$

where A and B are themselves complex variables. However, the required DFT, F_k, of the input series, over the range 0Hz to $f_s/2$, is in fact given by:

$$F_k = \sum_{n=0}^{N-1} s_{2n}.w^{2nk} + \sum_{n=0}^{N-1} s_{2n+1}.w^{(2n+1)k} \qquad 0 \leqslant k \leqslant N-1 \qquad (5\text{-}21)$$

Comparison of equs.(5-20b) and (5-21) shows that equ.(5-20b) requires corrective multiplication of the imaginary term B_k by the complex correction factor $-jw^k$ to achieve the form of equ.(5-21) and so produce the first N points of the DFT of the 2N-point real sequence. Recovery of F_k from Y_k is a two-stage process involving, firstly, the retrieval of the real and imaginary parts of both A_k and B_k of equ.(5-20c) from an application of equs.(5-18). The complex correction factor can then be applied to the complex components of each of the B_k, and the complex results summed together with the A_k to provide each complex value for the F_k.

In the FFT implementation of this algorithm, corrective multiplication can be achieved by the use of the same sine/cosine look-up tables as used in the conventional FFT which is first applied to produce the Y_k of equs.(5-20). Furthermore, the inverse Fourier transform can be achieved by carrying out the exact inverse procedures to those outlined above, and moreover requires overall simply a few sign changes in both the corrective multiplication process and the inverse FFT.

5.2.2 AUTOCORRELATION USING THE DISCRETE FOURIER TRANSFORM

Correction for the helium speech effect using the RELPC unscrambler system depends on the ability to accurately calculate the autocorrelation sequences, required in the construction of both the p.e.f. and the a.r.f. structures, via the Fourier transform.

Consider an N-point data sequence, s_n, and its discrete autocorrelation sequence, r(i), as calculated directly:

$$r(i) = \sum_{n=0}^{N-1} s_n \cdot s_{n-i} \qquad 0 \leqslant i \leqslant p \qquad (5\text{-}22)$$

where $s_{n-i} = 0$ for $n-i < 0$. Referring to the Weiner-Kinchine theorem and equs.(5-8), it is not, however, possible to simply take the N-point standard DFT of an N-point data sequence, multiply the resulting spectrum by its complex conjugate, and then take the inverse DFT in order to compute the autocorrelation sequence. The problem here is that the DFT assumes that the data which has been transformed is infinitely periodic in modulo N on either side of the associated time window. That is, $s_{n-i} = s_{kN+(n-i)}$, k any integer. This means that the autocorrelation sequence is cyclic when computed via the DFT as opposed to the linear sequence obtained via the direct method [207,208]. The linear autocorrelation of an N-point data sequence is of length 2N points in general, therefore it is not surprising that the N-point autocorrelation sequence calculated from the standard N-point inverse DFT exhibits an 'aliasing' effect in the time domain. This effect can be avoided by padding the original N-point data sequence with an additional N zeros to give a 2N-point data sequence which can then be transformed by a standard 2N-point DFT followed by the inverse 2N-point DFT of the resulting power spectrum, thereby yielding a 2N-point autocorrelation sequence whose first N points will correspond to the first N points of the

sequence as calculated by the direct method. In this way, a linear autocorrelation sequence is forced from a cyclic process.

Figure 5.11 compares the first 12 values of the autocorrelation coefficients given by direct calculation and by the DFT method outlined in section 5.2.1 for several windows of length N=256 points extracted at random from the waveform of Fig.5.4(a). The coefficient values in each comparison are normalised to the value of r(0) as given by the direct method, and the percentage error between the two sequences is also displayed, and can be seen to be rarely in excess of $\pm$0.005%.

The use of the FFT in estimation of the autocorrelation coefficients offers a small saving in computation, requiring some 2^{13} operations, as against the direct method, which requires some 2^{15} operations for a 256-point autocorrelation sequence. In AR processing, only the first few coefficients are in fact required, therefore the direct method would undoubtedly be much faster computationally since all autocorrelation coefficients must be calculated by the inverse FFT. However, this computational penalty is unavoidable by the nature of the RELPC technique, at least in the calculation of the autocorrelation coefficients corresponding to the corrected speech spectrum. Furthermore, the FFT route is not as robust as the direct method in terms of round-off errors [168]. Accurate estimation requires a high level of quantisation of the input data and, ideally, floating-point arithmetic in calculation of the FFT. However, processing speed is aided to some extent by the fact that there is absolutely no requirement to multiply the input data with any window functions. Additionally, padding the data sequence with zeros improves accuracy during spectral frequency relocation in correction for the helium speech effect and this topic is explained further in section 5.2.5.

Lag No.	Coefficient Value: Direct Method	DFT Method	% Difference
0	1.0	0.999999	-0.000089
1	0.725143	0.725144	0.000198
2	0.336633	0.336632	-0.00033
3	0.153584	0.153586	0.000813
4	0.0362206	0.0362194	-0.003303
5	-0.0559971	-0.0559959	-0.00215
6	-0.23409	-0.234092	0.00056
7	-0.623076	-0.623075	-0.000159
8	-0.845046	-0.845048	0.000175
9	-0.605858	-0.605857	-0.000163
10	-0.257757	-0.257758	0.000509
11	-0.083976	-0.083975	-0.001406

Lag No.	Coefficient Value: Direct Method	DFT Method	% Difference
0	1.0	1.0	0.000021
1	0.723249	0.723249	0.000022
2	0.337671	0.337671	-0.0000076
3	0.156312	0.156312	0.0001315
4	0.040641	0.040640	-0.0003081
5	-0.052785	-0.052785	-0.0002494
6	-0.234158	-0.234159	0.0000987
7	-0.614677	-0.614677	0.0000167
8	-0.823577	-0.823577	0.0000561
9	-0.588399	-0.588399	0.0000087
10	-0.252175	-0.252175	0.0000917
11	-0.083052	-0.083052	-0.0001392

Lag No.	Coefficient Value: Direct Method	DFT Method	% Difference
0	1.0	1.0	0.0
1	0.731148	0.731148	0.0000417
2	0.341731	0.341731	0.0000381
3	0.152611	0.152611	0.0000427
4	0.030296	0.030296	0.0000107
5	-0.073046	-0.073046	0.0000535
6	-0.25667	-0.25667	0.0000406
7	-0.636258	-0.636258	0.0000327
8	-0.853417	-0.853417	0.0000305
9	-0.619356	-0.619356	0.0000336
10	-0.271114	-0.271114	0.0000384
11	-0.092431	-0.092431	0.0000282

Fig. 5.11 Comparison of autocorrelation by the direct calculation and DFT methods for several windows of length N=256 points, selected at random from the waveform of Fig. 5.4(a). (Values in each list are normalised to r(0) as given by the direct method).

5.2.3 SPECTRAL MANIPULATION AND AUTOREGRESSIVE FILTER STABILITY

The helium speech effect has been shown, in Chapter III, to produce an upwards translation in frequency of the vocal tract frequency response and therefore correction for this distortion by manipulation of the power spectrum necessitates spectral compression. Assuming that FFT techniques are to be used to facilitate real time implementation of the RELPC system, then unless the overall reduction in spectral bandwidth is an exact power of 2, there will be an area in the high-frequency portion of the corrected spectrum about which no information is known. That is, assuming the real data FFT implementation of section 5.2.1, then if the overall spectral compression corresponds to a ratio R, there is now no information known about frequencies in the region $f_s/(2.R)$ to $f_s/2$, since the spectral area to which they correspond in the original helium speech spectrum, from $f_s/2$ to $R.f_s/2$, will have been excised by anti-aliasing requirements. The problem, unique to this application of AR signal processing to helium speech unscrambling, then arises regarding what values, if any, to assign to these 'unknown' frequencies and the effect of the choice of these values on the autoregressive synthesis filter which will ultimately be constructed.

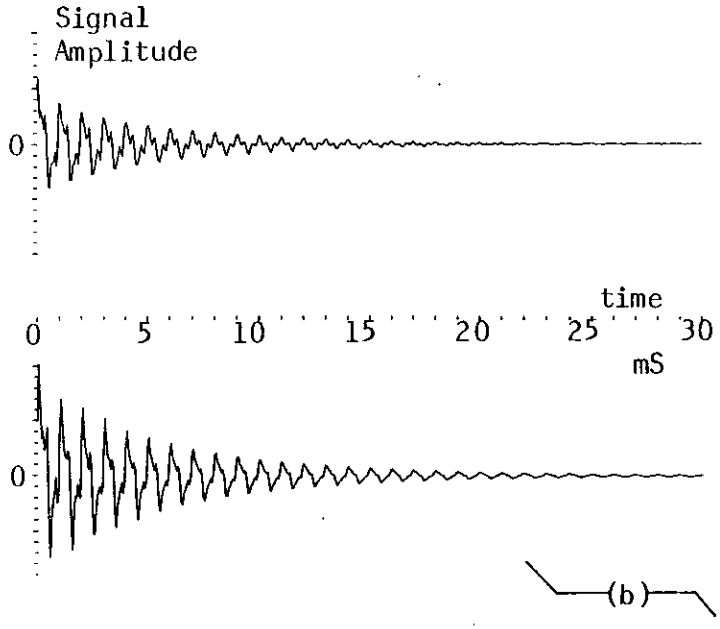
From a theoretical consideration of the RELPC system as a maximum entropy process, as detailed in section 5.1.3, it can be shown to be insufficient to simply set the unknown frequency region to zero amplitude. An exact inverse filter, exhibiting an ideal lowpass characteristic, applied to the signal described by the corrected power spectrum would theoretically produce a residual spectrum with the same region of unknown frequencies which, in the ideal lowpass case, would have zero power amplitude. In this case, the Paley-Weiner criterion [170] of equ.(5-11b) is violated, since $|\ln P_r(f)| = \infty$ for $f_s/(2.R) < f < f_s/2$.

Restated from a probability-entropy standpoint, the frequencies in this range, when set to zero amplitude, have zero probability of occurence in their corresponding time domain signal. This means that the entropy of the process in that region, instead of being maximised, is in fact minimised, see equ.(5-9), which is in direct conflict with the assumptions necessary in AR signal processing.

To investigate heuristically the effects of an ideal lowpass filter characteristic of varying bandwidth on the autoregressive filter, the signal described in section 5.1.3 and Fig. 5.4(a) was employed. The frequency response of the LTI filter system producing this signal exhibits nominally 3 formant peaks at 1kHz, 3kHz and 5kHz with bandwidths 80Hz, 100Hz and 150Hz and relative peak powers of 0dB, -6dB and -9dB respectively. The fundamental frequency of pulse excitation of the system is 150Hz, and the signal is sampled at 16kHz.

The signal was processed in contiguous sections 32mS in length. An unwindowed FFT was applied to each signal section of 512 points extended with 512 zeros in order to derive the power spectrum, using the technique of section 5.2.1. From this single power spectrum, four new power spectra were produced into which zero-amplitude components were forced in the range $f_c < f < f_s/2$ with f_c = 7kHz, 6kHz, 5kHz and 4kHz. The autocorrelation coefficients resulting from the inverse FFT of each spectrum were then derived, and the first 12 lags of each sequence, r(0) ... r(11) used to produce the corresponding p.e.f. coefficients, via Levinson recursion, from which an a.r.f. can be constructed, see Fig. 5.2. Each resulting a.r.f. was then excited by a single impulse, and the Fourier spectrum of the impulse response was then obtained.

Figures 5.12(a-c) show the corresponding impulse response-spectrum pairs for the several values of f_c stated above. It can be seen that



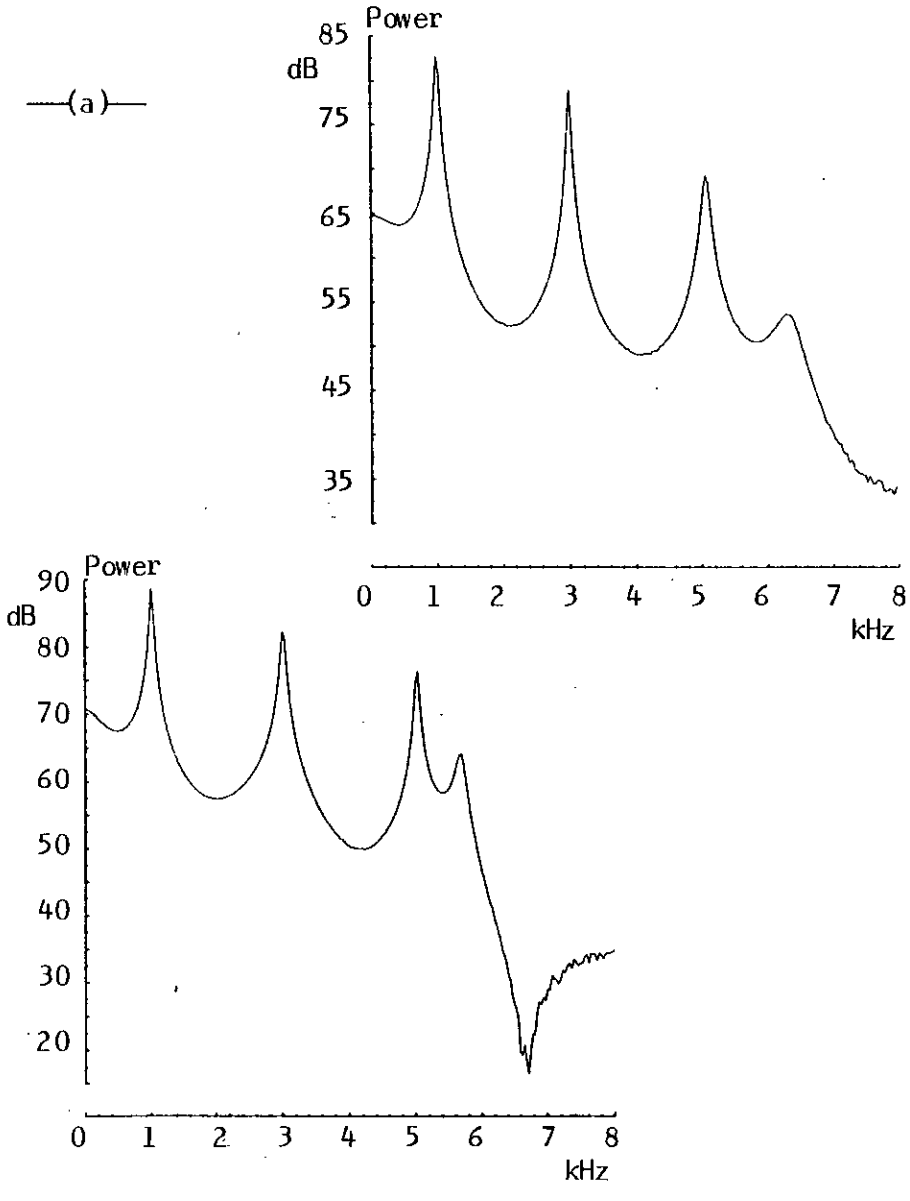


Fig. 5.12 Impulse response-spectrum pairs for autoregressive filters derived from the signal of Fig. 5.4(a) via the DFT autocorrelation method but with zero-amplitude frequency components forced from (a) f_c=7kHz and (b) f_c=6kHz.

(See over for Fig. 5.12(c))

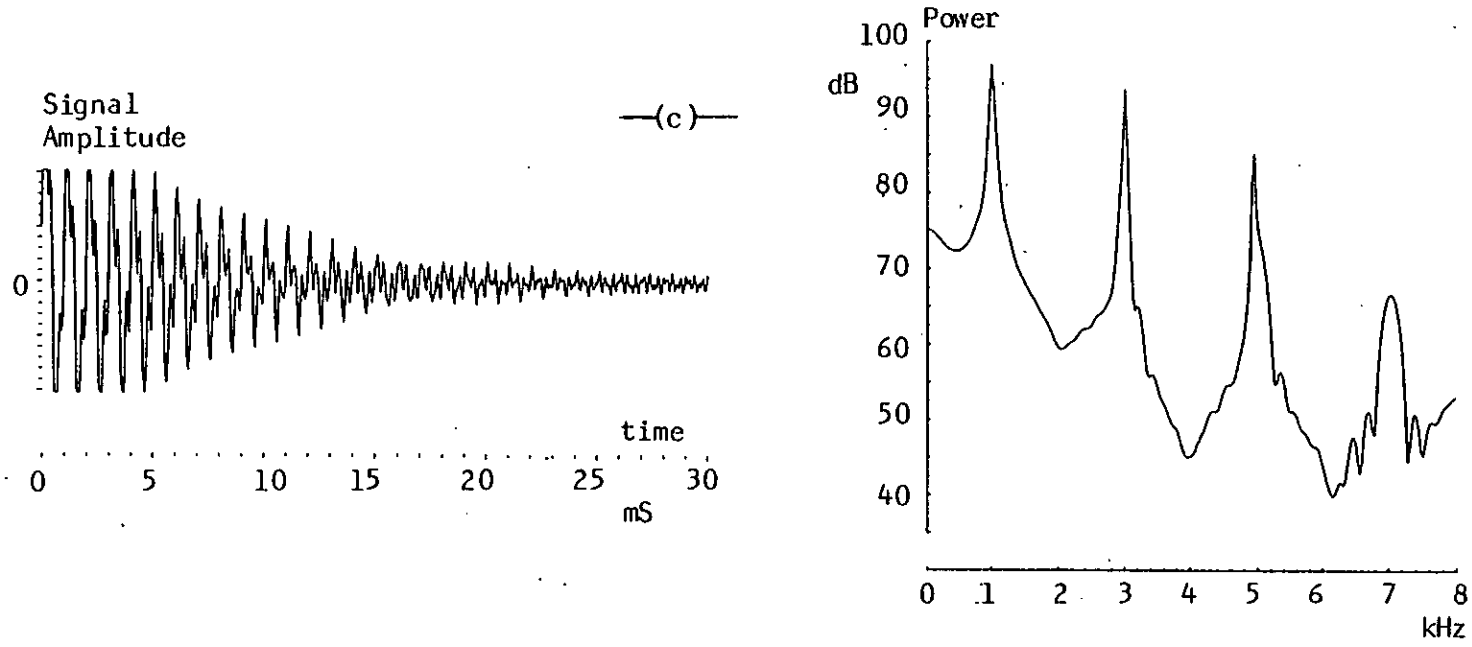


Fig. 5.12(c) Impulse response-spectrum pair for the autoregressive filter derived from the signal of Fig. 5.4(a) via the DFT autocorrelation method but with zero-amplitude frequency components forced from f_c=5kHz (note saturation clipping of impulse response).

as the bandwidth, f_c, of the ideal lowpass characteristic narrows in width, the resulting a.r.f. becomes gradually more unstable. The corresponding frequency responses also show severe distortion, although this may be due in part to the effects of peak clipping on the impulse response. For f_c = 4kHz, the impulse response diverged rapidly, halting the simulation due to floating-point overflow.

A solution to this problem of filter instability can be achieved within the power spectrum itself. Recalling that the problem relates to the, at present, minimum entropy in the unknown region of the spectrum, the entropy can nonetheless be maximised, consistent with the requirements of AR processing, by assigning a finite, constant power amplitude to each of the unknown frequency components; that is, the spectrum is forced to be maximally white over that region. Impulse response-spectrum pairs are shown using this technique in Figs. 5.13(a-d). The constant amplitude chosen for each unknown component is the last non-zero amplitude known at the cut-off frequency, f_c. Note that here, the impulse responses exhibit good stability, with little spectral distortion. Note in particular that, in this case, it was possible to achieve a stable impulse response where the original spectrum is obliterated from 4kHz, see Fig. 5.13(d).

Maximising the entropy with respect to half the spectrum in fact corresponds to the "worst case" condition of unknown frequencies. Recalling from section 5.1.3 that the fundamental frequency of excitation has no part in the characterisation of the signal in terms of the p.e.f. coefficients, then for an overall spectral compression of $R \geqslant 2.0$, the FFT frequency axis can simply be redefined as being from 0Hz - $f_s/4$ instead of 0 - $f_s/2$. In this way, the requirement to introduce unknown frequency amplitude estimates can be rendered minimal. Additionally,

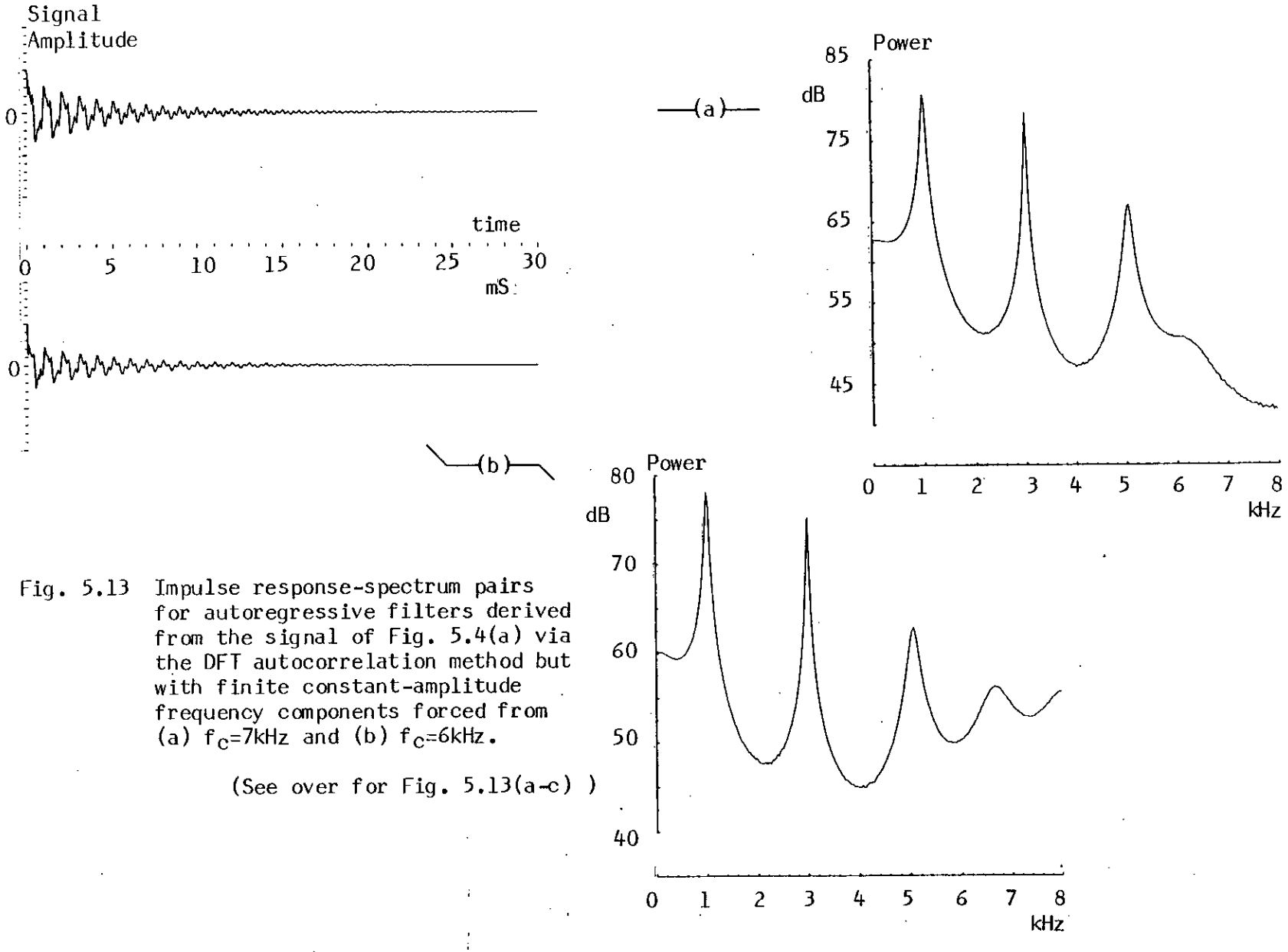


Fig. 5.13 Impulse response-spectrum pairs for autoregressive filters derived from the signal of Fig. 5.4(a) via the DFT autocorrelation method but with finite constant-amplitude frequency components forced from (a) f_c=7kHz and (b) f_c=6kHz.

(See over for Fig. 5.13(a-c))

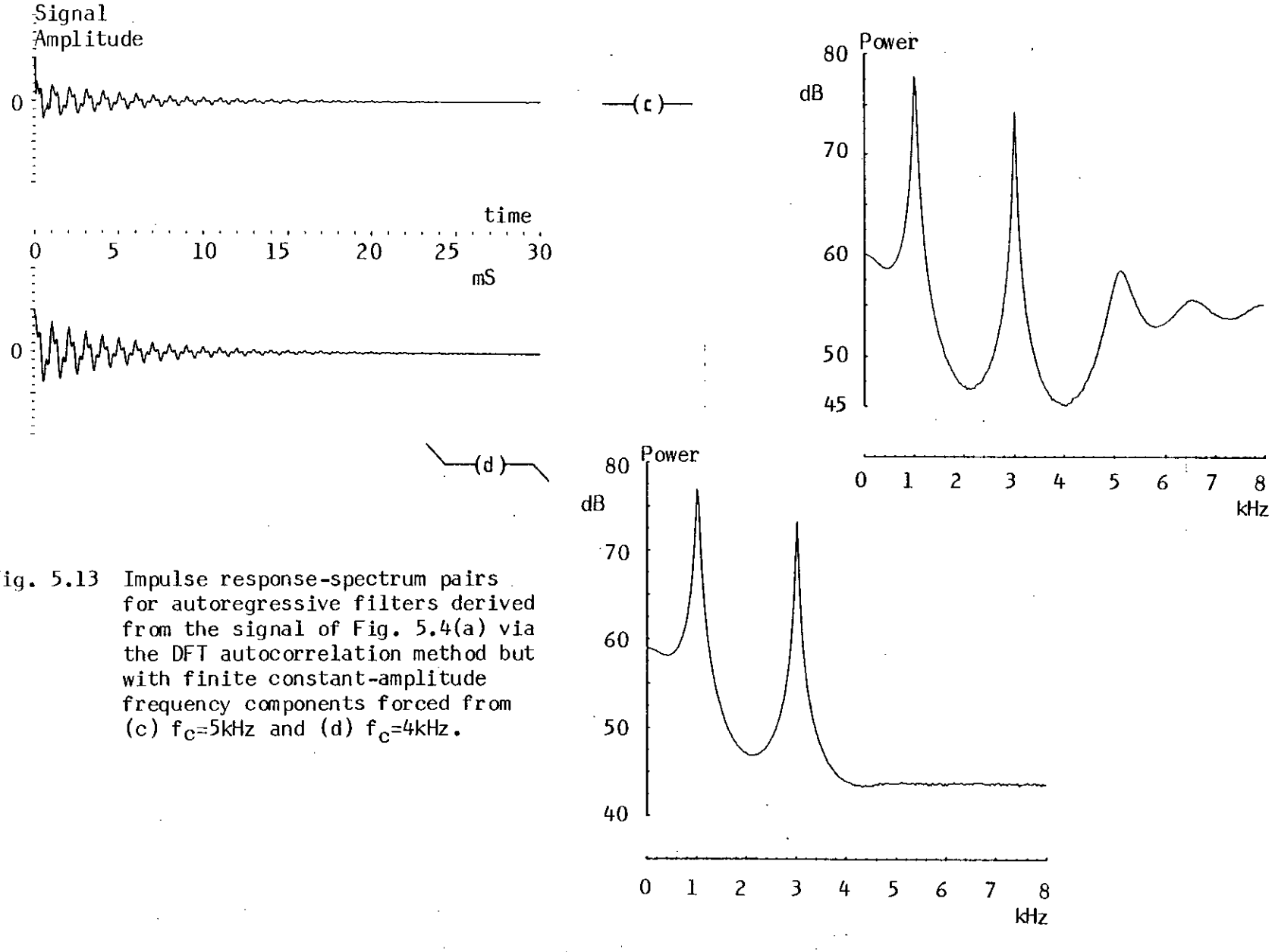


Fig. 5.13 Impulse response-spectrum pairs for autoregressive filters derived from the signal of Fig. 5.4(a) via the DFT autocorrelation method but with finite constant-amplitude frequency components forced from (c) f_c=5kHz and (d) f_c=4kHz.

the redefined frequency axis now has double the resolution, and the spectrum as a whole can be considered as that of an N/2-length data sequence, sampled at $f_s/2$Hz, to which have been appended 3N/2 extra zeros prior to performing the FFT. The inverse FFT of the power spectrum is still an approximation to linear autocorrelation, but the autocorrelation sequence must now be considered to be that of data sampled at $f_{so} = f_s/2$. This means that the residual signal must also be downsampled by a factor of 2 prior to excitation of the resulting autoregressive filter. Simple redefinition of the frequency axis in this way is particularly useful, not only because of the inherent downsampling and therefore bandwidth reduction of the resulting speech signal, but also because of the effective increase in spectral resolution, which improves the accuracy of spectral frequency relocation in correction for the helium speech effect. This latter topic is discussed further in section 5.2.5.

The pulse shape of the residual excitation signal has been shown to play an important rôle in the subjective interpretation of the acoustic waveform resulting from residual excitation of filters based on linear predictive coding [209,210]. In the present case, it can be shown that care must be taken in the downsampling operation to avoid introducing unsolicited acoustic aberrations.

The general method of downsampling a signal to some frequency f_{so}Hz from a signal sampled at frequency f_sHz consists of either digital prefiltering followed by data decimation or D/A conversion followed by analog filtering and resampling at f_{so}Hz, where the cut-off frequency of the prefilter is in both cases $f_{so}/2$Hz. Specifying that the frequency response of some LTI filter system must be compressed, using the RELPC processing method, by a factor of 2, say, then from the above discussion,

the only signal manipulation which requires to be carried out is to downsample the residual by a factor of 2. This process is applied to the signal of Fig. 5.12(a), which exhibits spectral resonances at 1kHz, 3kHz and 5kHz, and the residual signal resulting from inverse filtering of the input is shown in Fig. 5.14(a), where the input data is processed in contiguous segments each 32mS in length, sampled at 16kHz. Figure 5.14(b) shows the new residual signal (f_{so} = 8kHz) achieved by firstly applying a 1st-order digital lowpass filter with cut-off frequency 4kHz to the original residual sampled at 16kHz, followed by data decimation by a factor of 2.

Figure 5.15(a) demonstrates the signal generated by impulse excitation of the exact LTI filter with resonances nominally relocated at 0.5kHz, 1.5kHz and 2.5kHz, and whose bandwidths are halved compared to those corresponding to the signal in Fig. 5.12(a). Note however that the relative peak powers of each relocated resonance have been conserved. Figure 5.15(b) shows the signal resulting from excitation of the a.r.f., derived from the redefined power spectrum of the signal of Fig. 5.12(a), by the prefiltered residual signal of Fig. 5.14(b). It can be seen that the output resembles a signal with a fundamental frequency of approximately 38Hz, and indeed sounds subjectively very coarse.

Figure 5.14(c) shows the residual signal of Fig. 5.14(a) simply decimated by a factor of 2 with no prefiltering, and the corresponding output of the a.r.f. can be seen in Fig. 5.15(c), which can be seen to more closely resemble the ideal case of Fig. 5.15(a). This waveform compares particularly favourably when taken in comparison to the signal of Fig. 4.19(b) produced by the short-time Fourier transform processor.

Theoretically, there is no need to prefilter the decimated residual, since its spectrum is, ideally, flat in any case. Therefore here,

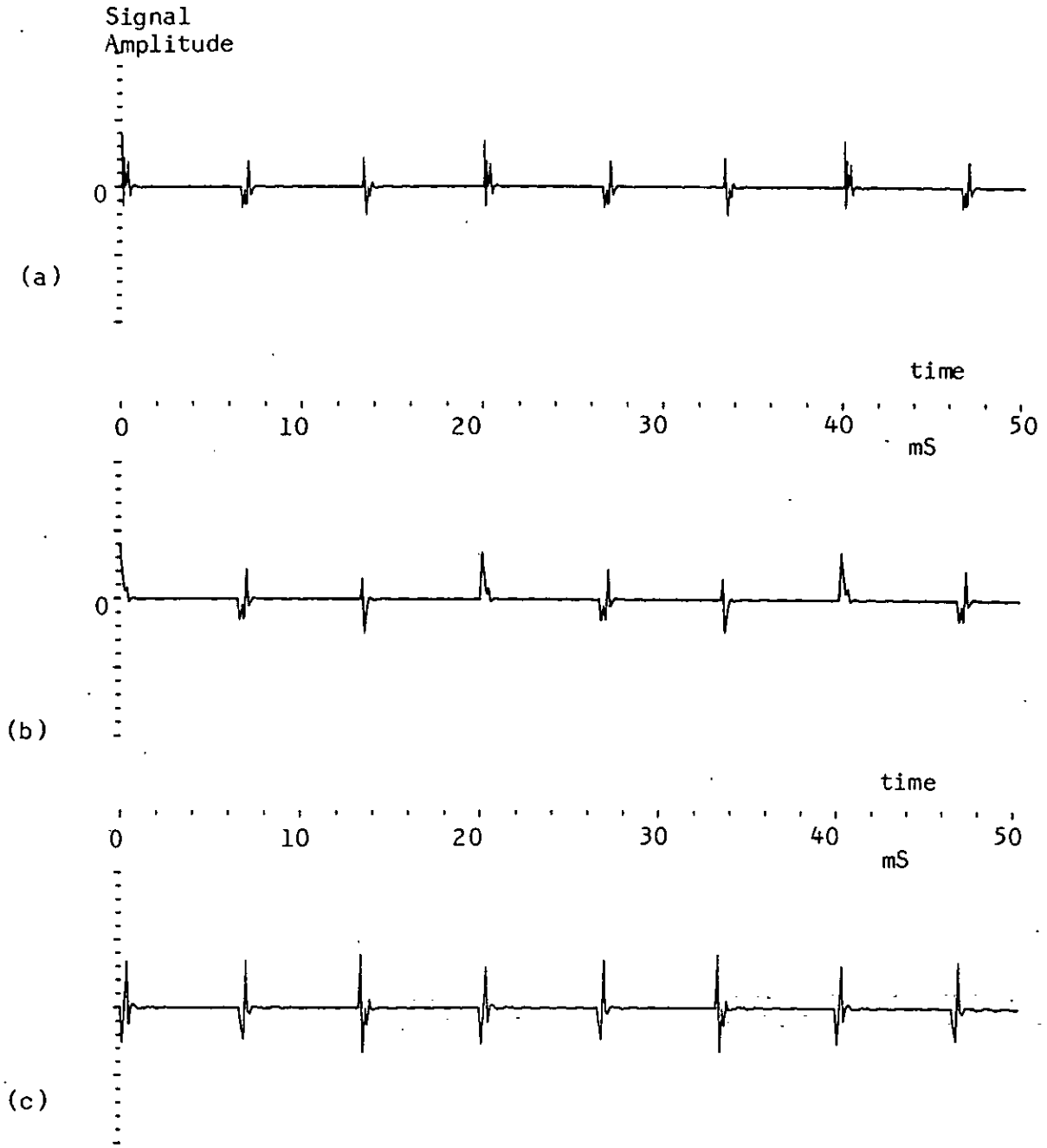


Fig. 5.14 (a) Residual signal from RELPC processing of the signal of Fig. 5.4(a); processed in contiguous 32mS sections, f_s=16kHz, 14-pole fit.
(b) Same residual as (a) but prefiltered with 1st-order lowpass filter (f_c=4kHz), then decimated by 2 (f_{so}=8kHz).
(c) Same residual as (a), no prefiltering, simple decimation by 2 (f_{so}=8kHz).

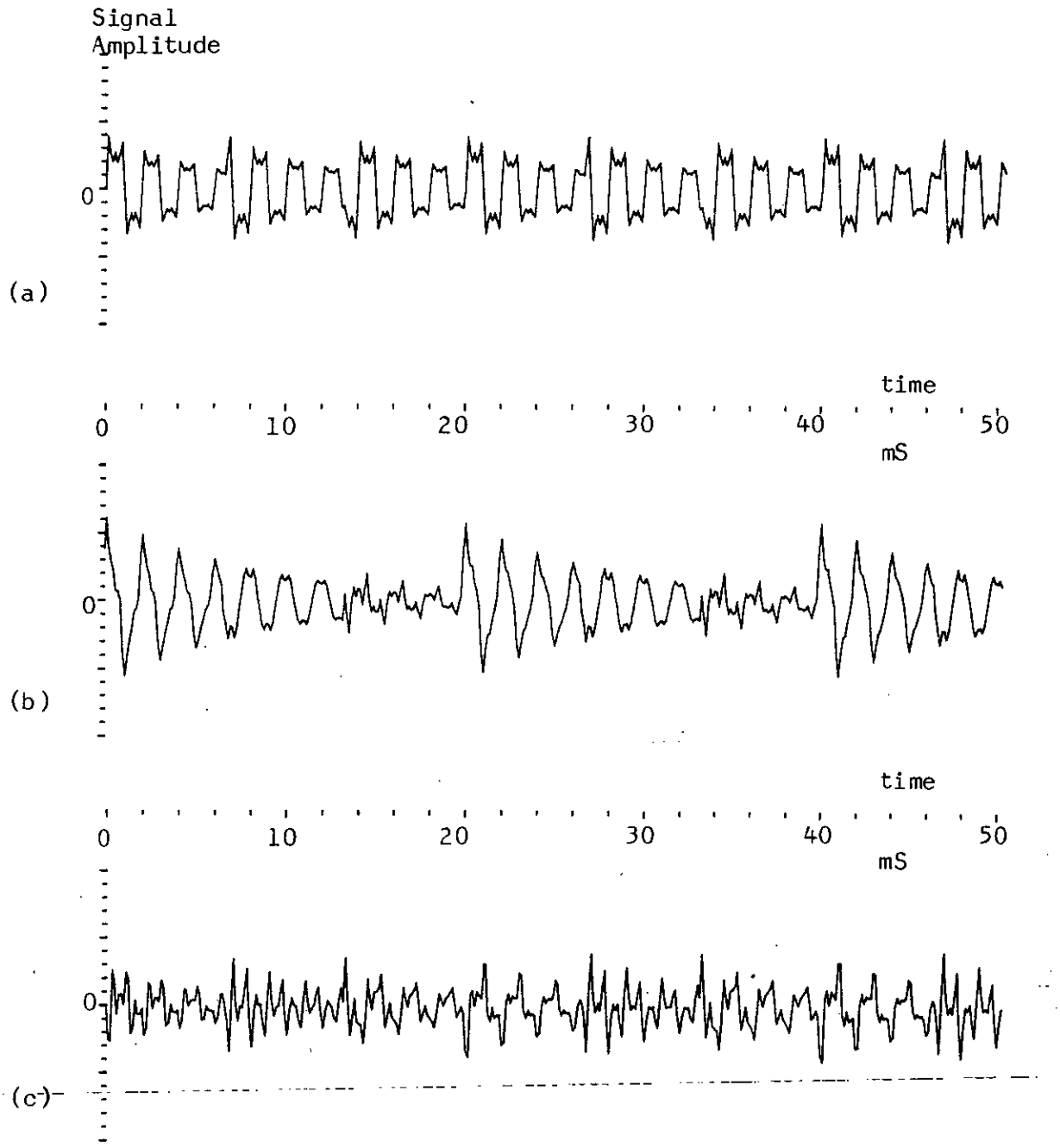


Fig. 5.15 (a) Signal from impulse-excited exact LTI filter system with frequency response having resonances nominally at 0.5kHz, 1.5kHz and 2.5kHz, bandwidths 40Hz, 50Hz and 75Hz, and relative peak power amplitudes of 0dB, -6dB and -9dB respectively.
(b) Signal synthesised by RELPC processing and a.r.f. excitation by the residual signal of Fig. 5.14(b).
(c) Same a.r.f. as in (b) but excited by the residual signal of Fig. 5.14(c).

prefiltering of the residual is strictly unnecessary and indeed adversely affects the resulting signal.

For comparison, some 5 seconds of the exact signal shown in Fig. 5.15(a) can be heard in rec. A13(a), and the signal resulting from processing by the RELPC system as in Fig. 5.15(c) can be heard in rec. A13(b). Notice that although rec. A13(b) sounds as if the fundamental has been altered slightly, the general timbre of the signal compares favourably with that of rec. A13(a). The reasons for the apparent subjective change in the fundamental are thought to be due to periodic over-excitation of the synthesis filter by one of the residual impulses. The effect is very noticeable in this case since the filter characteristics are exactly stationary over the 5 seconds period and the excitation function is exactly periodic with period = 6.7mS. It is expected that for speech, where the vocal tract filter is nonstationary in the long term and the excitation function is at best quasiperiodic, this effect will be minimised.

Power spectra for the exact signal of Fig. 5.15(a) and the signal of Fig. 5.15(c) as synthesised by RELPC processing can be compared in Fig. 5.16(a-b) respectively. It can be seen that RELPC processing has produced a signal whose spectrum demonstrates spectral resonances at the correct frequencies with the correct relationship between spectral peak power amplitudes. A comparison of the fine spectral detail also shows that the fundamental frequency has been conserved during processing.

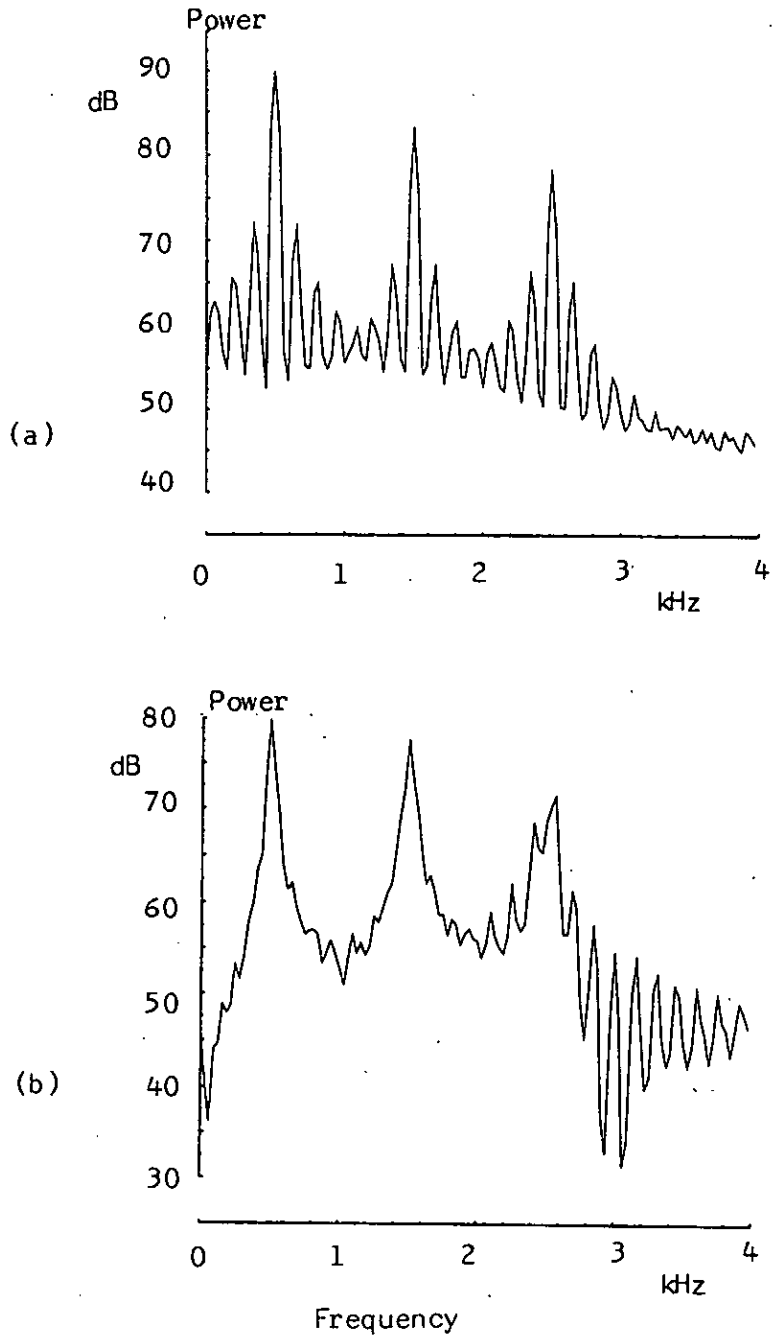


Fig. 5.16 (a) Power spectrum from windowed FFT of a 32mS section of the exact LTI system signal of Fig. 5.15(a).
(b) Power spectrum of same time section extracted from the RELPC synthesised signal of Fig. 5.15(c).

5.2.4 SIGNAL PREEMPHASIS AND HELIUM SPEECH SPECTRAL AMPLITUDE CORRECTION

It is well-established that computational considerations in autoregressive signal processing, as applied to speech, prefer the use of preemphasis in order to improve (reduce) the dynamic range of the speech spectrum [125]. Preemphasis is used here to improve estimation of the p.e.f. and a.r.f. coefficients, since the spectral whitening process of inverse filtering using LPC techniques tends to be most effective in areas of high spectral energy. Preemphasis filtering is therefore necessary to raise the power spectrum of speech at high frequencies since, for normal voiced speech in air, the effective spectral slope decay with frequency is of the order of -6dB/octave above the fundamental, see section 2.1.1. The signal power spectrum must therefore be elevated, prior to AR processing, by +6dB/octave if ill-conditioning is to be avoided.

In the continuous frequency/time domain, a power spectrum emphasis of +6dB/octave from some angular frequency w_u can be described, in the Laplace transform domain, by $(s + jw_u)$, where s is the complex frequency Laplace variable. The inverse transformation to the time domain yields an equation:

$$y(t) = \frac{dx(t)}{dt} + f_u \cdot x(t) \qquad (5\text{-}23)$$

where $x(t)$ is some signal and $y(t)$ is the resulting signal whose power spectrum is elevated by +6dB/octave, from frequency f_u, with respect to the spectrum of $x(t)$. In discrete time.

$$\frac{dx(t)}{dt} = \frac{x(nT) - x((n-1)T)}{T_n - T_{n-1}} \qquad (5\text{-}24a)$$

and

$$x(t) = x(nT) \qquad (5\text{-}24b)$$

But $1/(T_n - T_{n-1}) = 1/T_s = f_s$, where f_s is the sampling frequency.

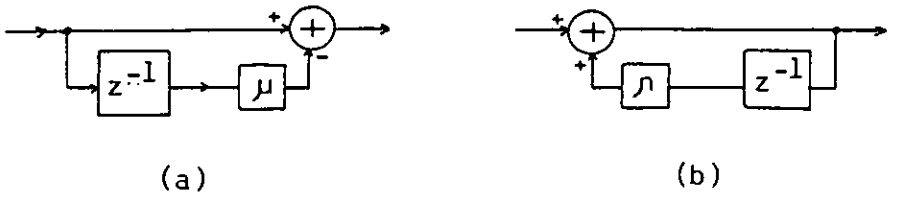


Fig. 5.17 (a) First-order digital preemphasis filter and (b) first-order digital deemphasis filter.

Therefore equation (5-23) becomes:

$$y_n = K[x_n - \mu . x_{n-1}] \qquad (5\text{-}25a)$$

and $$\mu = f_s/(f_s + f_u) \quad , \quad K = (f_s + f_u)/f_u \qquad (5\text{-}25b)$$

where μ is the preemphasis factor. Note that for f_u much less than f_s, then μ is very close to 1.0, and its exact value is found to have little effect on ill-conditioning providing $\mu \simeq 1.0$. Thus, some constant fixed value for f_u can be readily assumed, and is generally taken to be somewhere in the expected range of Fo values [32], $60\text{Hz} \leqslant f_u \leqslant 450\text{Hz}$. The practical implementation of equ.(5-25b) is usually taken to be a first-order differentiating filter, with transfer function $1-\mu z^{-1}$, of the form of Fig. 5.17(a), and the amplification factor K is generally not used. Note that with a quantised signal, then the level of y_n in equ.(5-25a) may be very small, particularly for f_s much greater than f_b, where f_b is the effective bandwidth of the signal, in which case $x_n \simeq x_{n-1}$. Thus, f_s should be chosen to be as low as is practically possible.

In terms of the RELPC unscrambler system, preemphasis is particularly advantageous since it was shown, in section 3.4.5, that the helium speech spectral amplitude decay for voiced speech was also of the order of -6dB/octave. Thus, if signal preemphasis is applied, spectral correction within the RELPC system for the helium speech effect can be

confined to formant frequency relocation and formant bandwidth manipulation.

The application of preemphasis does not invalidate the discussions of section 5.2.3 in respect of the 'unknown' frequencies on compression of the spectrum, and it is still necessary to assign a finite, constant power amplitude value to these frequencies. The residual signal resulting from the application of the inverse filter to the preemphasised signal is likewise unaffected: the aim of the processing is still to produce a residual whose spectrum is maximally white, and indeed preemphasis, from the above discussion, is specifically employed to improve this requirement.

The signal at the system output resulting from the residual excitation of the autoregressive filter will require a deemphasis of -6dB/octave to restore the required spectral decay corresponding to normal air speech, and this can be achieved by a digital integrating filter of the form of Fig. 5.17(b) with transfer function $\frac{1}{1-\rho z^{-1}}$, where ostensibly, $\rho=\mu$. Investigation [211] has shown, however, that an undesirable effect of preemphasis is to effectively attenuate the low frequency spectral region below f_u, thereby creating an area of low spectral energy which is poorly matched by the subsequent inverse filtering process. Recalling from section 3.3.1 that AR spectral analysis matches best to areas of high spectral energy (resonances), then too generous an estimate for the spectral energy is assumed in that region. Subsequent resynthesis of the signal through the corresponding a.r.f. therefore exaggerates the very low frequency signal components, which may produce discordant speech since human hearing is reported to be very acute at such low frequencies. However, a solution to this problem has been found by forcing $\rho<\mu$, so attenuating the very low frequency spectrum below f_u.

5.2.5 CORRECTION FOR THE HELIUM SPEECH FORMANT SHIFT CHARACTERISTIC

The formant frequency correction algorithm proposed for use with the RELPC unscrambler system, within the power spectrum, is identical to the FFT index mapping procedure implemented in the simulation of the STFT unscrambler system, as discussed in section 4.2.6. In this case, the spectral power is remapped using the linear interpolation procedure of equ.(4-24), which can be implemented by the use of two look-up tables, the first relating helium speech FFT index n to air index k, and the second giving the non-integer interpolation constant multiplying $P(n+1)-P(n)$, where $P(n)$ is the power amplitude at index n. Note that the requirement to produce the power spectrum using a double-length FFT padded with zeros, as discussed in section 5.2.2, inherently improves the interpolation process by effectively doubling the spectral resolution, which will give a more accurate representation of the corrected speech spectrum. The process of simple redefinition of the frequency axis for an overall spectral compression ratio $R \geqslant 2.0$ again improves frequency correction by similar enhancement of the spectral resolution.

The power spectrum of the voiced vowel "*er*" extracted from helium speech from "The Rainbow Passage" [121] at 100ft depth, sampled at 20kHz, can be seen in Fig. 5.18. The resulting normal air-speech spectrum, calculated from the same segment of speech after processing by the RELPC system, can be seen in Fig. 5.19. It can be seen that formants have been spectrally shifted down in frequency and that their amplitudes have been corrected for attenuation in the heliox environment by the use of preemphasis followed by post-processing deemphasis. Note in particular that the underlying fine harmonic structure of the original helium speech spectrum of Fig. 5.18 has been totally conserved. The frequency correction applied in this case corresponded to the average

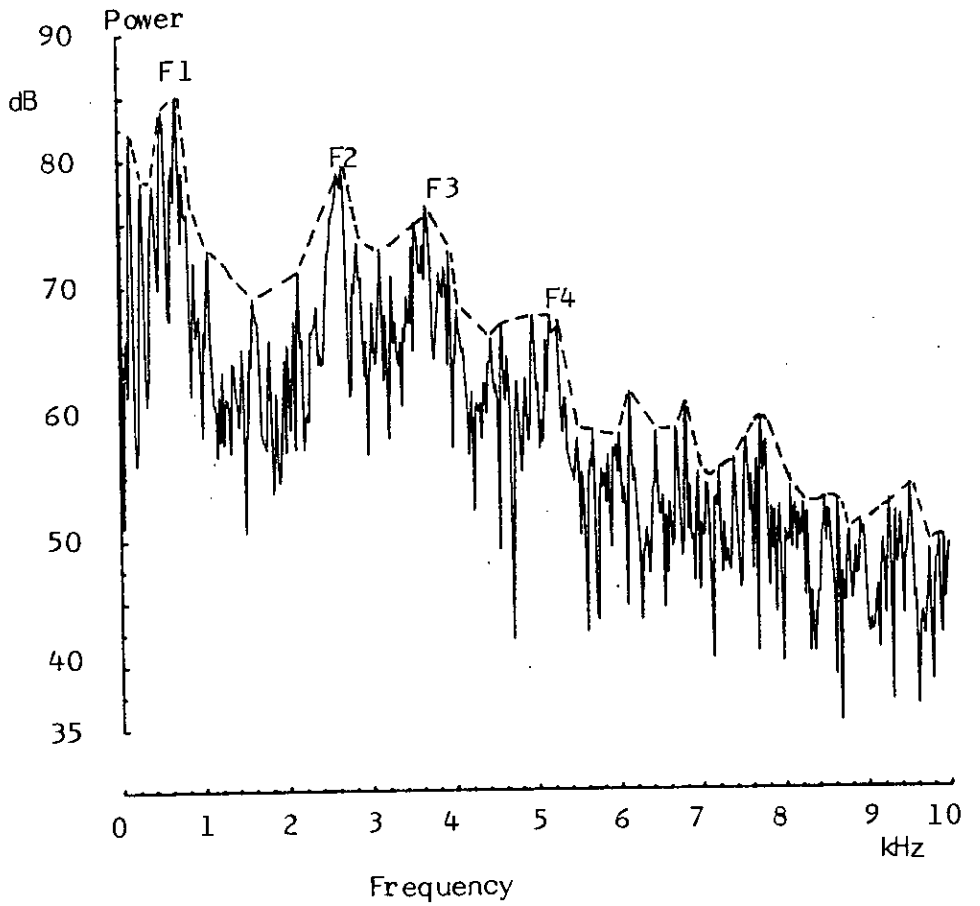


Fig. 5.18 Power spectrum of 51.2mS section of the vowel /ɚ/ spoken at 100ft depth in heliox, f_s=20kHz, windowed by Hamming window. (See Fig. 3.3 for heliox parameters).

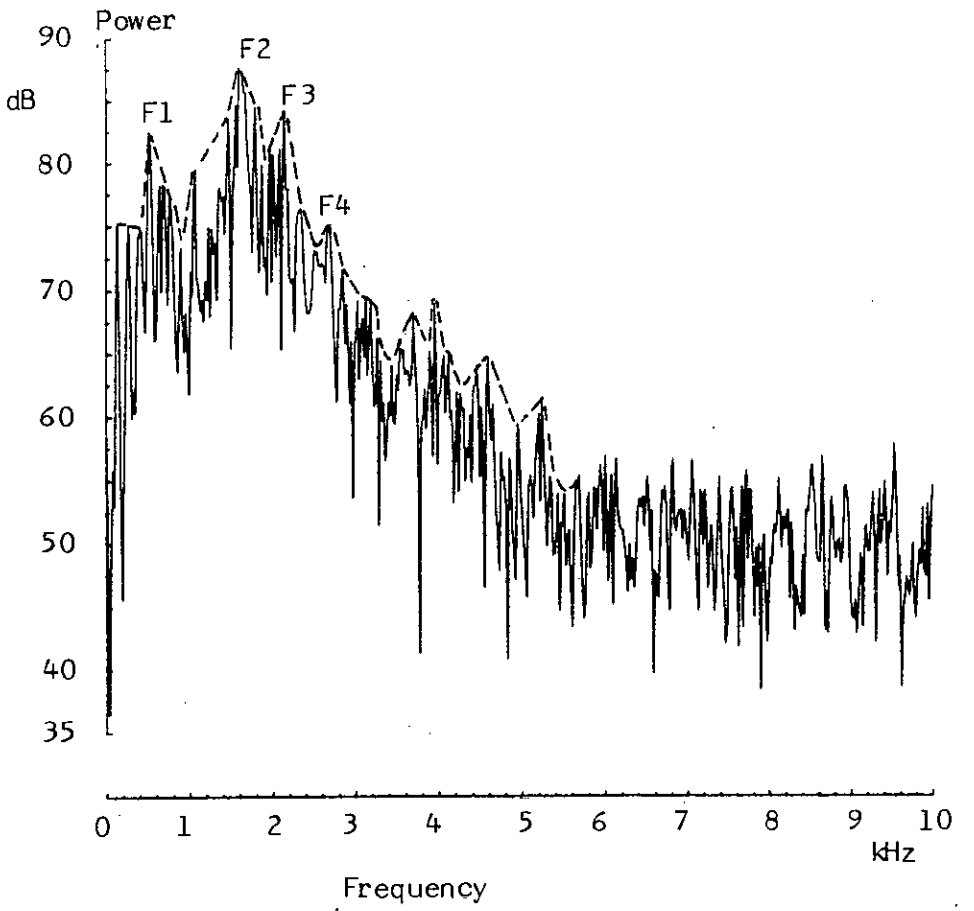


Fig. 5.19 Power spectrum of 51.2mS section of the vowel /er/ spoken at 100ft depth in heliox but corrected for the helium speech effect by RELPC processing. f_s=20kHz, windowed by Hamming window.
RELPC processing carried out in 25.6mS contiguous sections, 14-pole fit, μ=0.993, ρ=0.75, R=1.84.

of the piecewise linear shift characteristics for 100ft depth given in section 3.4.2, Fig. 3.13, that is, R = 1.84.

The proposed RELPC unscrambler system structure used to correct the helium speech effect is shown in Fig. 5.20, and the resulting unscrambled speech from its application to "The Rainbow Passage" spoken at 100ft in heliox (see Fig. 3.3), with sample rate and correction characteristic as specified above, can be heard in rec. A14.

5.3 CONCLUSIONS

A new system for unscrambling helium speech suitable for electronic implementation in real time has been presented, based on autoregressive (AR) signal processing. This new technique is consistent with the model of the speech mechanism as a linear time-invariant all-pole filter system in the short term over 10-30mS. The residually-excited linear predictive coding (RELPC) helium speech unscrambler system presented here improves on existing unscrambler systems by the complementary use of both AR and Fourier transform processing techniques. The fidelity of the unscrambled speech is improved both by maintaining continuity of the resulting speech, and by the use of the power spectrum together with AR filtering, which avoids the undesirable signal distortion otherwise resulting from direct use of the Fourier magnitude and phase spectra as detailed in section 4.3.

Ease of correction for the helium speech effect is promoted by direct manipulation of the power spectrum, which is used as an intermediate stage in the calculation of the AR synthesis filter coefficients. The use of the power spectrum in this way to implement the Weiner-Kinchine

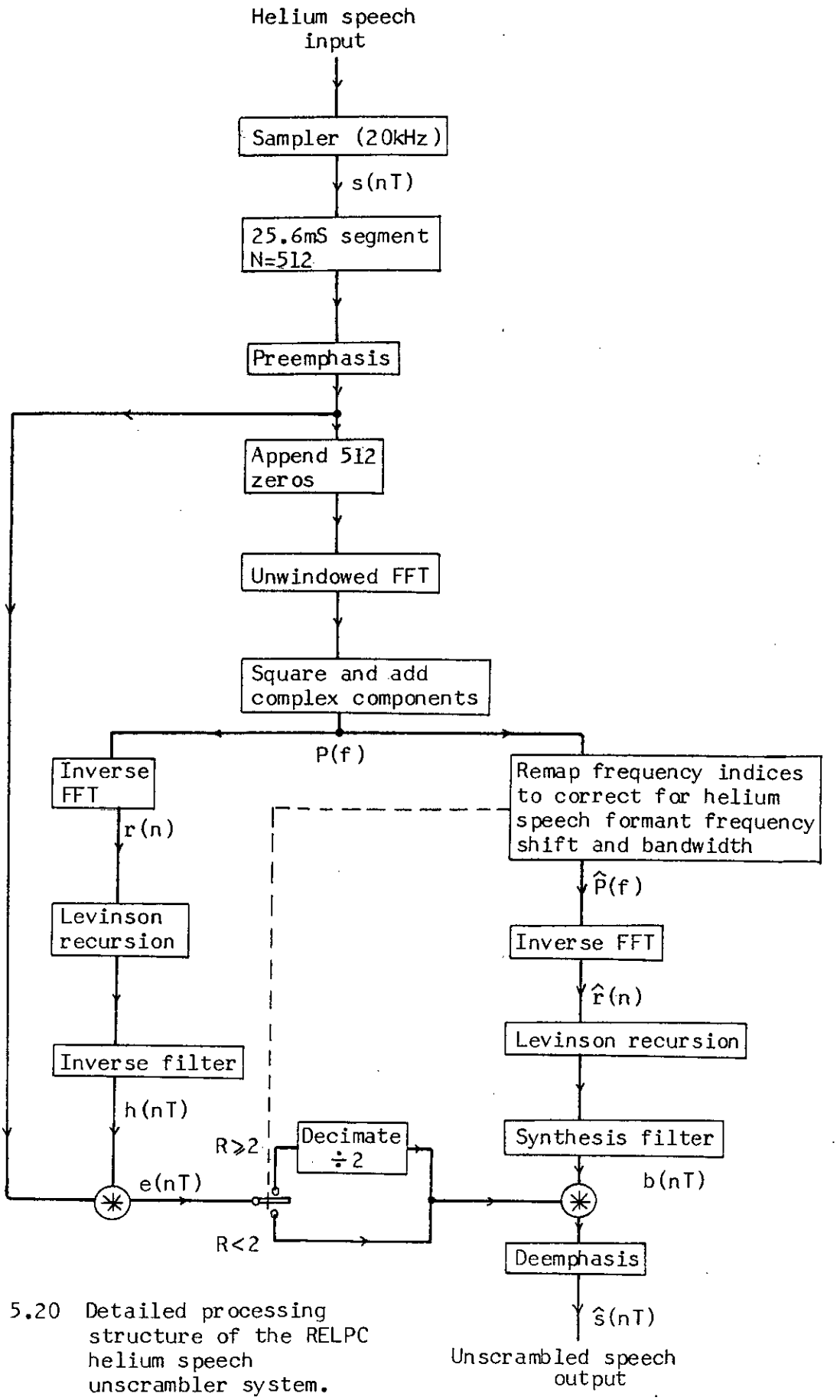


Fig. 5.20 Detailed processing structure of the RELPC helium speech unscrambler system.

theorem, so producing the autocorrelation sequence for the input helium speech data segment of length N points, demands the use of a discrete Fourier transform of length 2N, inclusive of N additional zeros appended to the input data segment. Whilst this necessarily complicates the DFT calculations, the resulting spectrum has effectively doubled the available resolution, see section 5.2.5, which improves the accuracy of the helium speech spectrum frequency correction (index mapping) algorithm.

Spectral correction can further be simplified by a consideration of the input speech data as a real-valued data sequence. A version of the fast Fourier transform has been proposed [205], for real time microprocessor-based implementation, which produces a spectrum with components in the range 0Hz - $f_s/2$Hz only, f_s = sample frequency, so obviating any consideration of the upper image half of the spectrum, thereby cutting by one half the operations otherwise required to achieve helium speech correction.

Autoregressive synthesis filter stability was shown, in section 5.2.3, to depend upon the values ascribed to the unknown frequency components in the upper regions of the corrected power spectrum. A consideration of the process from a maximum entropy standpoint, however, demonstrated that a.r.f. stability can be maintained by the relatively simple solution of setting the power amplitude of these frequencies to some finite non-zero value, so maximising the entropy in that region of the spectrum commensurate with the requirements for AR processing. The requirement to manufacture estimates for the unknown frequency components can be reduced, for overall spectral compression ratios $R \geqslant 2.0$, by simple redefinition of the frequency axis, which has the added advantage of doubling once more the available resolution, favouring frequency

correction by use of the index mapping algorithm detailed in section 4.2.6.

Signal preemphasis is a well-established requirement to increase the accuracy of AR spectral analysis, when applied to voiced speech signals. In the RELPC system, preemphasis followed by appropriate deemphasis of the unscrambled speech signal has the additional advantage of implicitly correcting formant amplitude with no need to apply explicit spectral amplitude correction algorithms in the frequency domain, which once again simplifies overall correction for the helium speech effect.

The application of preemphasis to unvoiced speech is, in normal circumstances, not required since the unvoiced speech spectrum exhibits a reduced formant amplitude decay, compared to voiced speech, due to the glottal excitation source characteristics. An analysis of the spectral characteristics of unvoiced helium speech has not been possible in the research presented in this thesis. However, it is possible to speculate on its relationship to RELPC system processing from a consideration of the discussions in section 2.2.3 and the results presented in sections 3.4.2-3 regarding voiced helium speech formant frequency translation.

It was shown, in section 2.2.4, that the time waveform of unvoiced helium speech is much reduced in intensity compared to voiced speech. Unvoiced speech in normal air contains spectral formants, significant to perception, in the range 3-7kHz. Assuming that the unvoiced speech spectrum is affected in a manner similar to the voiced speech spectrum, then these frequencies translate in high-percentage helium atmospheres to approximately 8-20kHz. Thus, certain high frequency unvoiced formants in helium speech may be outwith the range of normal human hearing, and in particular, may also be outwith the bandwidth of microphone and

recording equipment, which may account in part for the reduced intensity. Assuming that total spectral energy is conserved in the translation from unvoiced air to helium speech, then unvoiced formant energy will be smeared over a larger frequency range, so increasing unvoiced formant bandwidth. Whilst formant frequency relocation will restore bandwidth, spectral amplification will be required to restore relative power levels between voiced and unvoiced speech. This could be achieved by the preemphasis-deemphasis procedures, which would be beneficial in the case where unvoiced formants are shifted nonlinearly in frequency, and particularly where high frequency formants have been translated upwards in frequency by a greater amount than the low frequency formants.

In conclusion, taken in relation to the properties of helium speech and the corrections required to restore normal intelligibility, the RELPC system detailed here represents an improved solution to the helium speech unscrambling problem. A paper outlining the elementary theory of the RELPC helium speech unscrambler system is contained in Appendix A.

CHAPTER VI

SUMMARY AND SUGGESTIONS FOR FUTURE RESEARCH.

6.1 THE HELIUM SPEECH EFFECT AND RELATED ACOUSTIC ANALYSES

The requirement to carry out prolonged diving operations at depths below 150ft demands the use of helium-oxygen (heliox) respiratory mixtures in order to avoid potentially lethal physiological trauma. However, the physical properties of this extraordinary gas mixture, coupled with the effects of high ambient pressure, combine to distort the diver's speech, rendering it unintelligible in the main to the listener.

Since the inception and common use of heliox mixtures in the 1960's, the need to unscramble the diver's "helium speech" using electronic means in order to restore normal intelligibility to the speech waveform has been identified. It has been the purpose of the research presented in this thesis to determine the specifications for a new and improved helium speech unscrambler system in order to enhance speech intelligibility in respect of the listener.

As with all signal processing problems, it is important to specify the desirable characteristics of the end-product signal and also to identify the apposite processing strategy to be applied to the input signal to achieve these characteristics.

Chapter II has studied in detail the normal speech waveform in air in order to evince those criteria important in the production and perception of human speech. The speech mechanism can basically be

considered as a linear time-invariant (LTI) filter system in the short term over periods of time of the order of 10-30mS. The vocal tract, comprising the complex acoustic tube from the glottis to the lips and including the nasal cavities, can be modelled as an all-pole filter system which is excited either by a quasiperiodic impulse train for voiced speech or by a white noise source for unvoiced speech, or indeed by a combination of both types of excitation.

Speech perception has been shown to rely on a complex interplay of both temporal and spectral events in the acoustic waveform. For voiced speech, phonemic differentiation has been shown to depend on the relative frequency locations of spectral resonances (formants) irrespective of the fundamental frequency of pulse excitation of the vocal tract. Phonemic quality, on the other hand, depends on the relative peak powers and bandwidths of the formants of voiced speech. Perception of the unvoiced speech waveform can be shown also to depend to some extent on spectral formant structure, but is also dependent on other perceptual cues such as formant frequency transitions and the vowel-consonant intensity ratio.

From a survey of the existing literature on the helium speech phenomenon, the presently-known effects of pressure and high-percentage helium respiratory mixtures on the speech mechanism and speech perception have been summarised. It has been shown that the primary effects of a pressured heliox atmosphere on the speech waveform are that, firstly, type and fundamental frequency of glottal excitation are conserved and secondly, that the short and long term speech spectrum is shifted upwards in frequency. Whilst there is overall agreement as to the gross acoustic distortions of the helium speech effect, there are nonetheless areas of contention as regards the detailed nature of these distortions,

particularly in terms of linear/nonlinear formant translation, effect of translation on formant bandwidth and differential effects on voiced/unvoiced waveforms.

The brief summary of currently in-service helium speech unscramblers, which concludes Chapter II, demonstrates that all available unscrambler systems treat the speech mechanism as an LTI filter system in the short term, and that most systems provide only for linear correction of the helium speech effect, mainly to facilitate real time processing and cost-effective systems.

Despite the wide variety of processing algorithms employed in these systems, there has been cause to severely criticise their performance in terms of the speech intelligibility they afford. One immediate cause to which this inadequate performance can be ascribed is that perhaps there are erroneous assumptions made by these systems pertaining to the acoustic events in helium speech, or indeed they may be deficient in their provisions for processing certain acoustic attributes of the signal which have an important bearing on intelligibility.

This proposition has been explored in Chapter III, where new recordings of normal air and helium speech for the same subject have been procured. This allows a direct comparison between acoustic events in the several different atmospheres encountered during the experiment. The acoustic analysis has been confined to voiced speech only, however, since recording bandwidth considerations coupled with the expected translation of the speech spectrum preclude the possibility of a reliable analysis of unvoiced speech.

The results of the analysis of fundamental frequency have confirmed that there is no change in the rate of excitation of the vocal tract from a normal air to a pressured heliox environment. Explanations for

observed variations in fundamental frequency have been proposed in terms of psychoacoustic causes and, while necessarily conjectural in the sense that no corroborative data such as psychological stress could be measured at the time, are nonetheless founded on events known to occur in normal air speech.

Under the assumption of the speech mechanism as an LTI filter system in the short term, autoregressive (AR) spectral analysis techniques have been used to study the distortion of voiced formants from air to a heliox environment.

The overall translation of formant frequencies has been shown to be nonlinear in nature, although it is proposed here that a two-segment piecewise linear curve will adequately characterise the formant translation for the speech recorded at each depth. The research has, however, perhaps posed more questions in respect of the formant translations than it sought to resolve, and new phenomena have been reported from a consideration of the formant translations of each phoneme on an individual basis.

It has been demonstrated that each of the investigated phonemes is shifted differently from another but in a consistent manner from depth to depth. The immediate implication from these results is that speech recognition techniques must become an integral part of helium speech unscrambler systems, since each phoneme requires a different correction algorithm. However, from a consideration of the physical properties of the heliox mixture and known effects of modified auditory feedback in normal air, it is possible to speculate that simple preprocessing involving delayed auditory feedback of the diver's own voice may help linearise the phoneme-specific formant translations. Certainly further research is required in the first instance to fully characterise this

phenomenon, since the results presented here are related to the speech of a single subject.

New results have also been presented, again based on an investigation of individual phoneme spectra, which show that new formants occur at depth which cannot be directly related to the known spectra in air. The theoretical translation of these formants back to normal air values suggests that they may be the result of actual nasal resonance, although the exact mechanism of their production is a matter of some conjecture. Once again, these results require further investigation using several subjects before they can definitely be considered as a pandemic feature of the helium speech effect, except to say that, for the present, they support the stategy of helium speech unscrambling using direct spectral correction in the frequency domain.

Formant amplitude distortion in helium speech has been shown in Chapter III to correspond well to a simple translation of the vocal tract frequency response along the recognised slope decay of -6dB/octave, due to the combined effects of the glottal source and lip radiation characteristics, which supports the assumption that the glottal source characteristic is invariant from air to heliox. Helium speech formant amplitude correction has therefore been identified as being necessary to restore normal voice quality to the unscrambled speech. An investigation of helium speech formant bandwidth has also tended to support the hypothesis that formant Q-factor is conserved at high frequencies but reduced in the low frequency formant region, implying that additional formant bandwidth manipulation is required other than that achieved by helium speech formant frequency correction by spectral compression.

The recordings used for the analyses presented in Chapter III are contaminated with noise from a gas filter system. Analysis of the

characteristics of this noise has shown it to be Gaussian in nature, and this has facilitated an estimation of its effects on the foregoing AR spectral analyses. It has also been shown that, for visual identification of formant data, the critical signal-to-noise ratio for AR spectral analysis is of the order of 6dB, below which the smooth spectrum is so contaminated by extraneous spectral detail that formant identification becomes extremely difficult.

Taken overall, the results relating to the gross spectral distortions of the helium speech effect support a consideration of the speech mechanism as a short term LTI filter system, but require an unscrambling technique capable of providing nonlinear spectral correction to restore normal intelligibility.

6.2 HELIUM SPEECH UNSCRAMBLER SYSTEMS

Chapter IV has approached criticism of poor unscrambler performance from the point of view that there may be deficiencies in the signal processing strategies, implemented in unscrambler systems, which may affect the resultant unscrambled speech in some manner which is antagonistic to good intelligibility.

Two systems have been investigated in detail. The first is a simple, cost-effective unscrambler system based on time domain segmentation and expansion of the speech waveform in synchronism with the pitch period. The very nature of this device renders it essentially incompatible with the corrections required for the helium speech effect. It has been shown to be confined to only linear compression of the frequency spectrum, and whilst some formant amplitude compensation is

possible by simple prefiltering, modifications to the prefilter characteristic are not simple to achieve, which therefore excludes selective amplitude correction with depth. Additionally, the dependence of this technique on simple threshold peak detection to provide pitch synchronism aggravates intelligibility by producing discontinuities in the unscrambled speech. Modification of the device to provide continuity of speech output has been shown to improve intelligibility. However, this modification has been demonstrated to actually worsen intelligibility when significant levels of ambient noise contaminate the speech signal. The subject of ambient noise is an important topic, and is discussed again later in section 6.3.

The advanced unscrambler system based on the short-time Fourier transform (STFT) has also been examined in Chapter IV. This system affords continuity of speech output and nonlinear correction of formant frequency and amplitude characteristics, and is based on the assumption of the speech mechanism as a short term LTI filter system. Furthermore, no explicit time domain pitch detection is required, with type of glottal excitation and fundamental frequency being conserved by direct manipulation of the Fourier magnitude spectrum. Estimation of the spectral envelope has been shown to depend critically on the value adopted for the normalised slope factor. Correction of this magnitude envelope is then achieved by a frequency index remapping and interpolation technique. Indeed, all the correction for the helium speech effect is applied to the magnitude spectrum with no corrective operations whatsoever being exercised upon the associated Fourier phase spectrum, the implicit assumption being that the human auditory system is relatively insensitive to spectral phase. This has been shown to be a false premise, however, with the ear being insensitive only to linear

phase distortions. In contrast, the equivalent phase distortions produced by nonlinear correction of the magnitude spectrum with retention of the original phase spectrum are highly nonlinear, and can produce serious acoustic aberrations in the signal synthesised through the inverse Fourier transform, and may therefore affect the ultimate fidelity of the unscrambled speech afforded by this system.

Improvements in unscrambler design must therefore seek to maintain ease of nonlinear correction of the helium speech effect through the use of the signal spectrum and implicit conservation of the glottal excitation source. A method of improving the accuracy of frequency index remapping and interpolation is desirable, however, and also the enigma of spectral magnitude and phase must be resolved.

This challenge has been taken up in Chapter V, and has resulted in the proposal of a new unscrambler system based on AR signal processing techniques. The residually-excited linear predictive coding (RELPC) helium speech unscrambler system preserves the glottal excitation characteristic by inverse filtering of the speech waveform to produce a residual signal. This residual is then used to excite a synthesis filter, corresponding to the vocal tract, whose frequency response has been corrected for the helium speech effect.

Ease of correction for the helium speech effect is maintained through use of the Fourier transform. However, use is made of the power spectrum in this case to obtain the autocorrelation series necessary in construction of the synthesis filter. This method of performing autocorrelation is advantageous since the power spectrum is cophase, thereby obviating any consideration of spectral phase. Spectral resolution is also immediately doubled using this method by requirements to force linear autocorrelation from what is basically a cyclic autocorrelation

process, thereby improving the accuracy of frequency index remapping and interpolation.

From a consideration of the maximum entropy properties of AR processing, it has been shown that no special precautions are required to maintain the fine harmonic structure of the power spectrum when correction for the helium speech effect is applied. Synthesis filter stability has been shown to be assured by maintainance of the maximum entropy condition for those frequencies in the spectrum about which no information is otherwise known after spectral correction has been applied. In the application of the fast Fourier transform in obtaining the power spectrum, it has been shown that a consideration of the same maximum entropy conditions permit a doubling once again of the effective spectral resolution through a simple redefinition of the frequency axis for an overall spectral compression by a factor of 2 or greater, which again improves the accuracy of formant frequency correction by frequency index remapping and interpolation. This redefinition requires downsampling of the residual signal, however, and it has been demonstrated that care must be taken in the decimation procedure to avoid inadvertent distortion of the resulting unscrambled speech waveform.

Preemphasis is generally recommended to be applied to the speech signal in AR modelling to avoid ill-conditioning of the autocorrelation function as applied to calculation of the prediction error filter coefficients. In this application to correction for the helium speech effect, preemphasis not only improves the effectiveness of the AR technique, but additionally, when followed by appropriate deemphasis, automatically corrects for the formant amplitude distortions in helium speech, leaving only the tasks of formant frequency relocation and formant bandwidth manipulation to be performed on the spectrum.

In relation to the properties of helium speech and the corrections required to restore normal intelligibility, the RELPC unscrambler system represents an improved solution to the helium speech unscrambling problem.

6.3 CONCLUDING REMARKS

Electronic implementation of the RELPC unscrambler system in real time is considered to be possible, and particularly with the advent of very fast custom-design correlator [212] and FFT integrated circuits [213], together with advances in the design of high-speed pipelined microprocessor architectures [214], it is thought that significant processing bandwidth can be achieved, and it is suggested that this topic is a natural sequel to the research presented in this thesis.

The present research has considered enhancement of the intelligibility of helium speech from the characterisation of the acoustic phenomena of helium speech itself and improvements to the processing architectures applied to the speech waveform. An inherent problem of helium speech unscrambling, however, relates to the addition of at times high levels of background noise, and this subject has been touched upon in Chapter III.

Using the RELPC technique, it is possible to compensate for coloured noise possessing stationary characteristics by direct subtraction of the noise spectrum from the speech spectrum prior to correction for the helium speech formant frequency shift, and this technique is equally applicable to the STFT unscrambling method of Chapter IV. Similarly, the masks and helmets worn by divers are known to influence the spectral

characteristics of the divers' speech [215,216], and this effect too is therefore correctable by manipulation of the power spectrum.

An analysis of the ambient noise, in Chapter III, has shown that the background noise may, however, be Gaussian in nature, that is, having stationary characteristics, which renders compensation through spectral subtraction impossible. The use of AR processing in the RELPC system, on the other hand, opens up the possibility of noise cancellation using adaptive filtering techniques [217,218], which basically require a second signal input to the processor system from the noise source itself, therefore implying that the noise source must be accessible. To date, little work has been done in the identification and characterisation of the various noise sources to be encountered in the diving environment, and so this too is suggested as an important area for research which will help improve the performance, in terms of intelligibility, of unscrambler systems.

Man's excursion into the deep-sea environment necessitates the use of extraordinary respiratory mixtures in order to sustain life. Exploitation of the ocean floor to extract organic fuels and minerals is at present confined to working depths of some 600ft or so, and therefore helium-oxygen mixtures are physiologically suited to the well-being of man at these depths. It has been already been proposed, however, that current oil resources in the known sites will be extinguished by the mid-21st century, and therefore exploration will become necessary at depths far in excess of those commonly encountered at present. This may mean that the gas mixtures required to sustain human life at these depths will be radically different from the heliox mixtures used at present: already it has been shown [219] that the nitrogen which is narcotic at present working depths is in fact physiologically

necessary for excursions beyond 800ft or so, which changes the physical properties of the respiratory gas, and therefore the associated speech distortions.

Whilst the RELPC unscrambler system proposed in this thesis involves complex signal processing techniques, it conserves the possibility of ease of adaptation to new respiratory mixtures since elimination of voice distortion is effected by correction of the short term spectrum, and therefore changes in the distortion characteristic require only changes in the spectral correction algorithm. In conclusion, unscrambler systems of this kind, although complex and at present expensive to implement, offer advantages of adaptation to the signal processing problem which may outweigh complexity and cost considerations in the long term.

References

ABBREVIATIONS

Acous.	Acoustical
Assoc.	Association
ed.	editor
IEE	Institution of Electrical Engineers (U.K.)
IEEE	Institution of Electrical and Electronic Engineers (U.S.A.)
IERE	Institution of Electronic and Radio Engineers (U.K.)
Inst.	Institute
Int.	International
J.	Journal
no.	number
pp.	pages
Proc.	Proceedings
Res.	Research
Soc.	Society
Tech.	Technology
Trans.	Transactions
v.	volume

References

(1) E. Halley, 'The art of living underwater: or, a discourse concerning the means of furnishing air at the bottom of the sea, in any ordinary depths', *Philosophical Trans. of the Royal Soc. of London*, v.29, no.349, pp.492-499, Sept. 1716.

(2) A. Jaminet, 'Physical effects of compressed air and of the causes of pathological symptoms produced on man by atmospheric pressure employed for the sinking of piers in the construction of the Illinois - St. Louis bridge over the Mississippi River at St. Louis, Missouri' (R. and T. A. Enious, St. Louis, 1871)

(3) P. Bert, 'La pression de l'air et les êtres vivants', *La Revue Scientifique de la France et de l'Etranger*, no.3, pp.49-55, July 1876.

(4) J. M. Clark, 'Oxygen toxicity', in 'The physiology and medicine of diving', ed. D. H. Elliott (Baillere and Tyndall, London, 1982) .

(5) A. R. Behnke, R. M. Thompson and E. P. Motley, 'The psychologic effects of breathing air at 4 atmospheres', *American J. of Physiology*, v.112, pp.554-558, 1935.

(6) E. M. Case and J. B. S. Haldane, 'Human physiology under high pressure: I. effects of nitrogen, carbon dioxide and cold', *J. of Hygiene*, v.41, pp.225-249, 1941.

(7) D. H. F. Webster, 'The danger game' (Hale Publications, London, 1978).

(8) E. K. Hunter, 'Problems of diver communication', *IEEE Trans.*, v.AU-16, no.1, pp.118-120, March 1968.

(9) A. D. Baume, D. R. Godden and J. R. Hipwell, 'Procedures and language for underwater communication', UEG Technical Note no.26, Underwater Engineering Group, Jan. 1982.

(10) G. Fant, 'Acoustic theory of speech production' (Mouton, The Hague, 1960).

(11) E. N. Pinson, 'Pitch-synchronous time-domain estimation of formant frequencies and bandwidths', *J. of the Acous. Soc. of America*, v.35, pp.1264-1273, Aug. 1963.

(12) L. R. Rabiner and R. W. Schafer, 'Digital processing of speech signals' (Prentice-Hall, 1975).

(13) S. Wood, 'A radiographic analysis of constriction locations for vowels', *Working Papers, Phonetics Laboratory*, v.15, pp.101-131, (Lund University) 1977.

(14) J. Laver, 'The phonetic description of voice quality' (Cambridge University Press, 1980).

(15) Jw. van den Berg, 'Myoelastic/aerodynamic theory of voice production', *J. of Speech and Hearing Res.*, v.1, pp.227-244, 1958.

(16) A. E. Rosenberg, 'The effect of glottal pulse shape on the quality of natural vowels', *J. of the Acos. Soc. of America*, v.49, no.2, pp583-590, 1971.

(17) R. L. Miller, 'Nature of the vocal cord wave', *J. of the Acous. Soc. of America*, v.31, no.6, pp.667-677, 1959.

(18) J. L. Flanagan, 'Speech analysis, synthesis and perception' (Academic Press, New York, 1965).

(19) G. Fant, 'The acoustics of speech', *Proc. 3rd Int. Congress on Acoustics*, pp.188-201, 1959.

(20) J. Martony, 'On the vowel source spectrum', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.3-4, (Royal Inst. of Tech., Stockholm) 1964.

(21) P. M. Morse, 'Vibration and sound (2nd Edition)' (McGraw-Hill, New York, 1948).

(22) C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens and A. S. House, 'Reduction of speech spectra by analysis-by-synthesis techniques', *J. of the Acous. Soc. of America*, v.33, pp.1725-1736, Dec. 1961.

(23) J. L. Flanagan, 'Voices of men and machines', *J. of the Acous. Soc. of America*, v.51, pp.1375-1387, March 1972

(24) M. V. Mathews, J. E. Miller and E. E. David(Jr), 'Pitch synchronous analysis of voiced sounds', *J. of the Acous. Soc. of America*, v.33, no.2, pp179-186, 1961.

(25) J. L. Stewart, 'Fundamentals of signal theory' (McGraw-Hill, 1960).

(26) P. B. Denes and E. N. Pinson, 'The speech chain' (Bell Telephone Laboratories, 1963).

(27) G. Fairbanks and P. Grubb, 'A psychophysical investigation of vowel formants', *J. of Speech and Hearing Res.*, v.4, pp.203-219, 1961.

(28) H. Fujisaki and T. Kawashima, 'The roles of pitch and higher formants in the perception of vowels', *IEEE Trans.*, v.AU-16, pp.73-77, 1968.

(29) G. E. Peterson and H. L. Barney, 'Control methods used in the study of vowels', *J. of the Acous. Soc. of America*, v.24, pp.175-184, 1952.

(30) G. E. Peterson, 'The information-bearing elements of speech', *J. of the Acous. Soc. of America*, v.24, no.6, pp.629-637, 1952.

(31) A. Holbrook and G. Fairbanks, 'Dipthong formants and their movements', *J. of Speech and Hearing Res.*, v.5, pp.38-58, 1962.

(32) J. Laver and P. Trudgill, 'Phonetic and linguistic markers in speech', in 'Social markers in speech', ed. K. Scherer and H. Giles (Cambridge University Press, 1979).

(33) K. R. Scherer, 'Vocal indicators of stress', in 'The evaluation of speech in psychiatry', ed. J. Darby (Grune and Stratton, New York, 1981).

(34) C. E. Williams and K. N. Stevens, 'Emotions and speech: some acoustical correlates', *J. of the Acous. Soc. of America*, v.52, pp.1238-1250, 1972.

(35) L. Bjork, 'Velopharyngeal function in connected speech. Studies using tomography and cineradiography synchronised with speech spectrography', *Acta Radiographica*, v.56, p.399, 1961 (Book review).

(36) L. B. Lintz and D. Sherman, 'Phonetic elements and perception of nasality', *J. of Speech and Hearing Res.*, v.4, pp.381-396, 1961.

(37) A. S. House and K. N. Stevens, 'Analog studies of the nasalization of vowels', *J. of Speech and Hearing Disorders*, v.21, no.2, pp.218-231, 1956.

(38) A. S. House, 'Analog studies of nasal consonants', *J. of Speech and Hearing Disorders*, v.22, no.2, pp.190-204, 1957.

(39) O. Fujimura, 'Analysis of nasal consonants', *J. of the Acous. Soc. of America*, v.34, pp.1865-1875, 1962.

(40) J. Martony, 'The role of formant amplitudes in the synthesis of nasal consonants', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-3, pp.28-31, (Royal Inst. of Tech., Stockholm) 1964.

(41) C. F. Diehl and E. T. McDonald, 'Effect of voice quality on communication', *J. of Speech and Hearing Disorders*, v.21, pp.233-237, 1956.

(42) H. M. Moser, J. J. Dreher and S. Adler, 'Comparison of hyponasality, hypernasality and abnormal voice quality on the intelligibility of two-digit numbers', *J. of the Acous. Soc. of America*, v.27, pp.872-874, 1955.

(43) J. H. Heinz and K. N. Stevens, 'On the properties of voiceless fricative consonants', *J. of the Aous. Soc. of America*, v.33, pp.589-596, May 1961.

(44) 'Physical acoustics', ed. W. P. Mason (Academic Press, 1964).

(45) J. Martony, 'On the synthesis and perception of voiceless fricatives', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.17-22, (Royal Inst. of Tech., Stockholm) 1962.

(46) L. Lisker, J. Martony, B. Lindblom and S. Ohman, 'F-pattern approximations of voiced stops and fricatives', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.20-22, (Royal Inst. of Tech., Stockholm) 1960.

(47) J. Lindqvist and J. Lubker, 'Mechanisms of stop consonant production', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.1-2, (Royal Inst. of Tech., Stockholm) 1970.

(48) R. Carlson and B. Granstrom, 'Perception and synthesis of speech', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.1-16, (Royal Inst. of Tech., Stockholm) 1977.

(49) R. Carlson, B. Granstrom and S. Pauli, 'Perceptive evaluation of segmental cues', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.18-24, (Royal Inst. of Tech., Stockholm) 1972.

(50) P. C. Delattre, A. M. Liberman and F. S. Cooper, 'Acoustic loci and transitional cues for consonants', *J. of the Acous. Soc. of America*, v.27, no.4, pp.769-773, July 1955.

(51) K. N. Stevens and A. S. House, 'Studies of formant transitions using a vocal tract analog', *J. of the Acous. Soc. of America*, v.28, no.4, pp.578-585, July 1956.

(52) K. N. Stevens and D. H. Klatt, 'Role of formant transitions in the voiced-voiceless distinction for stops', *J. of the Acous. Soc. of America*, v.55, pp.653-659, 1974.

(53) B. Borovickova, V. Malac and S. Pauli, 'F2-transitions in the perception of Czech stops', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.14-15, (Royal Inst. of Tech., Stockholm) 1970.

(54) K. S. Harris, H. Hoffman, A. M. Liberman, P. C. Delattre and F. S. Cooper, 'Effect of third-formant transitions on the perception of the voiced stop consonants', *J. of the Acous. Soc. of America*, v.30, pp.122-126, 1958.

(55) G. Fairbanks and A. S. House, 'The influence of consonant environment upon the secondary acoustical characteristics of vowels', *J. of the Acous. Soc. of America*, v.25, pp.105-113, 1953.

(56) G. Fairbanks, A. S. House and K. N. Stevens, 'An experimental study of vowel intensities', *J. of the Acous. Soc. of America*, v.22, pp.457-459, 1950.

(57) B. L. Scott, 'Temporal factors in vowel perception', *J. of the Acous. Soc. of America*, v.60, no.6, pp.1354-1365, Dec. 1976.

(58) G. Fant, 'Sound, features and perception', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-2, pp.1-14, (Royal Inst. of Tech., Stockholm) 1967.

(59) S. E. G. Ohman, 'Perception of segments of VCCV utterances', *J. of the Acous. Soc. of America*, v.40, pp.979-988, 1966.

(60) A. M. Liberman, P. C. Delattre, F. S. Cooper and L. J. Gerstman, 'The role of consonant-vowel transitions in the perception of the stop and nasal consonants', *Psychological Monographs*, v.68, no.8, pp.1-13, 1954.

(61) R. R. Verbrugge, W. Strange, D. P. Shankweiler and T. R. Edman, 'What information enables a listener to map a talker's vowel space?', *J. of the Acous. Soc. of America*, v.60, no.1, pp.198-212, July 1976.

(62) P. Mermelstein, 'Difference limens for formant frequencies for steady-state and consonant-bound vowels', *J. of the Acous. Soc. of America*, v.63, no.2, pp.572-580, Feb. 1978.

(63) J. L. Flanagan, 'Difference limen for vowel formant frequency', *J. of the Acous. Soc. of America*, v.27, pp.613-617, 1955.

(64) B. Lindblom and A. Floren, 'Estimating short-term context-dependance of formant pattern perception: stimulus specification', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-2, pp.24-26, (Royal Inst. of Tech., Stockholm) 1965.

(65) G. Fairbanks and M. S. Miron, 'Effects of vocal effort upon the consonant-vowel ratio within the syllable', *J. of the Acous. Soc. of America*, v.29, pp.621-626, 1957.

(66) J. C. R. Licklider and I. Pollack, 'Effects of differentiation, integration and infinite peak-clipping upon the intelligibility of speech', *J. of the Acous. Soc. of America*, v.20, no.1, pp.42-51, 1948.

(67) G. Fairbanks, 'A physiological correlate of vowel intensity', *Speech Monograms*, v.17, pp.390-395, 1950.

(68) G. Fant, J. Lindqvist, B. Sonesson and H. Hollien, 'Speech distortion at high pressure', *Underwater Physiology, Proc. of the 4th Int. Congress on Diver Physiology*, pp.293-299, 1971.

(69) G. Fant and B. Sonesson, 'Speech at high ambient air pressure', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.9-21, (Royal Inst. of Tech., Stockholm) 1964.

(70) G. Fant and B. Sonesson, 'Divers' speech in comressed air atmosphere', *Military Medicine*, v.132, pp.434-436, 1967.

(71) H. Hollien, C. L. Thompson and B. Cannon, 'Speech intelligibility as a function of ambient pressure in a helium-oxygen atmosphere', *Aerospace Medicine*, v.44, pp.249-253, March 1973.

(72) Jw. van den Berg, 'Transmission of the vocal cavities', *J. of the Acous. Soc. of America*, v.27, pp.161-168, 1955.

(73) G. Fant, 'Vocal tract wall effects, losses and resonance bandwidths', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-2-3, pp.28-52, (Royal Inst. of Tech., Stockholm) 1972.

(74) G. Fant and J. Lindqvist, 'Pressure and gas mixture effects on divers' speech', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.1-17, (Royal Inst. of Tech., Stockholm) 1968.

(75) H. Suzuki, G. Ooyama and K. Kido, 'Analysis-conversion-synthesis system for improving naturalness and intelligibility of speech at high pressure in a helium gas mixture', *Publication SCS-74, Speech Communication Seminar, Stockholm*, pp.97-105, Aug.1974.

(76) R. L. Sergeant, 'Phonemic analysis of consonants in helium speech', *J. of the Acous. Soc. of America*, v.41, no.1, pp.66-69, 1967.

(77) R. G. Beil, 'Frequency analysis of vowels produced in a helium-rich atmosphere', *J. of the Acous. Soc. of America*, v.34, no.3, pp.347-349, March 1962.

(78) M. Copel, 'Helium voice unscrambling', *IEEE Trans.*, v.AU-14, no.3, pp.122-126, Spet. 1966.

(79) H. Hollien, W. Shearer and J. W. Hicks(Jr), 'Voice fundamental frequency levels of divers in helium-oxygen speaking environments', *Undersea Biomedical Res.*, v.4, no.2, pp.199-207, June 1977.

(80) C. T. Morrow, 'Speech in deep submergence atmospheres', *J. of the Acous. Soc. of America*, v.50, no.1, pp.715-728, 1971.

(81) M. Nakatsui, J. Suzuki, T. Tagasugi and R. Tanaka, 'Nature of helium speech and its unscrambling', *Conference Records of the IEEE Conference on Engineering in the Ocean Environment*, pp.137-140, July 1973.

(82) M. Nakatsui, 'Comment on helium speech - insight into speech event needed', *IEEE Trans.*, v.ASSP-22, pp.472-473, 1974.

(83) R. Tanaka, M. Nakatsui, T. Tagasugi and J. Suzuki, 'Source characteristics of speech produced under high ambient pressure', *J. of the Radio Res. Laboratories of Japan*, v.21, pp.269-373, 1974.

(84) R. Tanaka, M. Nakatsui and J. Suzuki, 'Formant frequency shifts under high ambient pressures', *J. of the Radio Res. Laboratories of Japan*, v.21, pp.261-267, 1974.

(85) M. Nakatsui and J. Suzuki, 'Observations of speech parameters and their daily variations in a helium-nitrogen-oxygen mixture at a depth of 30m', *J. of the Radio Res. Laboratories of Japan*, v.18, pp.221-225, May 1971.

(86) L. J. Gerstman, G. R. Gamertsfelder and A. Goldberger, 'Breathing mixture and depth as separate effects on helium speech', *J. of the Acous. Soc. of America*, v.40, no.5, p.1283(Abstract) 1966.

(87) K. Holywell and G. Harvey, 'Helium speech', *J. of the Acous. Soc. of America*, v.36, pp.210-211, 1964.

(88) R. L. Sergeant, 'Speech during respiration of a mixture of helium and oxygen', *Aerospace Medicine*, v.34, pp.826-828, 1963.

(89) C. W. Nixon and H. C. Sommer, 'Subjective analysis of speech in a helium environment', *Aerospace Medicine*, v.39, pp.139-144, Feb.1968.

(90) T. A. Giordano, H. B. Rothman and H. Hollien, 'Helium speech unscramblers - a critical review of the state of the art', *IEEE Trans.*, v.AU-21, no.5, pp.436-444, Oct. 1973.

(91) E. O. Belcher and S. Hatlestad, 'Analysis of isolated vowels in helium speech', *Report 26-82, Norsk Undervannsteknologisk Senter*, March 1982.

(92) M. A. Richards, 'Helium speech enhancement using the short-time Fourier transform', *IEEE Trans.*, v.ASSP-30, no.6, pp.841-853, Dec. 1982.

(93) R. A. Flower and L. J. Gerstman, 'Correction of helium speech distortions by real-time electronic processing', *IEEE Trans.*, v.COM-19, no.3, pp.362-364, June 1971.

(94) D. J. Maclean, 'Analysis of speech in a helium-oxygen mixture under pressure', *J. of the Acous. Soc. of America*, v.40, no.3, pp.625-627, 1966.

(95) T. Tagasugi, M. Nakatsui and J. Suzuki, 'Long-term speech spectrum in a helium-nitrogen-oxygen mixture at a depth of 30m', *J. of the Radio Res. Laboratories of Japan*, v.18, pp.227-231, May, 1971.

(96) G. Fant, 'Glottal source excitation analysis', *Quarterly Progress and Status Report, Speech Transmission Laboratory*, v.STL-QPSR-1, pp.85-107, (Royal Inst. of Tech., Stockholm) 1979.

(97) J. D. Speakman, 'Physical analysis of speech in helium environments', *Aerospace Medicine*, v.39, pp.48-53, Jan. 1968.

(98) J. Suzuki and M. Nakatsui, 'Articulation of monosyllables uttered in a helium-nitrogen-oxygen mixture at a depth of 30m', *J. of the Radio Res. Laboratories of Japan*, v.18, pp.233-237, May 1971.

(99) R. S. Brubaker and J. W. Wurst, 'Spectrographic analysis of divers' speech during decompression', *J. of the Acous. Soc. of America*, v.43, no.4, pp.798-802, 1968.

(100) E. F. Fluur and J. Adolphson, 'Hearing in hyperbaric air', *Aerospace Medicine*, v.37, pp.783-785, Aug.1966.

(101) G. Fairbanks, 'A theory of the speech mechanism as a servosystem', *J. of Speech and Hearing Disorders*, v.19, pp.133-139, 1954.

(102) M. Verzeano, 'Time-patterns of speech in normal subjects', *J. of Speech and Hearing Disorders*, v.15, pp.197-201, 1950.

(103) B. S. Lee, 'Effects of delayed speech feedback', *J. of the Acous. Soc. of America*, v.22, pp.824-826, 1950.

(104) G. Fairbanks, 'Selective vocal effects of delayed auditory feedback', *J. of Speech and Hearing Disorders*, v.20, pp.333-346, 1955.

(105) J. Suzuki and M. Nakatsui, 'Perception of helium speech uttered under high ambient pressures', *Publication SCS-74, Speech Communication Seminar, Stockholm*, pp.97-105, Aug. 1974.

(106) W. R. Stover, 'Technique for correcting helium speech distortion', *J. of the Acous. Soc. of America*, v.41, pp.70-74, 1967.

(107) J. S. Gill et al., British patent no.1321313, June 1970.

(108) M. A. Jack, A. D. Milne, L. E. Virr and R. Hicks, 'Miniature helium speech unscrambler for diver-borne use', *Conference on Electronics for Ocean Technology*, v.51, pp.13-18, (IERE Conference Proceedings) 1981.

(109) R. M. Golden, 'Improving naturalness and intelligibility of helium speech, using vocoder techniques', *J. of the Acous. Soc. of America*, v.40, no.3, pp.621-624, May 1966.

(110) J. F. Zurcher, 'Voice transcoder', British patent no.1561918, March 1980.

(111) M. A. Jack and G. Duncan, 'The helium speech effect and electronic techniques for enhancing intelligibility in a helium-oxygen environment', *The Radio and Electronic Engineer*, v.52, no.5, pp.211-223, May 1982.

(112) R. F. Quick, 'Helium speech translation using homomorphic techniques', *J. of the Acous. Soc. of America*, v.48, p.130(Abstract), 1970.

(113) H. Hollien and H. B. Rothman, 'Evaluation of helium speech unscramblers under controlled conditions', *MTS Journal*, v.8 no.9, pp.35-44, Oct. 1974.

(114) D. D. Brown and S. H. Feinstein, 'An evaluation of three helium speech unscramblers to a depth of 1000ft', *J. of Sound and Vibration*, v.48, no.1, pp.123-135, 1976.

(115) H. Hollien and H. B. Rothman, 'The effectiveness of divers' work with and without the aid of communication systems', *MTS Journal*, v.9, no.8, pp.3-10, Sept. 1975.

(116) A. M. Noll, 'Cepstrum pitch determination', *J. of the Acous. Soc. of America*, v.41, no.2, pp.293-309, 1967.

(117) H. Akaike, 'Power spectrum estimation through autoregressive model fitting', *Annals of the Institute of Statistical Mathematics*, v.21, pp.407-419, 1969.

(118) S. M. Kay and S. L. Marple (Jr), 'Spectrum analysis - a modern perspective', *Proc. IEEE*, v.69, no.11, pp.1380-1418, Nov. 1981.

(119) J. Makhoul, 'Linear prediction - a tutorial review', *Proc. IEEE*, v.63, pp.561-581, Apr. 1975.

(120) L. J. Lehiste and G. E. Peterson, 'Acoustic intelligibility', *J. of the Acous. Soc. of America*, v.31, pp.280-286, 1959.

(121) G. Fairbanks, 'Voice and articulation drill book (2nd edition)', (Harper Brothers, New York, 1940).

(122) B. Gold and L. R. Rabiner, 'Parallel processing techniques for estimating pitch periods of speech in the time domain', *J. of the Acous. Soc. of America*, v.46, no.2, pp.442-448, Aug. 1969.

(123) L. R. Rabiner, M. R. Sambur and C. E. Schmidt, 'Applications of a non-linear smoothing algorithm to speech processing', *IEEE Trans.*, v.ASSP-23, no.6, pp.552-557, Dec. 1975.

(124) R. W. Schafer and L. R. Rabiner, 'System for automatic formant analysis of voiced speech', *J. of the Acous. Soc. of America*, v.47, pp.634-648, Feb. 1970.

(125) J. D. Markel and A. H. Gray (Jr), 'Linear prediction of speech', (Springer-Verlag, Berlin, 1967).

(126) L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, 'A comparative performance study of several pitch detection algorithms', *IEEE Trans.*, v.ASSP-24, no.5, pp.399-417, Oct. 1976.

(127) 'Interactive Laboratory System (ILS) – Users' Guide', (Signal Technology Inc., Santa Barbara, 1981).

(128) K. R. Scherer, 'Vocal indicators of stress', in 'The evaluation of speech in psychiatry', ed. J. Darby (Grune and Stratton, New York, 1981).

(129) T. J. Ulrych and M. Ooe, 'Autoregressive and mixed autoregressive-moving average models and spectra', in 'Topics in applied physics', ed. S. Haykin (Springer-Verlag, Berlin, 1979).

(130) B. S. Atal and S. L. Hanauer, 'Speech analysis and synthesis by linear prediction of the speech wave', *J. of the Acous. Soc. of America*, v.50, no.2, pp.637-655, 1971.

(131) L. J. Griffiths, 'Rapid measurement of digital instantaneous frequency', *IEEE Trans.*, v.ASSP-23, no.2, pp.207-222, 1975.

(132) R. T. Lacoss, 'Data adaptive spectral analysis methods', *Geophysics*, v.36, pp.661-675, 1971.

(133) P. M. Chirlian, 'Signals, systems and the computer', (Intext Educational Publishers, New York, 1973).

(134) P. J. Gawthorpe, 'Linear prediction and parameter estimation', in 'Digital signal processing', ed. N. B. Jones (Peter Peregrinus Ltd, 1982).

(135) G. U. Yule, 'On a method of investigating periodicities in disturbed series, with particular reference to Wolfer's sunspot numbers', *Philosophical Trans. of the Royal Soc. of London*, v.226, pp.267-298, 1927.

(136) G. Walker, 'On periodicity in series of related terms', *Proc. of the Royal Soc. of London*, v.131, pp.518-532, 1931.

(137) N. Levinson, 'The Weiner r.m.s.-error criterion in filter design and prediction', *J. of Mathematical Physics*, v.25, no.4, pp.261-278, 1947.

(138) N. Weiner, 'Generalized harmonic analysis', *Acta Mathematica*, v.55, pp.117-258, 1930.

(139) H. Akaike, 'Fitting autoregressive models for prediction', *Annals of the Institute of Statistical Mathematics*, v.21, pp.243-247, 1969.

(140) E. Parzen, 'Some recent advances in time series modelling', *IEEE Trans.*, v.AC-19, no.6, pp.723-730, 1974.

(141) H. Akaike, 'A new look at the statistical model identification', *IEEE Trans.*, v.AC-19, no.6, pp.716-723, 1974.

(142) T. E. Landers and R. T. Lacoss, 'Some geophysical applications of autoregressive spectral estimates', *IEEE Trans.*, v.GE-15, no.1, pp.26-33, 1977.

(143) M. J. Rutter, 'The theory, design and application of gradient adaptive lattice filters', *Ph.D. Thesis* (University of Edinburgh, Dept. of Electrical Engineering, 1983).

(144) E. O. Brigham, 'The fast Fourier transform', (Prentice-Hall, Eaglewood Cliffs, 1974).

(145) D. R. Dickson, 'An acoustic study of nasality', *J. of Speech and Hearing Res.*, v.5, pp.103-111, 1962.

(146) H. K. Dunn, 'Methods of measuring vowel formant bandwidths', *J. of the Acous. Soc. of America*, v.33, pp.1737-1746, Dec. 1961.

(147) E. N. Pinson, 'Pitch-synchronous time-domain estimation of formant frequencies and bandwidths', *J. of the Acous. Soc. of America*, v.35, pp.1264-1273, Aug. 1963.

(148) H. Helfrich, 'Age markers in speech' in 'Social markers in speech', ed. K. Scherer and H. Giles (Cambridge University Press, 1979).

(149) S. Hattori, K. Yamamoto and O. Fujimura, 'Nasalization of vowels in relation to nasals', *J. of the Acous. Soc. of America*, v.30, pp.267-274, 1958.

(150) B. L. Hicks, L. D. Braida and N. I. Durlach, 'Pitch invariant frequency lowering with nonuniform spectral compression', *Proc. of the IEEE Int. Conference on ASSP*, pp.121-124, Boston, 1981.

(151) L. A. O'Neill, 'The representation of continuous speech with a periodically sampled orthogonal basis', *IEEE Trans.*, v.AU-17, pp.14-21, March 1969.

(152) M. R. Portnoff, 'Short-time Fourier analysis of sampled speech', *IEEE Trans.*, v.ASSP-29, no.3, pp.364-373, June 1981.

(153) M. R. Portnoff, 'Time-frequency representation of digital signals and systems based on short-time Fourier analysis', *IEEE Trans.*, v.ASSP-28, no.1, pp.55-69, Feb. 1980.

(154) J. B. Allen, 'Applications of the short-time Fourier transform to speech processing and spectral analysis', *Proc. of the IEEE Int. Conference on ASSP*, pp.1012-1015, Paris, 1982.

(155) J. B. Allen and L. R. Rabiner, 'A unified approach to short-time Fourier analysis and synthesis', *Proc. of the IEEE*, v.65, pp.1558-1564, Nov. 1977.

(156) J. B. Allen, 'Short term spectral analysis, synthesis and modification by discrete Fourier transform', *IEEE Trans.*, v.ASSP-25, pp.235-238, June 1977.

(157) C. A. McGonegal, L. R. Rabiner and A. E. Rosenberg, 'A subjective evaluation of pitch detection methods using LPC synthesised speech', *IEEE Trans.*, v.ASSP-25, pp.221-229, 1977.

(158) M. A. Jack, A. D. Milne and L. E. Virr, 'Compact helium speech unscrambler using charge transfer devices', *Electronics Letters*, v.15, no.18, pp.548-550, Aug. 1979.

(159) M. A. Jack, A. D. Milne and W. Donaldson, 'Final report and operating manual for compact helium speech unscrambler equipment', *Wolfson Microelectronics Institute*, (University of Edinburgh) 1979.

(160) B. W. Kernighan and D. M. Ritchie, 'The C Programming language' (Prentice-Hall, 1978).

(161) S. R. Bourne, 'The Unix system' (Addison-Wesley, 1982).

(162) F. F. Kuo, 'Network analysis and synthesis (2nd edition)' (J. Wiley and Sons Inc., New York, 1966).

(163) J. J. D'Azzo and C. H. Houpis, 'Feedback control systems analysis and synthesis' (McGraw-Hill, 1966).

(164) G. M. Jenkins and D. G. Watts, 'Spectral analysis and its applications' (Holden-Day, San Francisco, 1968).

(165) A. V. Oppenheim, A. S. Willsky and I. T. Young, 'Signals and systems' (Prentice-Hall, New Jersey, 1983).

(166) F. J. Harris, 'On the use of windows for harmonic analysis with the discrete Fourier transform', *Proc. of the IEEE*, v.66, no.1, pp.51-83, Jan. 1978.

(167) J. D. Bruce, 'Discrete Fourier transforms, linear filters and spectrum weighting', *IEEE Trans.*, v.AU-16, pp.495-499, Dec. 1968.

(168) C. K. Yuen and D. Fraser, 'Digital spectral analysis' (Pitman, London, 1979).

(169) M. S. Bartlett, 'An introduction to stochastic processes with special reference to methods and applications' (Cambridge University Press, 1953).

(170) A. Papoulis, 'The Fourier integral and its applications' (McGraw-Hill, New York, 1962).

(171) T. J. Healy, 'Convolution revisited', *IEEE Spectrum*, v.6, no.4, pp.87-93, April 1969.

(172) E. O. Belcher, 'The cause and enhancement of helium speech', *Publication of the Norwegian Underwater Technology Centre*, 1982.

(173) C. M. Harris and M. R. Weiss, 'Pitch extraction by computer processing of high-resolution Fourier analysis data', *J. of the Acous. Soc. of America*, v.35, p.339, 1963.

(174) J. Makhoul, 'Methods for nonlinear spectral distortion of speech signals', *Proc. of the IEEE Int. Conference on ASSP*, pp.87-90, April 1976.

(175) H. Nyquist, 'Certain topics in telegraph transmission theory', *Trans. of the American Inst. of Electrical Engineers*, v.47, pp.617-644, April 1928.

(176) G. E. Cook and M. Bernfeld, 'Radar signals: an introduction to theory and application' (Academic Press, New York, 1967).

(177) J. W. Cooley, P. A. W. Lewis and P. D. Welch, 'The finite Fourier transform', *IEEE Trans.*, v.AU-17, no.2, pp.77-85, June 1969.

(178) C. Bingham, M. D. Godfrey and J. W. Tukey, 'Modern techniques of power spectrum estimation', *IEEE Trans.*, v.AU-15, no.2, pp.56-66, June 1967.

(179) G. D. Bergland, 'A guided tour of the fast Fourier transform', *IEEE Spectrum*, pp.41-52, July 1969.

(180) B. Arambepola, 'General discrete Fourier transform and fast Fourier transform algorithm', in 'Signal processing: theories and applications', ed. M. K. Kunt and F. de Coulon (North Holland Publishing Company, 1980).

(181) E. O. Brigham and R. E. Morrow, 'The fast Fourier transform', *IEEE Spectrum*, pp.63-70, Dec. 1967.

(182) R. E. Crochiere, 'A weighted overlap-add method of short-time Fourier analysis/synthesis', *IEEE Trans.*, v.ASSP-28, pp.99-102, Feb. 1980.

(183) W. Gersch and D. R. Sharpe, 'Estimation of power spectra with finite-order autoregressive models', *IEEE Trans.*, v.AC-18, pp.367-369, Aug. 1973.

(184) R. C. Mathes and R. L. Miller, 'Phase effects in monaural perception', *J. of the Acous. Soc. of America*, v.19, p.780, 1947.

(185) G. A. Miller and J. C. R. Licklider, 'The intelligibility of interrupted speech', *J. of the Acous. Soc. of America*, v.22, no.2, pp.167-173, 1950.

(186) E. de Boer, 'A note on phase distortion in hearing', *Acustica*, v.11, pp.182-184, 1961.

(187) J. F. Schouten, R. J. Ritsma and B. L. Cardozo, 'Pitch of the residue', *J. of the Acous. Soc. of America*, v.34, no.8, pp.1418-1424, 1962.

(188) T. J. F. Buunen, J. M. Festen, F. A. Bilsen and G. van den Brink, 'Phase effects in a three-component signal', *J. of the Acous. Soc. of America*, v.55, no.2, 297-303, 1974.

(189) R. J. Ritsma and F. L. Engel, 'Pitch of frequency modulated signals', *J. of the Acous. Soc. of America*, v.36, no.9, pp.1637-1644, 1964.

(190) J. L. Goldstein, 'Auditory nonlinearity', *J. of the Acous. Soc. of America*, v.41, no.3, pp.676-689, 1967.

(191) M. H. Hayes, J. S. Lim and A. V. Oppenheim, 'Signal reconstruction from phase or magnitude', *IEEE Trans.*, v.ASSP-28, no.6, Dec. 1980.

(192) B. Yegnanarayana and A. Dhayalan, 'Noniterative techniques for minimum phase signal reconstruction from phase or magnitude', *Proc. of the IEEE Int. Conference on ASSP*, pp.639-642, Boston, 1983.

(193) S. H. Nawab, T. F. Quatieri and J. S. Lim, 'Algorithms for signal reconstruction from short-time Fourier transform magnitude', *Proc. of the IEEE Int. Conference on ASSP*, pp.800-803, Boston, 1983.

(194) A. V. Oppenheim and J. S. Lim, 'The importance of phase in signals', *Proc. of the IEEE*, v.69, no.5, pp.529-541, May 1981.

(195) E. O. Belcher and K. Andersen, 'Helium speech enhancement by frequency-domain processing', *Proc. of the IEEE Int. Conference on ASSP*, v.3, pp.1160-1163, Boston, 1983

(196) G. Duncan and M. A. Jack, 'Residually excited LPC processor for enhancing helium speech intelligibility', *Electronics Letters*, v.19, no.18, pp.710-711, Sept. 1983.

(197) T. J. Ulrych and T. N. Bishop, 'Maximum entropy spectral analysis and autoregressive decomposition', *Reviews of Geophysics and Space Physics*, v.13, no.1, pp.183-200, Feb. 1975.

(198) C. K. Un and D. T. Magill, 'The residually-excited linear prediction vocoder with transmission rate below 9.6kbit/s', *IEEE Trans.*, v.COM-23, no.12, pp.1466-1473, Dec. 1975.

(199) J. Makhoul, 'Spectral linear prediction: properties and applications', *IEEE Trans.*, v.23, no.3, pp.283-296, June 1975.

(200) S. Beet and C. C. Goodyear, 'Making helium speech intelligible', *IEE Colloquium on the Digital Processing of Speech*, pp.11/1-11/5, April 1983.

(201) T. J. Terrell, 'Introduction to digital filters', (The MacMillan Press Ltd., London, 1980).

(202) C. E. Shannon, 'A mathematical theory of communication', *Bell System Technical Journal*, v.27, no.3, pp.379-423, July 1948.

(203) C. E. Shannon, 'A mathematical theory of communication', *Bell System Technical Journal*, v.27, no.4, pp.623-656, Aug. 1948.

(204) A. Papoulis, 'Maximum entropy and spectral estimation: a review', *IEEE Trans.*, v.ASSP-29, no.6, pp.1176-1186, Dec. 1981.

(205) G. D. Bergland, 'A fast Fourier transform algorithm for real-valued series', *Communications of the Assoc. for Computing Machinery*, v.11, no.10, pp.703-710, Oct. 1968.

(206) H. J. Nussbaumer, 'Fast Fourier transform and convolution algorithms', (Springer-Verlag, New York, 1981).

(207) J. W. Cooley, P. A. Lewis and P. D. Welch, 'Applications of the fast Fourier transform to computation of Fourier integrals, Fourier series, and convolution integrals', *IEEE Trans.*, v.AU-15, no.2, pp.79-84, June 1967.

(208) C. M. Rader, 'An improved algorithm for high speed autocorrelation with applications to spectral estimation', *IEEE Trans.*, v.AU-18, no.4, pp.439-441, Dec. 1970.

(209) A. P. Varga and F. Fallside, 'Multi-pulse excitation in linear predictive synthesis of speech', *IEE Colloquium on the Digital Processing of Speech*, pp.2/1-2/4, April 1983.

(210) M. R. Sambur, A. E. Rosenberg, L. R. Rabiner and C. A. McGonegal, 'On reducing the buzz in LPC synthesis', *J. of the Acous. Soc. of America*, v.63, no.3, pp.918-924, March 1978.

(211) D. Y. Wong, C. C. Hsiao and J. D. Markel, 'Spectral mismatch due to preemphasis in LPC analysis/synthesis', *IEEE Trans.*, v.ASSP-28, no.2, pp.263-264, April 1980.

(212) S. K. Kawahara, R. P. O'Connell and J. G. Peterson, 'A one-micron bipolar VLSI convolver', *Proc. of the IEEE Int. Conference on Solid-State Circuits*, pp.226-227, Feb. 1981.

(213) I. R. MacTaggart and M. A. Jack, 'Radix-2 FFT butterfly processor using distributed arithmetic', *Electronics Letters*, v.19, no.2, pp.43-44, Jan. 1983.

(214) 'Signal processing with the TMS320', (Texas Instruments Inc., 1983).

(215) C. T. Morrow and A. J. Brouns, 'Speech communications in diving masks. I: Acoustics of microphones and mask cavities', *J. of the Acous. Soc. of America*, v.50, no.1, pp.1-9, 1971.

(216) C. T. Morrow and A. J. Brouns, 'Speech communications in diving masks. II: Communication in mask cavities', *J. of the Acous. Soc. of America*, v.50, no.1, pp.10-22, 1971.

(217) B. Widrow, J. R. Glover, J. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong Jr. and R. C. Goodlin, 'Adaptive noise cancelling: principles and applications', *Proc. of the IEEE*, v.63, no.12, pp.1692-1716, Dec. 1979.

(218) C. F. N. Cowan, J. Mavor, J. W. Arthur and P. B. Denyer, 'An evaluation of analogue and digital adaptive filter realisations', *IEE Int. Specialist Seminar on Case Studies in Advanced Signal Processing*, pp.178-183, Sept. 1979.

(219) P. B. Bennett, G. D. Blenkarn, J. Roby and D. Youngblood, 'Suppression of the high pressure nervous syndrome in human deep dives by $He-N_2-O_2$', *Undersea Biomedical Res.*, v.1, pp.221-237, 1974.

Appendix A

Publications

M. A. Jack and G. Duncan, 'The helium speech effect and electronic techniques for enhancing intelligibility in a helium-oxygen environment', *The Radio and Electronic Engineer*, v.52, no.5, pp.211-223, May 1982.

- and -

G. Duncan and M. A. Jack, 'Residually excited LPC processor for enhancing helium speech intelligibility', *Electronics Letters*, v.19, no.18, pp.710-711, Sept. 1983.

UDC 534.78 : 621.37/8 : 626.02
Indexing Terms: Diving apparatus, Helium–oxygen systems, Signal processing, Speech analysis and synthesis

The helium speech effect and electronic techniques for enhancing intelligibility in a helium–oxygen environment

M. A. JACK, B.Sc., M.Sc., Ph.D.*
and
G. DUNCAN, B.Sc.†

SUMMARY

This paper considers the nature of the speech mechanism, and the effects on the speech spectrum of a high pressure helium–air environment. A comparison is made between the characteristics of speech distortions in a helium–air mixture and certain well-known characteristics of speech in normal air which give rise to similar effects.

The criteria for good intelligibility are related to the performance of various helium speech unscrambling techniques which have been used. These unscrambling techniques are classified here into two main categories: those essentially using signal processing in the frequency domain and those using signal processing in the time domain. Consideration is also given to waveform coding techniques which involve a combination of both of these classes of signal processing.

The ability of these categories to incorporate the various features required for good intelligibility in unscrambling helium speech is discussed, in order to highlight the potential importance of frequency domain approaches to future unscrambling developments.

* *Department of Electrical Engineering, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JL.*
† *Wolfson Microelectronics Institute, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JL.*

1 Introduction

Deep sea diving is a specialist industry which has grown dramatically in importance within Europe as a direct result of oil exploration and production in the North Sea. Divers provide an essential service without which underwater construction, inspection and maintenance tasks could not be carried out. At the same time diving is a difficult and hazardous occupation and in consequence a great deal of time and money has been spent to improve operational efficiency and the personal safety of the divers. The present widespread use of saturation diving techniques means that divers can now work in the sea at great depths for long periods of time and working dives to 700 metres are now possible. The frequency, complexity and cost of such dives have focused attention on the need for improved diver-to-diver and diver-to-surface communication.

Saturation diving refers to the technique whereby the diver can be considered as 'saturated' with gas for a specific depth and can remain at such depths without danger. Decompression time is necessary because dissolved gases in the body must be given sufficient time to reach a new equilibrium with the reducing ambient pressure during ascent from a dive. Failure to decompress the diver at a slow enough rate causes the condition known as 'the bends', in which bubbles of trapped gas are released in the tissues. In extreme cases, bubbles can appear in the bloodstream itself with possibly fatal consequences. Nitrogen, which constitutes 78% by volume of air, is an especially troublesome gas in this respect and requires long decompression times, even at shallow depths.

In saturation diving beyond approximately 50 metres, breathing normal air under pressure causes enhanced absorption of nitrogen into the bloodstream, and promotes a dangerous state of narcosis in the diver, reducing his capacity to work accurately and increasing the possibility of a fatal error in judgement.

The use of a gas mixture with reduced nitrogen content obviates this dangerous situation. The simple solution of increasing the proportion of oxygen in the breathing mixture is however to be avoided since oxygen gas dissolved in the bloodstream can cause convulsions if the partial pressure of the oxygen in the mixture exceeds two atmospheres. Further, at depth, the density of normal breathing mixtures would render breathing difficult and therefore use is made of a gas mixture containing high proportions of a gas with low molecular weight. Helium, which has the additional advantage of being an inert gas, is normally used. The use of helium in the respiratory gas mixture (of the order of 95% helium, 3% nitrogen, 2% oxygen), thus reduces the probability of these dangerous physiological and mental problems, but it creates a serious communication problem for those who must attempt to understand the 'squeaky' voice emissions of the diver. It is these distorted voice sounds which are termed 'helium speech'.

The distortions of speech sounds uttered in helium, when compared to similar sounds in air, render the speech unintelligible and the need for electronic techniques to 'unscramble' the helium speech has been identified.

The Radio and Electronic Engineer, Vol. 52, No. 5, pp. 211–223, May 1982
0033-7722/82/050211+13 $1.50/0

This paper considers in detail the nature of the speech mechanism, and the effects on the speech spectrum of pressure and of a helium environment are discussed. A comparison is made between the characteristics of speech utterances in helium and certain well-known characteristics of speech in normal air which give rise to similar effects.

The criteria for good intelligibility both with respect to the listener and with respect to the speaker himself are discussed, and these are related to the fidelity of the various unscrambling techniques which have been used.

A number of helium speech unscrambler techniques have been proposed to date, and these are classified here into two main categories: those essentially using signal processing in the frequency domain, and those using signal processing in the time domain. Consideration is also given to waveform coding techniques which involve a combination of both of these classes of signal processing.

Finally, the factors necessary for good intelligibility from unscrambled helium speech are re-iterated, and related to the performance in terms of intelligibility of present-day unscrambler devices.

2 The Speech Mechanism

In normal speech in air, the structure of the speech waveform is determined by a complex interplay between the shape of the vocal tract from the vocal cords to the lips and nose, by the shape of the excitation waveform injected into the vocal tract, and by the speed of sound in the exhaled gas mixture (Fig. 1). There are three main

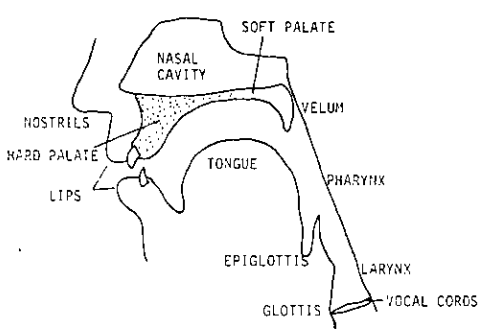


Fig. 1. Section through human vocal apparatus.

ways in which the airflow from the lungs is converted into an acoustic signal with components in the audio range. Firstly, the vocal cords situated in the larynx are adducted, or drawn together, while air is blown from the lungs through the glottis. Very rapidly, sub-glottal pressure builds up as the lungs continue to expel air, until the pressure necessary to blow the vocal cords apart again is reached. The cords separate under this pressure, allowing an impulse of air to be injected from the over-pressured sub-glottis through the gap between the vocal cords. This gap effectively takes the form of a Venturi tube due to the fashion in which the cords separate and as a result of the local drop in air pressure in the constricted passage between the cords, coupled with the elastic tensions acting in and on the vocal cords and the relieved sub-glottal pressure, the vocal cords are forced back towards each other. The instant of glottal closure is usually the point at which excitation of the vocal tract is most powerful.[1–3] This process is repeated at a rate of between 50 and 400 times a second in the voiced speech uttered during ordinary conversation. The spectrum of the vocal cord excitation exhibits a harmonic structure in which the strongest component is normally the fundamental repetition rate and the fall-off with increasing frequency is around 12 dB/octave. Vowel sounds such as 'ah' or 'ee' are produced by this mechanism.

The second main sound source consists of air turbulence produced at a constriction somewhere in the vocal tract. Since it involves a continuous stream of air, the spectrum of this sound source exhibits neither periodicity nor any harmonic structure, and resembles white noise with an essentially flat power spectrum. Sounds formed in this way are generally characterized by a 'hissy' quality, and are termed fricatives. Examples of fricatives include 'ss' or 'ff'.

The third type of sound source results from the build-up of pressure which occurs when the vocal tract is closed at some point, such as by the lips or tongue. A sudden release of pressure causes a transient excitation of the vocal tract which results in a sudden onset of sound. The speech resulting from this type of excitation can be classified into two types: voiceless stop consonants, where the vocal cords are not vibrating during the closure and the onset is preceded by silence (such as 'p' or 't'); and voiced stop consonants, where the vocal cords are vibrating during the closure and the onset is preceded by a low intensity voiced sound, such 'b' or 'd'.

Sound energy from these sources enters the vocal tract, whose configuration, as governed by the positions of the tongue, jaw, velum and lips (Fig. 1), determines its acoustic properties and modifies the spectrum of the sound source. The resonances of the vocal tract cause concentrations of energy at certain frequencies which are known as formants. The coupling at the point of radation (lips/nostrils) produces a high-frequency emphasis of 6 dB/octave.

3 Effects of Ambient Pressure on the Speech Spectrum

Although the velocity of sound is essentially independent of air pressure, when a diver speaks in high pressure air

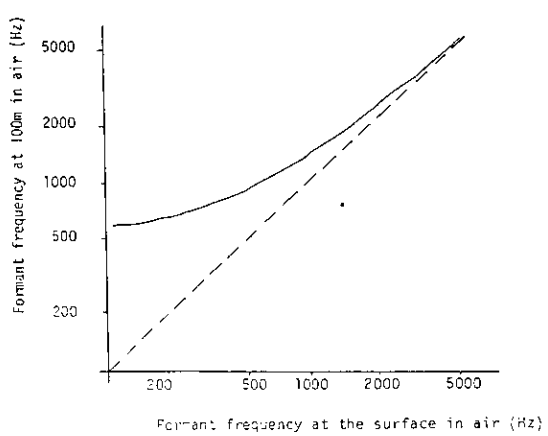


Fig. 2. Formant shift in high pressure air (from Ref. 6).

his speech exhibits a pronounced nasal quality and a steady decrease in speech intelligibility[4,5] is produced with increasing depth and air pressure. Published comparisons[6] of voice formant frequencies in air at atmospheric pressure and in air at high pressure (3 atmospheres) indicate a non-linear shift in formant frequencies (Fig. 2). This effect is considered to be a result of a variation in acoustic impedance mismatch between the vocal tract walls and the high density air.[7,8]

4 Effects of Helium–Oxygen Mixtures on Pitch Period

Published results[9,10] suggest that the fundamental period of repetition of the vocal cords (the pitch (or larynx) period) does not vary as a function of the composition or pressure of the breathing environment.

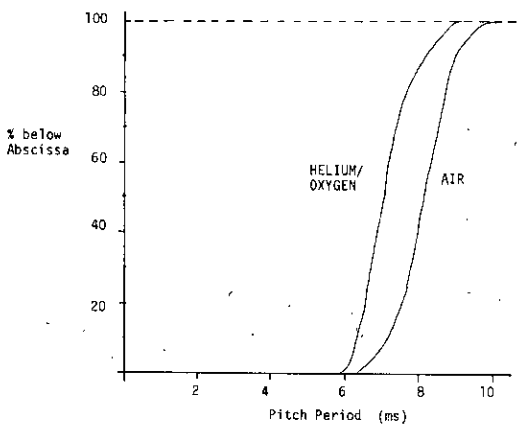


Fig. 3. Cumulative distribution of pitch period (from Ref. 7).

Although early results, shown in Fig. 3, demonstrate a slight decrease in pitch period when breathing helium gas under laboratory conditions, this effect has been attributed to a physical contraction of the larynx muscles since the helium gas was colder than room temperature.[7] Acoustic theory would predict that no increase in pitch period should occur as a function of increased helium concentration or increased ambient pressure. Published data[11] relating the pitch period for the same utterances

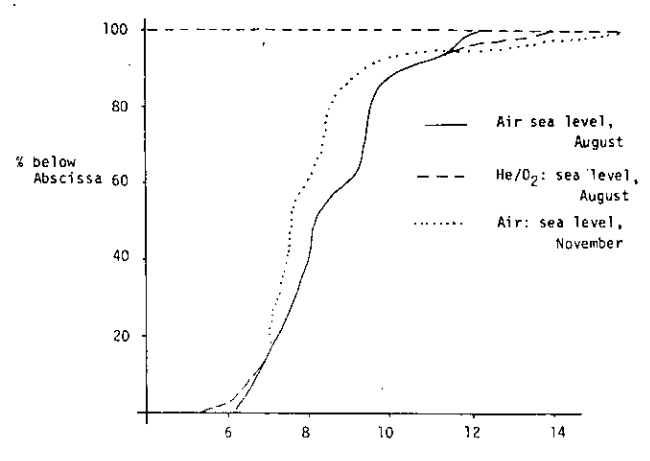


Fig. 4. Cumulative distribution of fundamental periods for air and helium/oxygen mixture at sea level (from Ref. 11).

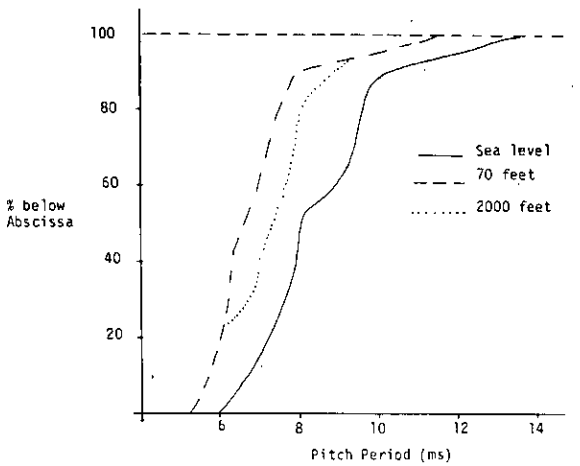


Fig. 5. Cumulative distribution of pitch period for helium/oxygen mixture vs. depth (from Ref. 11).

made in air and in a (pressured) helium environment are shown in Fig. 4 and demonstrate close correlation between the pitch distribution in air and helium at sea level on the same day. However, the pitch distribution measured some three months later for the same utterance in air at sea level shows a noticeable change. The results suggest that the effect of helium gas present in the vocal tract has no distinct effect upon pitch period, but however, the pitch distribution for a given speaker and a given utterance is subject to change with time. Results[11] shown in Fig. 5 demonstrate that pitch period in helium is reduced in comparison to the pitch period in air for the case where the helium is respired at depth. Notice here, however, that the reduction in pitch period does not vary directly with depth since the reduction is greater for 70 feet depth than for 200 feet depth, and that such changes fall within the range of expected pitch variations for normal speech in air. Several possible causes for these observed pitch changes at depth have been suggested.

First, at depth, and especially in a diving habitat, it has been observed that divers tend to speak with increased vocal intensity (loudness) in an attempt to overcome background noise levels experienced in such a chamber. Such increases in vocal intensity are normally accompanied by a reduction in pitch period. Secondly, environmental effects on the diver's speech mechanism, such as changes in the acoustic loading of the vocal tract might be expected to produce some change in pitch period. Finally, the diver may invoke modifications to his speech to alter his pitch period, in an attempt to enhance the intelligibility of speech as he himself judges.[12] Although changes in pitch period have been measured at depth, it has been shown[13] that changes of the magnitudes described in Fig. 5 have minimal effect on speech intelligibility. Pressure and gas composition appear to produce little effect on the spectrum of the vocal cord source.

5 Effects of Helium–Oxygen Mixtures on Voiced Sounds

Helium speech is best understood in relation to voiced vowel sounds. Vowel sounds are produced by excitation of the vocal tract by the vocal cords alone. The vocal

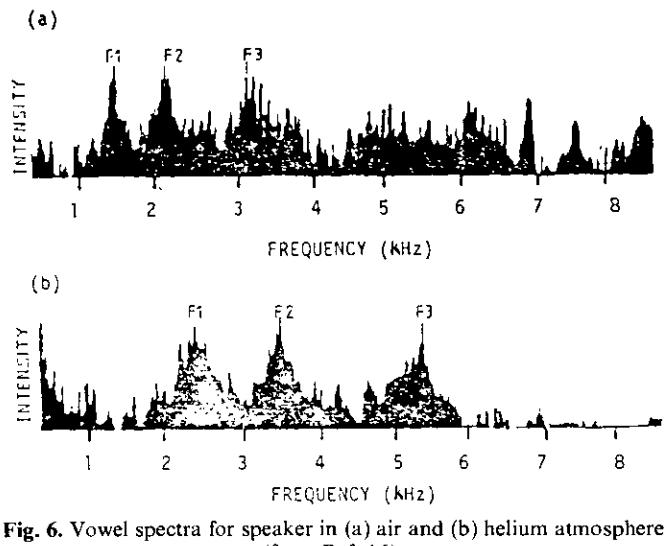


Fig. 6. Vowel spectra for speaker in (a) air and (b) helium atmosphere (from Ref. 16).

tract resonances produce peaks in the spectrum of the speech signal at frequencies determined by the positions of the movable elements (tongue, lips) of the vocal tract and a simple frequency translation in these formant frequencies might therefore be expected as a result of change in the speed of sound produced when breathing a gas of low molecular weight. Such a simple frequency translation effect applies in general; however, more detailed investigations have shown that the formant shift, in particular for the first formant, is non-linear and these non-linearities have been attributed to physiological effects produced by the change in gas density and by the increase in ambient pressure.[6,14,15] Figure 6 shows two speech spectra of the same vowel uttered by a speaker in air, (a), and in a helium environment, (b), (79% helium : 16% oxygen) at normal atmospheric pressure.[16] The frequency transposition ratios for formant centre frequencies F_1, F_2 and F_3 are approximately 1·65 : 1, 1·57 : 1, 1·56 : 1, illustrating the non-uniformity of frequency translation especially for the first formant. An increase in formant bandwidth is also

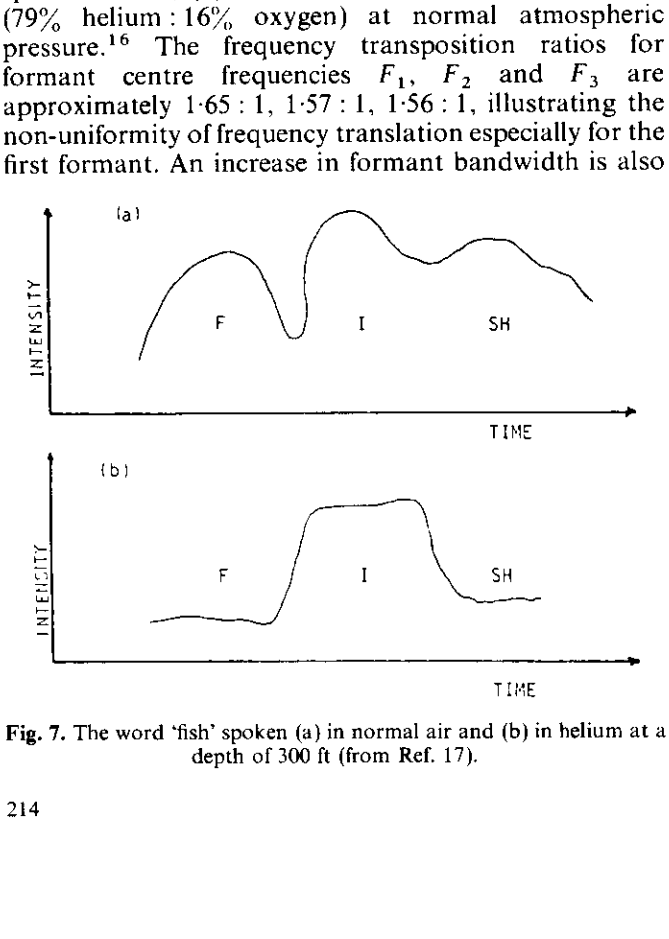


Fig. 7. The word 'fish' spoken (a) in normal air and (b) in helium at a depth of 300 ft (from Ref. 17).

apparent in Fig. 6(b) and, in particular, there is severe attenuation of high frequency components above 6 kHz. Thus, to compensate for the frequency expansions experienced in a helium–oxygen mixture, which correspond to a fall-off in excitation with frequency of 6 dB/octave, high-frequency pre-emphasis is indispensable in unscrambler design.

6 Effects of Helium–Oxygen Mixtures on Unvoiced Sounds

The attenuation of high-frequency components in helium leads to a dominance of low-frequency voiced sounds over high-frequency unvoiced sounds. This has a serious effect on the intelligibility of the speech signal since many consonant sounds in words fall into the latter category.

An example of consonant attenuation in helium speech can be seen in Fig. 7, which compares vocal intensity as a function of time for the word 'fish' taken at the surface, Fig. 7(a), and then at 300 feet, Fig. 7(b), in a helium atmosphere.[17]

7 Nasality Effects in Helium Speech

The loss of high-frequency energy and the non-linear formant frequency shifts present in helium speech can be directly attributed to an increase in the nasality of the helium speech.[14,15]

Nasality is normally due to a resonance of the nasal cavity. However, nasality can also be produced by any 'sidebranch' present in the vocal tract.[1] The main features of nasality in air at normal pressure are a general broadening of formant bandwidths, an overall loss in spectral energy, and a significant drop in intensity of the higher formants. The correlation between these features of nasality in air and the previously discussed parameters of the helium speech effect leads directly to the conclusion that nasal cavity resonance contributes to the main non-linear effects in helium speech.

In a helium environment, the likely path for transmission of energy into the nasal cavity is through the tissue of the soft palate (Fig. 1) whose acoustic impedance becomes more closely matched to that of a helium–oxygen mixture under pressure than to normal air.

8 Intelligibility of Helium Speech to the Listener

One of the most important accepted criteria for speech intelligibility is the ability to distinguish speech sounds by their relative formant frequency ratios. This is the property which allows distinction of the same sound spoken by a small boy, whose absolute formant frequencies are quite high, and a grown man, whose absolute formant frequencies will be much lower. In view of the non-linear formant shifts inherent in a pressured helium atmosphere, vowels and certain consonants involving voiced speech can therefore be expected to become more difficult to identify since their relative formant frequency ratios will not be maintained in the helium atmosphere.

Another important factor degrading intelligibility of helium speech is due to the enhancement of vowel sound intensity relative to the intensity of consonant sounds (see Fig. 7). This feature is important since consonants

and the transitions between consonants and vowels provide cues to the next sound to be produced by a particular speaker.[17,18] Transitions of the first and second formant frequencies have been highlighted[19] as playing an important role in identifying the onset of, for example, nasal sounds. Hence, in the absence of such cues, the listener's perceptual system may be caught unawares and will effectively lag behind changes in the speech, and it may misinterpret and confuse one transition for another, thereby reducing intelligibility. The importance of vowel–consonant transitions has been widely reported and it has been shown that the vowel–consonant intensity ratio is an important vehicle upon which many speakers reduce articulatory effort, using instead the variation of the intensity ratio to impose a message structure.[18]

9 Self-intelligibility in a Helium Atmosphere

The problem of self-intelligibility for the diver himself is a more complex affair. It has been shown that the helium atmosphere itself affects the diver's auditory ability and provides a high-frequency emphasis of +10 dB/octave for those frequencies above 5 kHz and an attenuation for those below 5 kHz.[15] In addition to this inherent compensation in the auditory system some divers appear to adapt their voices (usually over a matter of days) in order to sound more intelligible to themselves, with varying opinions as to the success or otherwise of such manipulations with respect to the listener.[6,14]

on the part of the diver suggest that the speech mechanism can be regarded as a closed-loop feedback system (Fig. 9). Feedback path P1 represents acoustic pathways to the ear through the surrounding environment, and paths P2 and P3 represent tactile feedback, or feedback by sensation through body tissues. The sample-and-hold unit represents short persistence memory, and stores information relating to the next sound which the speaker intends to make. This sampled information passes immediately to the feedback nodes of the error processor which analyses by how much the actual sound output differs from the expected output. It is thought that the decision element works in a predictive mode, projecting ahead to the zero-error condition and permitting input of the next speech unit on this basis.[20]

10 Helium Speech Unscrambling

The earliest attempts to unscramble helium speech involved the use of tape recorders where the helium speech was recorded at a fast speed and was subsequently played-back at a lower speed at which the resulting speech was more intelligible.[21] A modified version of this technique which enabled real-time processing involved use of a continuous tape loop on which the helium speech signal was recorded, and was subsequently read by pick-up heads which were themselves rotating past the tape loop. However, these tape methods inherently involved bulky, moving mechanisms and, more importantly, were limited in their

	Vowel 1			Vowel 2			Vowel 3			Vowel 4		
	F1	F1:F2	F1:F3	F1	F1:F2	F1:F3	F1	F1:F2	F1:F3	F1	F1:F2	F1:F3
Early in experiment	1000	1.9	3.25	1100	1.91	3.27	950	2.3	3.3	1450	1.93	2.97
Late in experiment	900	2	3.28	950	2	3.47	800	2.6	3.56	1100	2.09	3.36
In air	550	3.09	4.45	600	3	4.17	400	3	5.75	700	2	3.36

Fig. 8. Self-adaptation of formant ratios in a helium environment towards normal (air) values.

Formant frequency ratios for four different vowel sounds uttered by the same subject are shown in Fig. 8 and suggest that relative formant ratios can be varied in a conscious manner by the diver in a helium environment.[14] The subject spent several days in a deep diving chamber and the respective vowel formant frequencies were measured in air before the experiment, in the chamber at the start of the dive and, finally, prior to leaving the chamber after several days in the helium environment.

The trends of Fig. 8 show that the first formant frequency F_1 has ultimately tended towards its value in air and that the relative formant frequency ratios ($F_1:F_2$, $F_1:F_3$) have also demonstrated a consistent trend towards their values in air.

Such observations relating to the self-compensation ultimate fidelity since, in addition to shifting the formant frequencies of the speech signal, the fundamental frequency was also shifted.

The search to produce unscrambling methods which would enable correction of only the formant frequencies of the helium speech spectrum led to the development of a series of real-time helium speech unscramblers based on a variety of signal processing algorithms. These signal processing algorithms can be divided into frequency-domain and time-domain techniques. Frequency-domain techniques involve unscrambling of the speech signal by real-time frequency manipulation, usually by some form of band-pass filtering. These techniques offer the intrinsic advantage of preserving basic pitch information which is important to intelligibility. Time-domain techniques involve processing of the time-

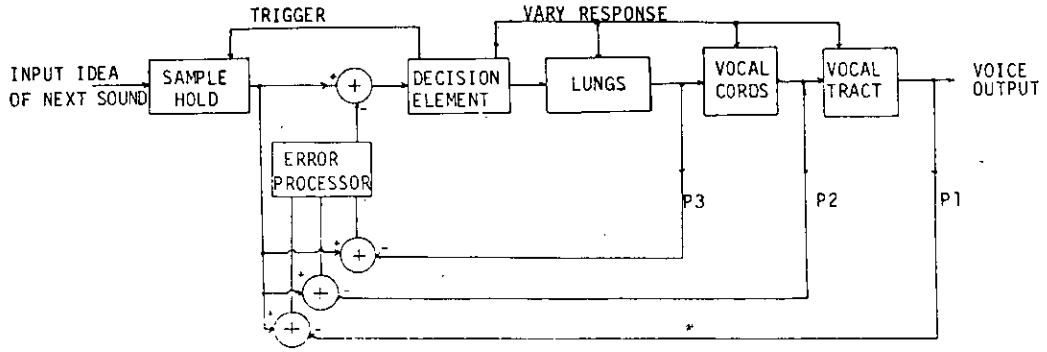


Fig. 9. The speech mechanism as a closed-loop feedback system.

varying helium speech signal directly, the most widely used technique being to segment the helium speech signal either asynchronously or in synchronism with the pitch period, prior to subsequent time expansion of each speech segment by the required correction ratio. Time-domain techniques thus need special precautions to maintain pitch information.

Recent developments in unscrambler design involve the use of unscrambling techniques based on waveform encoding methods, which are a mixture of time and frequency domain manipulations. These encoding techniques, however, invariably involve computer processing, which is currently too slow to allow realization of a real-time unscrambler. These recent developments are reviewed here after a consideration of the three techniques used in unscrambler design, two of which can be regarded as frequency-domain techniques and the third being a time-domain processing technique.

10.1 Unscrambling by Frequency Subtraction

The first of the frequency domain techniques involves frequency subtraction by heterodyning, where the input helium speech is first converted upwards in frequency by a balanced modulator, and is subsequently down-converted by a balanced demodulator of different (and variable) frequency.[11] Use of such a variable-frequency mixing oscillator allows some form of fine tuning and hence caters for variations in helium mixture.[22]

An extension of this principle which incorporates dual-band frequency subtraction (Fig. 10), permits each band to be shifted downwards by different amounts, thereby improving intelligibility. Note that every frequency

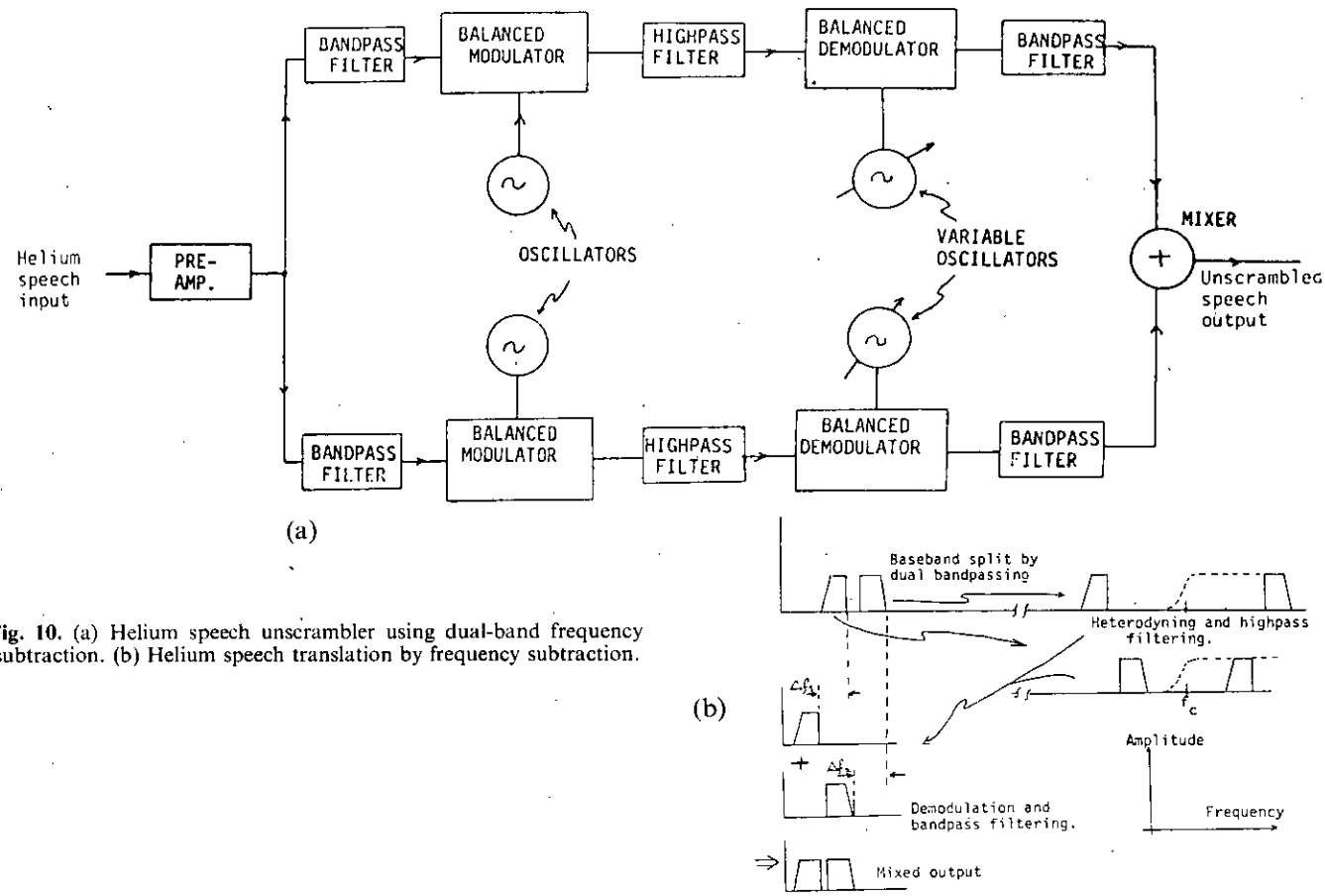


Fig. 10. (a) Helium speech unscrambler using dual-band frequency subtraction. (b) Helium speech translation by frequency subtraction.

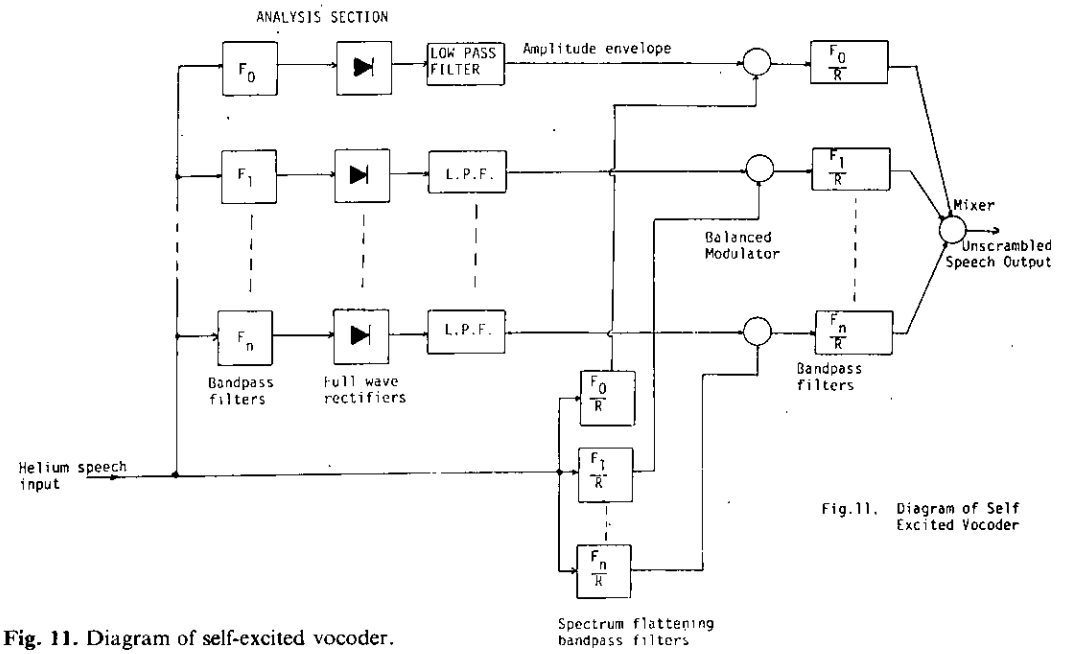


Fig.11. Diagram of Self Excited Vocoder

Fig. 11. Diagram of self-excited vocoder.

within the individual bands will be shifted by the same amount from its original position in the helium speech spectrum, thus the original formant bandwidths are conserved, and since helium speech is, by nature, very nasal in quality with an attendant increase in formant bandwidth, this nasal quality will be conserved in the unscrambled output, thereby limiting the fidelity of this technique.

10.2 Unscrambling by Frequency Multiplication

The second of the frequency domain techniques involves a variant of the analysis-synthesis (vocoding) scheme.[3] This technique offers unscrambling by scaling the frequency spectrum of the helium speech by ratio multiplication as opposed to the heterodyne technique, which used arithmetic subtraction. Use of the ratio multiplication technique means that formant bandwidths can be scaled by a geometric ratio and hence improved performance in terms of reduced nasality of the unscrambled helium speech can be achieved.

In general, in the analysis operation the input speech signal is decomposed into several contiguous passbands and the amplitude envelope (spectral energy) in each passband is extracted. The system employed for this is

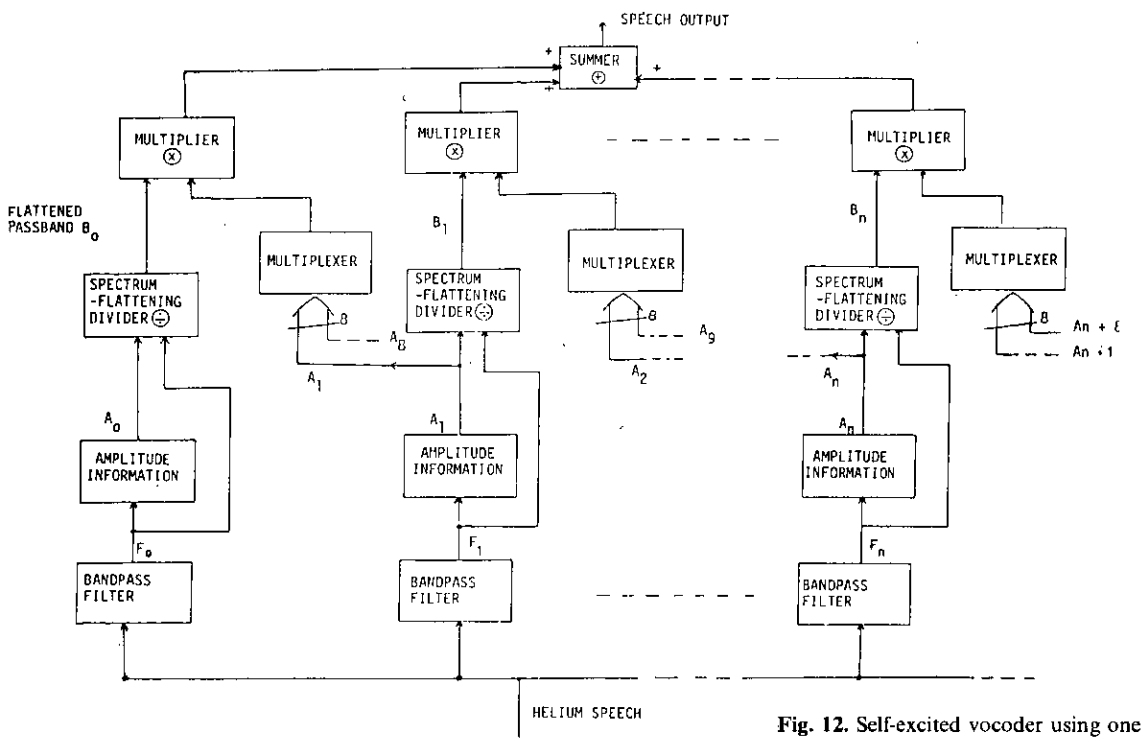


Fig. 12. Self-excited vocoder using one set of bandpass filters.

similar to that shown in the analysis section of Fig. 11. In classical vocoder design, additional circuitry exists to determine if the input speech is voiced or unvoiced and, when voiced, the fundamental pitch period is extracted. The amplitude information from each passband, the pitch information and the voiced/unvoiced decision are subsequently transmitted to a synthesizer, where a 'voiced' excitation source (driven at the pitch period) or an 'unvoiced' random signal source is connected to a second set of bandpass filters. The excitation level for each bandpass filter in the synthesizer is dictated by the energy level of the corresponding analysis filter. An extension to this analysis/synthesis scheme has been applied to the helium speech problem in the form of the Self-Excited Vocoder (SEV) system[23] (see Fig. 11). The main difference between this arrangement and the classical vocoder design outlined above lies in the excitation source used to drive the synthesis filters. Here, instead of using an artificial voiced or unvoiced source, the input helium speech is spectrally flattened (uniform spectral intensity) and is itself used as the excitation source to drive the synthesis filters, thereby preserving the natural quality of the speech. For the purpose of unscrambling helium speech, the centre frequencies of the synthesis bandpass filters and the corresponding analysis bandpass filters are related by the factor R which is defined as the ratio of the velocities of sound in the helium mixture and in air respectively. Thus for an analysis filter having centre frequency F_c, the amplitude information it derives relates to the synthesis filter whose centre frequency is equal to F_c/R. The synthesizer outputs are then summed to produce the formant-shifted speech output.

A recent, further development[24] of the SEV principle uses only one set of contiguous, bandpass filters to derive the signals required for both analysis and synthesis tasks (Fig. 12). The principle of operation depends on there being a mathematical relationship between both adjacent centre frequencies of the contiguous bandpass filters and also the velocity ratio R for certain helium–oxygen mixtures. Here, the permissible velocity ratios which the unscrambler can handle form a geometric series, as does the sequence of bandpass filter centre frequencies, both series being linked by the same geometric ratio. This means that a simple multiplexing arrangement can be employed to combine the appropriate synthesis signal (amplitude envelope) with the corresponding spectrally flattened passband (excitation signal) and hence essentially provide a shifted segment of the original helium speech spectrum.

Both the number of bandpass filters required in the system and the amount of different velocity ratios which can be catered for are directly dependant on the choice of value for the geometric ratio. A convenient value which has been chosen for this system constant[24] is 1·21.

Since the maximum velocity ratio possible is ≈ 3 (100% helium), then the system may cope with six different helium–oxygen mixtures in the range 1·21–3, ($1{\cdot}21^6 \approx 3$). Also, twenty-two contiguous bandpass filters are required with this choice of system constant in order to unscramble helium speech successfully in the range 250 Hz–16·5 kHz. (250 Hz $\times$ $1{\cdot}21^{22} = 16{\cdot}5$ kHz).

The main advantages of this system are that it provides a selectable range of helium–oxygen mixtures over which it will operate and, since it uses only one set of bandpass filters to provide analysis and synthesis signals, the number of electronic components is reduced thereby reducing the cost of the system. The very nature of operation of this system, however, precludes the ability to correct for non-linear formant shifting in the speech spectrum, although the possibility of correcting for high-frequency attenuation exists by selective manipulation of gain within passbands.

10.3 Unscrambling by Time Expansion

The most widely used real-time signal processing technique for helium speech unscrambling involves time expansion of pitch-synchronous segments of the speech signal. The precursor to this technique involved the use of a tape recorder, where playback of the recorded helium speech at half-speed was reported to increase intelligibility.[21] However, with this tape recorder method, fundamental pitch period is also shifted, and the technique is difficult to implement in real time. The time expansion process involves segmenting the helium speech signal in time using the start of each pitch period as the segment marker. The signal within a segment is sampled at a fast rate commensurate with Nyquist requirements; the signal is then stored and subsequently read out of the storage register at a slower speed. There are many available means of determining the start of a pitch period.[25–30]

Several unscramblers have been designed which employ signal processing techniques based on bandwidth compression by simple waveform time-base expansion in pitch synchronism with each pitch period.[31–33] Such signal processing demands use of electronic storage and systems have tended to employ digital memory components for this function. However, developments in analogue memory components now permit the realization of unscramblers which employ the basic pitch-synchronous time-expansion technique but are based on analogue charge transfer devices for waveform storage and c.m.o.s. digital circuitry for control logic functions. Use of low-power analogue technology in the unscrambler permits realization of a compact unscrambler system which can operate in an underwater environment.[34,35]

A schematic block diagram of a helium speech unscrambler based on the time-expansion technique is shown in Fig. 13. After high-frequency equalization in

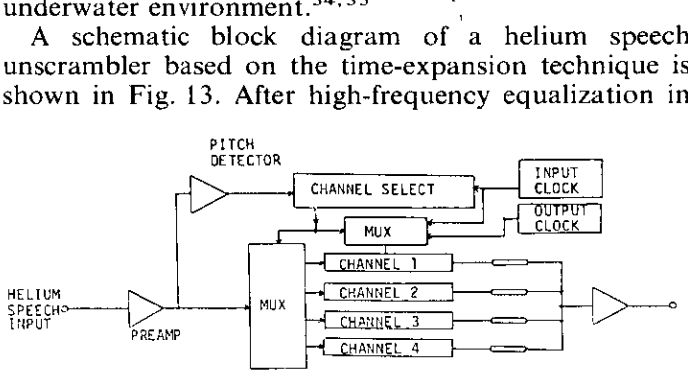


Fig. 13. Diagram of helium speech unscrambler using the pitch synchronous time expansion technique.

the pre-amplifier, the helium speech signal is input to the data stores.

From acoustic theory, the speed of sound in a gas is given by:

$$C = (\gamma P/\rho)^{\frac{1}{2}}$$

where γ is the adiabatic constant (ratio of specific heats), P is the gas pressure and ρ is the density of the gas. The adiabatic constants and the densities can be calculated using the following relationships:

$$\gamma = \sum_i Q_i \cdot \gamma_i$$

$$\rho = \sum_i Q_i \cdot \rho_i$$

where Q_i is the percentage of the ith gas in a mixture. For air the main constituents are nitrogen and oxygen in the proportions 78% : 21% by volume. A measure of the worst-case condition can be made by considering a pure helium environment in which case the velocity ratio is given by:

$$\frac{C_{He}}{C_{Air}} = \left(\frac{\gamma_{He}}{\gamma_{Air}} \times \frac{\rho_{Air}}{\rho_{He}}\right)^{\frac{1}{2}} = \left(\frac{1{\cdot}66 \times 1{\cdot}27}{1{\cdot}40 \times 0{\cdot}179}\right)^{\frac{1}{2}} = 2{\cdot}9.$$

Since the maximum time expansion required on the inter-pitch stored waveform is 3 : 1, four independent storage channels are required so that, in the worst case, three channels can be reading signal out whilst the fourth channel is available for reading signal in, thus preventing loss of any pitch intervals.

The pitch detector used in available unscramblers is typically an amplitude peak detection circuit with hysteresis. The increased speed of sound in the helium/oxygen mixture—in comparison to air—produces a faster decay time-constant for the helium speech waveform envelope, forcing the pitch peaks to become more pronounced than in air, and simplifying the pitch detection circuitry so that a peak detector circuit can be successfully employed. With unvoiced speech, characterized by a noise-like waveform, the pitch detector operates repetitively.

The output from the peak detector circuit defines the start of a pitch period and a segment of the input helium speech, of bandwidth up to 16 kHz, is stored in one of the four channels at a fast sample rate. At the end of the frame, the clock frequency for this store is reduced by a factor which is dependent on the gas mixture being used by the diver. The stored signal is then read out at a lower rate with a bandwidth compression to the normal 3–4 kHz speech bandwidth. On detection of the start of the subsequent pitch period, the clock and signal multiplexers change and another store reads in the helium speech. Typical stored segments are in the range 2·5–20 ms permitting operation with diver fundamental frequencies up to around 400 Hz. This is acceptable for most male voices, whose pitch frequency will on average vary from 70 Hz to 150 Hz in normal speech, and for most female voices which will vary from 100 Hz to 400 Hz.

More than one channel may be producing an output at any one time, a feature which is acceptable since, in normal speech, the vocal tract response due to a glottal pulse has not died away before the next response appears.

A section of input helium speech waveform (85 metres depth) is shown in Fig. 14(a), with Fig. 14(b) indicating the pitch detector output defining a 3 ms storage interval from the pitch peak. The loss of signal in discarded sections of each pitch period causes minimal degradation in the output speech. The pitch period in Fig. 14(a) is seen to be of the order of 6·5 ms. Another section of input helium speech waveform (Fig. 15(a), 85 metres depth) is compared with the corresponding unscrambled output (Fig. 15(b)). Note that the pitch interval (approx. 6 ms) remains unchanged although the stored signal is expanded in time.

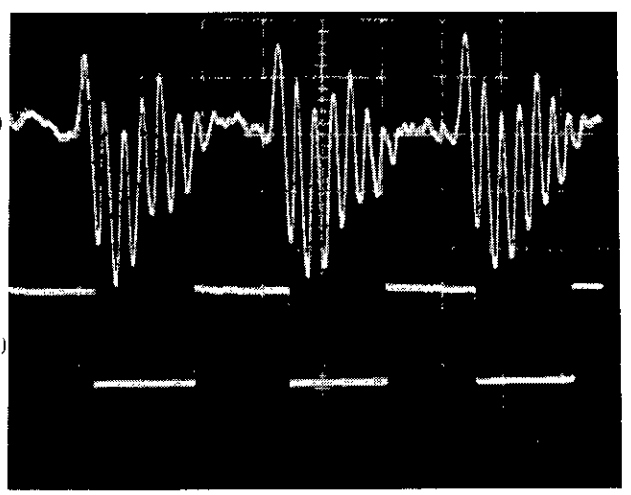


Fig. 14. Pitch detector operation showing 3 ms frame. (a) Helium speech waveform 20 mV/div. (b) Pitch detector output 10 V/div horizontal scale 1 ms/div.

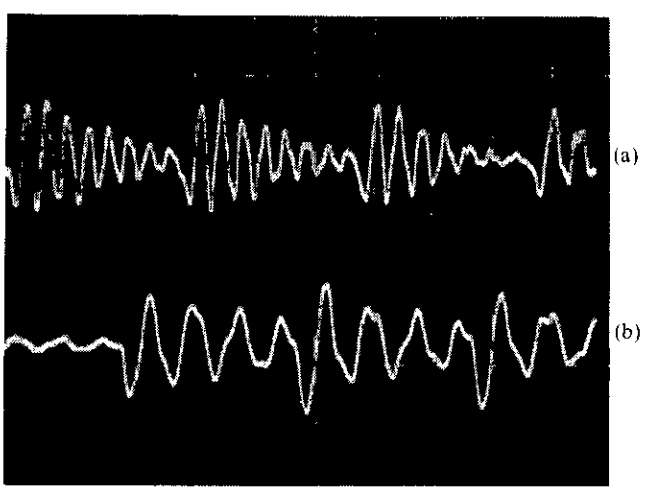


Fig. 15. Unscramble output. (a) Helium speech input waveform 50 mV/div. (b) Final unscrambler output 1 V/cm horizontal scale 2 ms/div.

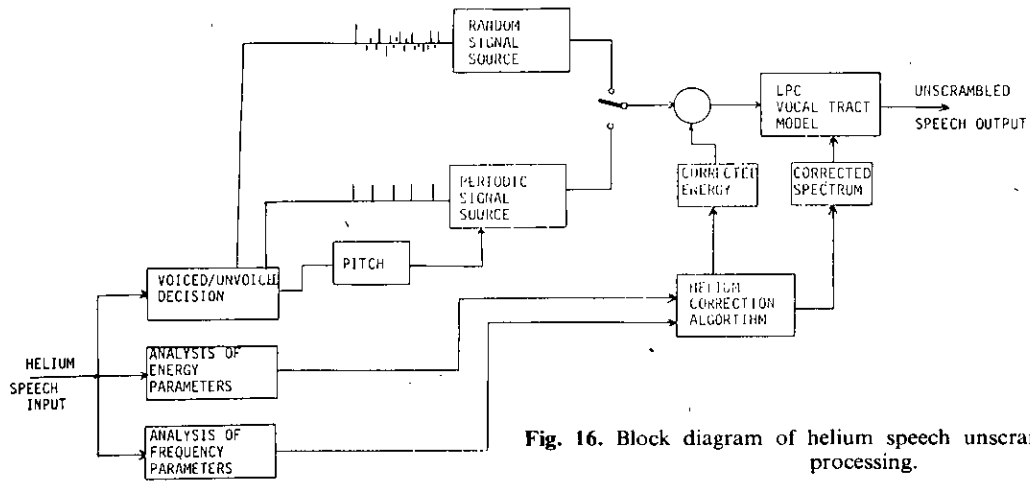


Fig. 16. Block diagram of helium speech unscrambler using LPC processing.

10.4 Unscrambling Techniques Based on Waveform Encoding

In addition to the previously described techniques which have all been implemented practically, linear predictive coding[36] and homomorphic deconvolution[37] have also been applied to the problem of helium speech unscrambling, although neither has yet been implemented practically in a real-time unscrambler. Although lengthy and time-consuming algorithms are at present necessary for their implementation, these techniques offer the potential to perform non-linear frequency shifting for future high performance unscrambling systems.[38]

10.4.1 *Unscrambling by linear predictive coding*

In linear predictive coding (LPC), the speech signal is sampled and analysed to yield data such as pitch period, voiced or unvoiced decision, voice intensity and vocal tract response parameters. The latter parameters can be used to control the response of a time-variant, multi-stage, recursive, digital filter which is excited by a voiced/unvoiced excitation source (Fig. 16). The filter represents a model of the vocal tract, and is capable of predicting the current speech sample based on a weighted linear combination of previous speech samples. The system works on the assumption that the vocal tract characteristics remain constant over a 20–25 ms period so that new filter values from the analysis section are computed and corrected according to some helium–air correction algorithm every 20 ms.

Although computer-simulated unscrambling using LPC analysis has been evaluated,[39] no real-time device has yet been used for helium unscrambling based on this technique, and the successful use of any such LPC system will depend heavily upon the ability to model successfully the helium speech effect in order to apply dynamic (non-linear) correction to the filter coefficients.

10.4.2 *Unscrambling by homomorphic deconvolution*

The speech signal is the result of a convolution operation between the glottal air excitation injected through the vocal cords into the vocal tract and the impulse response of the vocal tract. Signal processing by homomorphic deconvolution[37] involves a transformation which yields a function known as the cepstrum. The cepstrum contains information relating to fundamental pitch period and frequency/amplitude information for the speech signal. If the cepstrum of the glottal air excitation itself is then known or can be approximated, it can be subtracted directly to yield a modified cepstrum which contains only information relating to the deconvolved vocal tract response. The vocal tract response is assumed constant over the 20 ms analysis period of the cepstrum computation. Figure 17 shows the homomorphic (cepstrum) process applicable to this situation.

The required non-linear frequency processing to achieve unscrambling of the helium speech can be applied in computation of the inverse cepstrum.[40] The resulting output of the homomorphic analysis system is

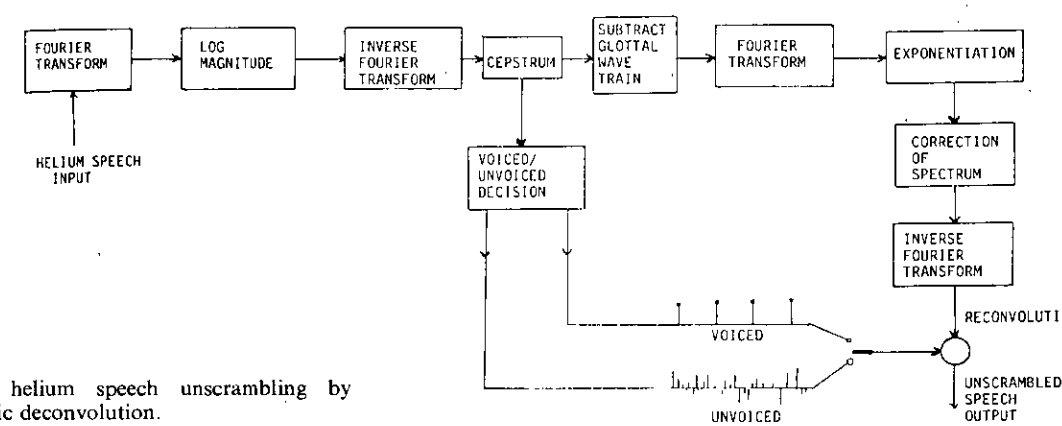


Fig. 17. Block diagram of helium speech unscrambling by homomorphic deconvolution.

then reconvolved with a periodic pitch impulse source or unvoiced random signal source. This proposed technique is ultimately limited by the requirement to compute Fourier transformations in real time, even when using fast Fourier transform (FFT) techniques.[41,42] In addition, the system assumes prior knowledge of the glottal air excitation waveform in helium to permit deconvolution at the cepstrum stage. Thus the derived vocal tract impulse response may be corrupted by remnants of the incorrectly deconvolved glottal air excitation function, thereby limiting the performance of the technique.

11 Factors Degrading the Intelligibility of Unscrambled Helium Speech

Many of the above unscrambling techniques involve some form of voiced/unvoiced sound detection based on a pitch extraction algorithm whereby if a pitch cycle is detected, the sound is assumed voiced, if not, it is assumed unvoiced. On the basis of this judgement the time domain 'expanded segment' principle stores each pitch-synchronous segment on detection of a pitch cycle. In the absence of a pitch decision, the unscrambler segments the speech in a repetitive fashion, or alternatively, may output white noise. In the waveform coding techniques of linear predictive coding and homomorphic deconvolution, either a pitch-synchronous impulse train or a random signal source is used to resynthesize the speech. This strategy tends to form an important source of degradation of intelligibility, since the air excitation waveform injected through the vocal cords, although up to this point assumed to be a perfect Dirac impulse, is in fact a complex waveform.[43]

In view of the fact that pitch detection is carried out on the emitted helium speech signal (which is the convolution of the vocal cord excitation and vocal tract impulse response) and especially since the diver may be working a high ambient noise environment, it seems reasonable to assume that there may be a significant probability of pitch extraction errors in helium speech unscrambler operation.[29]

In analysis/synthesis unscrambler systems, an unvoiced decision causes a random noise source to be used as the synthesis excitation source and, for the pitch-synchronous time-expansion unscrambler, the sampled and expanded waveform loses pitch synchronism for such an unvoiced decision. Assuming that such unvoiced decision errors occur amidst true voiced speech, these errors can be viewed as 'noise' intervals, as far as the human ear is concerned. Figure 18 shows results of the effects on intelligibility of alternating intervals of equal length of speech and white noise.[44] The intelligibility is measured in terms of the percentage of words heard correctly, and is plotted against frequency of interruption. Interruption frequencies ranged from 0·1 to 10000 per second, and the signal-to-noise ratio was varied from −18 to +9 dB. Shown also is the response, marked 'Quiet', when the intervals between speech were silent, with no added noise.

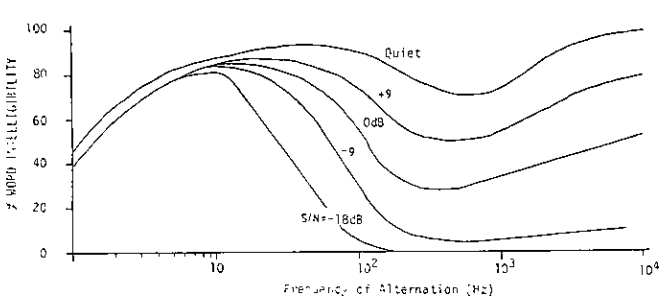


Fig. 18. Word intelligibility as a function of the frequency of alternation between speech and noise, with signal-to-noise-ratio in dB as the parameter (from Ref. 44).

It can be seen from Fig. 18, that, in the range 10 to 560 noise interruptions per second, intelligibility is increasingly impaired. The pitch frequencies of most male and female speakers fall within this range; hence if a succession of unvoiced pitch decisions are made in error of actual voiced speech sounds, then that sound will be effectively masked in noise as the pitch detector causes unvoiced sound to be output. In addition to masking the actual voiced speech sound, it is conceivable that vowel-consonant transitions, which have already been identified as playing an important role in intelligibility,[17,18] may be degraded.

Another possible source of degradation of intelligibility and loss of voice quality can be appreciated from the facts, stated earlier, that in a helium atmosphere the frequency spectrum is shifted in a non-linear fashion and that, due to amplitude spectrum distortion, voiced sounds are enhanced at the expense of unvoiced sounds (high frequency energy is heavily attenuated). All of the unscramblers developed to date which are based solely on time or frequency domain techniques involve a shift of the frequency spectrum in a piecewise linear fashion and make no attempt to correct amplitude spectrum distortion. However, since it has been shown that vowels are recognized by their formant frequency ratios, and that speech intelligibility is a function of vowel–consonant intensity ratio fluctuations, then a correlation exists between these facts and loss of intelligibility when using present unscramblers in that formant frequency ratios may not be conserved by simple linear frequency transformations, and since no amplitude compensation is applied, vowel–consonant ratios are adversely affected. Hence, although the technology used in present-day unscramblers continues to advance in terms of reduction of unscrambler size and minimization of power consumption, device performance continues to be non-optimal in terms of intelligibility.[45,46]

12 Conclusions

The mechanisms affecting speech intelligibility in normal air have been identified. These include the dependence of vowel recognition on the relative formant frequency ratios, and the recognition of nasality by the attenuation of formant amplitudes and the broadening of formant bandwidths. Transitions of formant frequencies and the fluctuation of vowel–consonant intensity ratios have likewise been identified as playing an important role in speech intelligibility.

Similarly, the effects of a helium atmosphere upon the speech spectrum have been discussed. These include a

non-linear shift of the frequency spectrum, particularly at low frequencies, which destroys relative formant frequency ratios, thereby confusing the recognition of vowel sounds; broadening of formant bandwidths due to alterations in the acoustic impedance of the vocal tract, hence causing the characteristic nasal quality of helium speech; and severe attenuation of high-frequency sound, thereby affecting vowel–consonant intensity relationships which are important in providing transitional cues from one sound to another.

Of the various unscrambling techniques reviewed in this paper, those techniques which involve frequency-domain processing either as the main source of unscrambling or as part of some composite strategy, as in waveform coding techniques, offer the potential of providing the non-linear shift necessary for helium speech unscrambling. Here, segmentation of the helium speech spectrum and subsequent piecewise relocation of each passband within the normal speech frequency range restores formant frequencies to their original areas of the spectrum. However, correction of individual formant bandwidths, and hence successful denasalization of the helium speech, is not necessarily implicit. Frequency domain techniques in which unscrambling is carried out by scaling each segment of the helium speech by a geometric ratio, such as in the Self-excited Vocoder, provide the best opportunity of denasalizing speech since the formant bandwidths are also scaled. On the other hand, techniques involving frequency subtraction, such as by heterodyning, have diminished capability of correcting nasality since the helium speech formant bandwidths are conserved. Unscrambling techniques based on frequency domain processing strategies, including waveform coding techniques, offer the possibility of manipulating selectively the amplitude response of each passband to achieve correction of the high frequency attenuation present in helium speech, thus restoring the vowel–consonant intensity relationships which have been identified as having an important bearing upon speech intelligibility. None of the unscramblers developed to date, however, has incorporated any useful means by which the amplitude spectrum may be corrected for this high frequency attenuation.

Time-domain processing techniques such as pitch-synchronous segmentation and expansion are inherently unable to provide non-linear frequency shifting and lack the capability to fully-correct the amplitude spectrum of helium speech. Notwithstanding the limitations of the basic technique, however, this type of unscrambler is currently in widespread use, due mainly to its simplicity, low cost and reliability.

Future developments in unscrambler design appear to be evolving on two major fronts. First, advances in implementation technology have already led to a miniaturization of the time-domain technique into a compact form for diver-borne use, with a view to further development to single-chip form, opening up the possibility for diver-to-diver and diver-to-surface through-water communications. Secondly, improvements in unscrambler system design, based on frequency domain techniques, will lead to a more faithful reproduction of the diver's voice. Techniques which use frequency-domain processing are the only options with the capability of including all the features necessary for good intelligibility of the unscrambled helium speech.

Thus, over the next few years, both new technological implementations and new systems developments can be expected in helium speech unscrambling. These new systems will result in a new generation of unscramblers, with high quality performance in terms of intelligibility of the unscrambled helium speech, increased diving efficiency and above all, increased diver safety.

13 Acknowledgments

The authors wish to acknowledge the help of Dr J. Laver, Department of Linguistics and Phonetics, University of Edinburgh; Dr L. E. Virr, Admiralty Marine Technology Establishment, Experimental Diving Unit, HMS *Vernon*, Portsmouth, and Mr C. S. Roper, Dive Equipment Technology Ltd, Aberdeen. This work has been carried out under MoD sponsorship.

14 References

1 Laver, J. D., 'The Phonetic Description of Voice Quality' (Cambridge University Press, 1980).
2 Van den Berg, J., 'Myoelastic/aerodynamic theory of voice production', *J. Speech and Hearing Res.*, 1, no. 3, pp. 227–44, 1958.
3 Holmes, J. N., 'Speech Synthesis' (Mills and Boon, London, 1972).
4 Fant, G., Lindqvist, J., Sonesson, B. and Hollien, H., 'Speech distortion at high pressure', 'Underwater Physiology', Proc. 4th International Congress on Diver Physiology, pp. 293–9, 1971.
5 Fant, G. and Sonesson, B., 'Speech at High Ambient Air Pressure', Quarterly Progress Report, Speech Transmission Lab, 5IL–QPSR, Stockholm, pp. 9–21, 1964.
6 Giordano, T. A., Rothman, H. B. and Hollien, H., 'Helium speech unscramblers—a critical review of the state of the art', *IEEE Trans.*, **AU-21**, pp. 436–44, October 1973.
7 Beil, R. G., 'Frequency analysis of vowels produced in a helium-rich atmosphere', *J. Acoust. Soc. Am.*, **34**, pp. 347–9, 1962.
8 Nakatsui, M., 'Comment on helium speech-insight into speech event needed', *IEEE Trans.*, **ASSP-22**, pp. 472–3, 1974.
9 Hollien, H., Shearer, W. and Hicks, J. W., 'Voice fundamental frequency levels of divers in helium–oxygen speaking environments', *Undersea Biomed. Res.*, **4**, pp. 199–207, June 1977.
10 Mathews, M. V., Miller, J. E. and David, E. E., 'Pitch synchronous analysis of voiced sounds', *J. Acoust. Soc. Am.*, **33**, pp. 179–86, 1961.
11 Copel, M., 'Helium voice unscrambling', *IEEE Trans.*, **AU-14**, no. 3, pp. 122–6, September 1966.
12 Shilling, C. W., Werts, M. F. and Schardel-Meier, M. R., 'Underwater communications', in 'The Underwater Handbook' (Wiley, New York, 1976).
13 Nakatsui, M., Suzuki, J., Tagasugi, T. and Tanaka, R., 'Nature of helium speech and its unscrambling', Conf. Rec. IEEE Conf. on Engineering in the Ocean Environment, pp. 137–40, July 1973.
14 Maclean, D. J., 'Analysis of speech in a helium–oxygen mixture under pressure', *J. Acoust. Soc. Am.*, **40**, no. 3, pp. 625–7, 1966.
15 Morrow, C. T., 'Speech in deep submergence atmospheres', *J. Acoust. Soc. Am.*, **50**, no. 3, pt. 1, pp. 715–28, 1971.
16 Sergeant, R. L., 'Speech during respiration of a mixture of helium and oxygen', *Aerospace Medicine*, **34**, pp. 826–8, September 1963.
17 Brubaker, R. S. and Wurst, J. W., 'Spectrographic analysis of divers speech during decompression', *J. Acoust. Soc. Am.*, **43**, pp. 798–802, 1968.
18 Fairbanks, G. and Miron, M. S., 'Effects of vocal effort upon the consonant–vowel ratio within the syllable', *J. Acoust. Soc. Am.*, **29**, pp. 621–6, 1957.

19 Suzuki, J. and Nakatsui, M., 'Perception of helium speech uttered under high ambient pressures', Publication SCS-74, Speech Communication Seminar, Stockholm, pp. 97–105, August 1974.
20 Fairbanks, G., 'Experimental Phonetics: Selected Articles' (University of Illinois Press, Urbana, 1966).
21 Hollywell, K. and Harvey, G., 'Helium speech', *J. Acoust. Soc. Am.*, **36**, pp. 199–207, 1964.
22 Gerstman, L. J., *et al.*, 'Breathing mixture and depth as separate effects on helium speech' (72nd meeting of the Acoustical Society of America, 1966), *J. Acoust. Soc. Am.*, **40**, no. 5, pp. 1283, 1966 (Abstract).
23 Golden, R. M., 'Improving naturalness and intelligibility of helium speech, using vocoder techniques', *J. Acoust. Soc. Am.*, **40**, no. 3, pp. 621–4, 1966.
24 Zurcher, J. F., 'Voice transcoder', British Patent specification 1561918, 5th March 1980.
25 Ananthapadmabha, T. V. and Yegnanarayana, B., 'Epoch extraction of voiced speech', *IEEE Trans.*, **ASSP-23**, no. 6, pp. 562–9, December 1975.
26 Moorer, J. A., 'The optimum comb method of pitch-period analysis of continuous digitized speech', *IEEE Trans.*, **ASSP-22**, no. 5, pp. 330–8, October 1974.
27 Noll, A. M., 'Cepstrum pitch determination', *J. Acoust. Soc. Am*, **41**, no. 2, pp. 293–309, 1967.
28 Tucker, W. H. and Bates, R. H. T., 'Efficient pitch estimation for speech and music', *Electronics Letters*, **13**, no. 12, pp. 357–8, June 1977.
29 Rabiner, L. R., Cheng, M. J., Rosenberg, A. E. and McGonegal, C. A., 'A comparative performance study of several pitch detection algorithms', *IEEE Trans.*, **ASSP-24**, no. 5, pp. 399–417, October 1976.
30 Hess, W. J., 'Pitch determination—an example for the application of signal processing methods in the speech domain', in 'Signal Processing: Theories and Applications', ed. Kunt, M. F., and de Coulon, F. (North Holland, Amsterdam, 1980).
31 Stover, W. R., 'Technique for correcting helium speech distortion', *J. Acoust. Soc. Am.*, **41**, pp. 70–4, 1967.
32 Gill, J. S., *et al.*, British patent 1,321,313. June 1970.
33 Flower, R. A. and Gerstman, L. J., 'Correction of helium speech distortion by real-time electronic processing', *IEEE Trans.*, **COM-19**, no. 3, pp. 362–4, June 1971.
34 Jack, M. A., Milne, A. D. and Virr, L. E., 'Compact helium speech unscrambler using charge transfer devices', *Electronics Letters*, **15**, no. 18, pp. 548–50, August 1979.
35 Jack, M. A., Milne, A. D., Virr, L. E. and Hicks, R., 'Compact helium speech unscrambler for diver-borne use', Conference on Electronics for Ocean Technology, pp. 13–18. IERE Conference Proceedings no. 51, 1981.
36 Makhoul, J., 'Linear prediction—a tutorial review', *Proc. IEEE*, **63**, pp. 561–81, April 1975.
37 Oppenheim, A. V. and Schafer, R. W., 'Digital Signal Processing' (Prentice-Hall, Englewood Cliffs, 1975).
38 Schafer, R. W., 'A survey of digital speech processing techniques', *IEEE Trans.*, **AU-20**, no. 1, March 1972.
39 Suzuki, H., Ooyama, G. and Kido, K., 'Analysis-conversion-synthesis system for improving naturalness and intelligibility of speech at high pressure in a helium gas mixture', Publication SCS-74, Speech Communication Seminar, Stockholm, pp. 97–105, August 1974.
40 Quick, R. F., 'Helium speech translation using homomorphic deconvolution' (79th meeting of the Acoustical Society of America, 1970), *J. Acoust. Soc. Am.*, **48**, p. 130, 1970 (Abstract).
41 Bergland, G. D., 'A guided tour of the Fast Fourier Transform', *IEEE Spectrum*, **6**, pp. 41–52, July 1969.
42 Girgis, A. A. and Ham, F. H., 'A quantitative study of pitfalls in the FFT', *IEEE Trans.*, **AES-16**, no. 4, pp. 434–9, July 1980.
43 Miller, R. L., 'Nature of the vocal cord wave', *J. Acoust. Soc. Am.*, **31**, no. 6, pp. 667–77, 1959.
44 Miller, G. A. and Licklider, J. C. R., 'The intelligibility of interrupted speech', Bell Laboratories Internal Memorandum, 1950.
45 Brown, D. D. and Feinstein, S. H., 'An evaluation of three helium speech unscramblers to a depth of 1000 feet', *J. Sound Vibration*, **48**, no. 1, pp. 123–35, 1976.
46 Hollien, H. and Rothman, H. B., 'Evaluation of helium speech unscramblers under controlled conditions', *Marine Technol. J.*, **8**, no. 9, pp. 35–44, October 1974.

Manuscript first received by the Institution on 20th July 1981 and in final form on 23rd September 1981
(Paper No. 2020/AMMS 109)

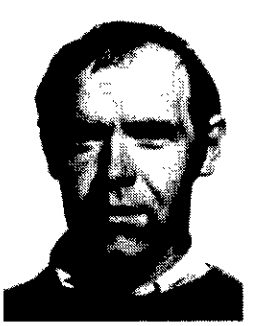
The Authors

Mervyn Jack received the B.Sc. degree in electronic engineering and the M.Sc. degree in digital techniques from the Heriot-Watt University, Edinburgh, in 1971 and 1975, respectively, and the Ph.D. degree from the University of Edinburgh in 1978.

From 1971 to 1975 he worked as a project engineer with Microwave and Electronic Systems, Edinburgh, where he was responsible for the design and development of security systems based on passive infra-red and microwave Doppler intruder detectors. In 1975 he took up a Research Fellowship at the University of Edinburgh to study the design and application of Fourier transform processors based on surface acoustic wave and charge coupled devices. In 1979 Dr Jack was appointed to a lectureship in the Department of Electrical Engineering at Edinburgh University, where he is currently involved with teaching and research into v.l.s.i. architectures for digital signal processing.

After receiving the B.Sc. degree in electrical and electronic engineering from the Heriot-Watt University, Edinburgh, in 1978, **George Duncan** worked in the Libyan desert as a Field Seismologist for Seismograph Service, Bromley. During the academic year 1979–1980, he was a Research Engineer in the Ecole Supérieure d'Ingénieurs en Electronique et Electrotechnique (ESIEE) in Paris, engaged in research into the microprocessor-based control of heat-pump systems.


M. JACK


G. DUNCAN

Since September 1980 Mr Duncan has been a Research Associate with the Wolfson Microelectronics Institute, University of Edinburgh, engaged in research into advanced techniques for unscrambling helium speech, and is reading for a Ph.D. degree.

UDC 621.3.049.771.14: 621.377.622.25
Indexing Terms: Circuit design, logic, Integrated circuits

Optimizing gate interconnections in four-phase dynamic logic m.o.s. l.s.i. technology

D. C. PATEL, B.Sc., M.Sc., Ph.D.*

SUMMARY

Dynamic logic circuit technology has been used widely to implement large-scale integrated (l.s.i.) circuits using metal–oxide–silicon field effect transistors (m.o.s.f.e.t.). One of the factors which determines the overall dimensions of a custom-designed random logic l.s.i. circuit is the number of interconnection tracks as they occupy a large part of the chip between functional modules. This paper describes a multiplexing technique, which allows a reduction in the number of interconnection tracks between modules in l.s.i. circuits implemented using the four-phase dynamic logic technology of the major–minor configuration. It is shown that the operating speed or performance of the circuit is not affected by this technique.

* *Department of Electronic and Electrical Engineering, University of Surrey, Guildford GU2 5XH.*

1 Introduction

With the advent of large-scale integrated circuits, it is feasible to implement complex functions on a single silicon chip. A typical l.s.i. circuit may incorporate several hundred gates or several thousand transistors. Advances in semiconductor fabrication technology and in mask making techniques have led to a reduction in the geometrical dimensions of each individual transistor; thereby increasing the packing density and the circuit complexity.

When realizing l.s.i. random logic circuits the interconnection tracks between gates are an important factor in determining the chip size. A typical circuit has data highways containing many lines and each line runs over a significant length. Interconnection tracks occupy significantly more area of the chip when compared with the area occupied by the active transistors. Consequently, the overall packing density of gates per unit area is less with random circuits than with regular circuits such as memories. This paper describes a multiplexing technique which reduces the number of lines in the data highways for customized m.o.s.l.s.i. circuits, using the major–minor configuration of the four-phase dynamic logic circuit implementation.[1] The operating speed or the performance of the circuit is not affected in any way although there is a reduction in the chip dimension and hence an increase in the overall packing density in terms of gates per unit area of the chip. The advantages of four-phase dynamic logic technology are that the power dissipation is reduced as there is no d.c. path between the supply and ground, the circuit may be implemented using the minimum size device thereby increasing the packing density, and several functions such as shift-register can be implemented using fewer devices. Because clocks are used to synchronize the operation of the circuit, the design is free of race hazards.

2 Optimization of Data Highway

To facilitate the implementation of an l.s.i. circuit, the designer partitions the logic into functional modules. Those with many interconnection wires are located adjacent to each other to minimize the chip area. Because of the very nature of the custom-designed integrated circuits, the layout is complex and random; even after optimum partitioning, a significant chip area is occupied by data highways containing many interconnected tracks. Modules which have a smaller number of interconnection tracks may be separated topologically by large distances and each interconnection wire increases the chip dimensions. Figure 1 shows part of a

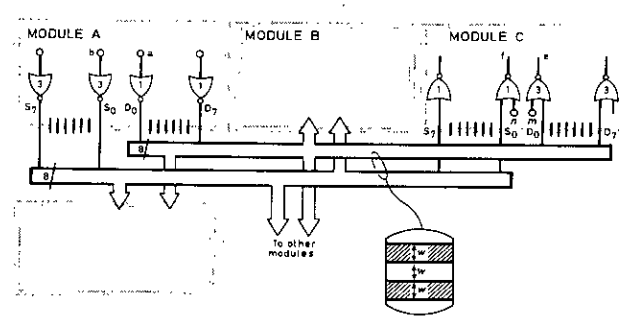


Fig. 1. Chip partitioning. Modules and data organization.

The Radio and Electronic Engineer, Vol. 52, No. 5, pp. 224–226, May 1982

0033-7722/82/050224+03 $1.50/0

RESIDUALLY EXCITED LPC PROCESSOR FOR ENHANCING HELIUM SPEECH INTELLIGIBILITY

Indexing terms: Signal processing, Speech processing

A new approach to restoring the intelligibility of divers' speech uttered in a high-pressure atmosphere containing high percentages of helium gas is presented. The system processing architecture consists of a residually excited linear predictive coder (RELPC) with a novel implementation of time/frequency-domain relationships to simplify processing.

Introduction: The alleviation of physiological discomfort in deep-sea saturated diving operations demands the use of helium-oxygen (heliox) breathing mixtures containing large amounts of helium gas, whose effect on the speech waveform is to shift spectral resonances (formants) by a factor nominally equal to the ratio of the velocity of sound in the heliox mixture to that in air, although the actual spectral shift is nonlinear, resulting in the degraded intelligibility of helium speech.[1] However, temporal features of the speech waveform, such as fundamental frequency of vocal tract excitation (pitch), are conserved.

The residually excited linear predictive coding (RELPC) unscrambler system detailed in this letter permits manipulation of the helium speech waveform to improve intelligibility by the use of autoregressive signal modelling, in which the temporal features of the speech signal are conserved, but speech formant data can be corrected nonlinearly.

System architecture: The system processing strategy reported here is based on an all-pole filter model whose poles correspond to the formats produced by the resonating cavities of the vocal tract. In the RELPC processor, Fig. 1, use is made of

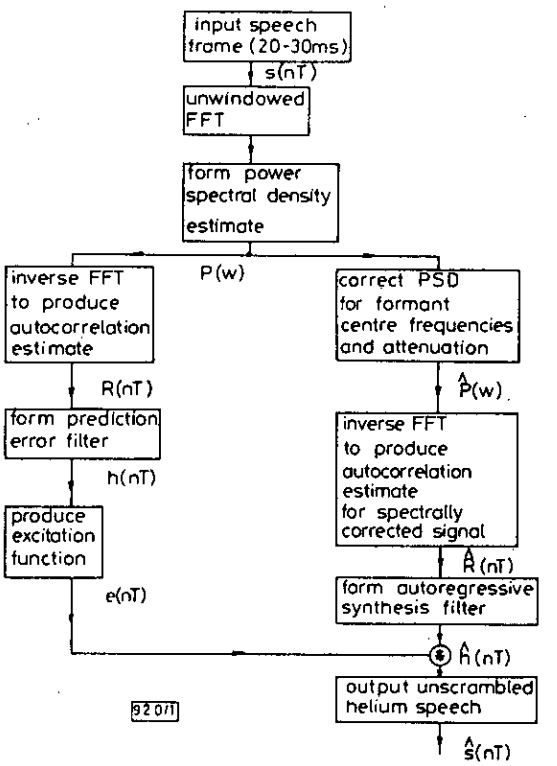


Fig. 1 *Block diagram of residually excited LPC (RELPC) helium speech unscrambler*

a prediction error (analysis) filter which extracts information on the positions of the vocal tract formants from the helium speech signal $s(nT)$, where the discrete time operator nT is used to indicate sampled data. The output (residual) signal $e(nT)$ from this filter is the vocal tract excitation function, which is either a periodic pulse train for voiced speech or random noise for unvoiced speech, and the analysis filter coefficients represent the transfer function $h(nT)$ of the vocal tract. Since the temporal spacing of the excitation function series is unaffected, only the coefficients of the analysis filter need be

altered according to some predetermined correction algorithm in order to eliminate the helium speech distortion. Thereafter, these corrected coefficients, which now represent the spectrally corrected vocal tract filter function $\hat{h}(nT)$, are used in the autoregressive (synthesis) filter, which is excited by the residual signal $e(nT)$ from the analysis filter. The resultant output of the system, $\hat{s}(nT)$, is then the unscrambled helium speech signal:

$$\hat{s}(nT) = e(nT) * \hat{h}(nT) \tag{1}$$

where the symbol (*) denotes convolution. Thus, the pitch of the original speech waveform is conserved, but the spectral harmonic content has been corrected for the helium speech distortion.

An example demonstrating the correction of helium speech is shown in Fig. 2. Fig. 2*a* shows the power spectral density

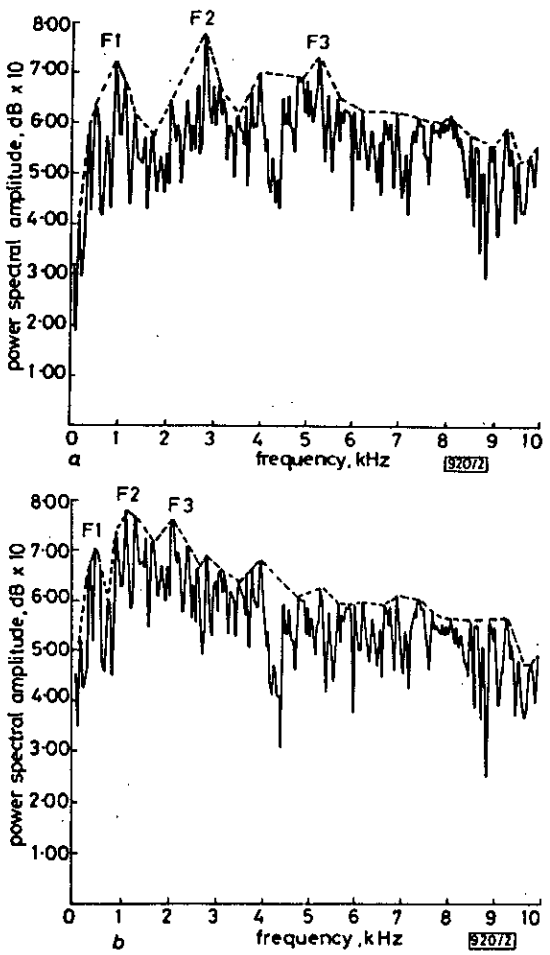


Fig. 2

a Helium speech power spectral density for the vowel /a/ (25 ms) before processing by the RELPC unscrambler
(Note: pre-emphasis factor of 0·98 applied to input data)
b Resultant power spectral density corrected for helium speech distortion

(PSD) of a 25 ms section (20 kHz sample rate) through the voiced vowel /a/ in the word 'had' spoken by a diver at 100 ft depth in a respiratory environment consisting of 91·6% helium and 8·4% oxygen. The PSD for the same section of speech, corrected by the RELPC processor for the helium speech distortion, is shown in Fig. 2*b*. Note that formants F1–F3 have been shifted downwards in frequency to their normal air values, thereby restoring intelligibility; furthermore, the underlying fine spectral detail relating to pitch and source excitation spectral characteristics has been totally conserved.

The route to both the analysis and synthesis filter coefficients in the RELPC unscrambler system is through the PSD distribution $P(w)$. Basically, in order to obtain values for either the analysis or the synthesis filter coefficients, use is made of the Levinson recursion relationship,[2] which relates the autocorrelation function (ACF) $R(nT)$ of the time signal to the prediction error filter coefficients $h(nT)$ in a computationally efficient manner.

As an alternative to computing the ACF explicitly from the time series, the approach adopted in this system is to estimate the ACF from the PSD utilising the Weiner–Kinchine theorem:[3]

$$R(\tau) = \int_0^\infty P(w) \exp(jw\tau)\, dw \tag{2}$$

This strategy is convenient since the PSD implicitly contains spectral information relating to the vocal tract formants and can therefore itself be corrected for the helium speech characteristic. The resulting PSD estimate $\hat{P}(w)$ corresponding to the corrected helium speech signal $\hat{s}(nT)$ can be used to form the autocorrelation function $\hat{R}(nT)$ of the spectrally corrected signal, from which the synthesis filter structure corresponding to the corrected vocal tract transfer function $\hat{h}(nT)$ can be computed. The route to both analysis and synthesis filters can be made very expedient by the use of fast Fourier transform (FFT) processors, and is particularly efficient since the forward FFT to obtain the initial estimate of the PSD requires no windowing of the time series. Notice also that construction of both the analysis and synthesis filters can be carried out simultaneously.

Conclusion: A system has been discussed which permits helium speech correction through the use of a residually excited linear predictive coding (RELPC) processor. Simulation results for the system have been judged subjectively in informal listening tests to produce an improved intelligibility and naturalness in the unscrambled speech output. The system conserves the temporal features of the input waveform but corrects spectral features in a nonlinear fashion commensurate with the speech distortions produced in a pressured heliox environment, thereby producing an intelligible speech signal.

Acknowledgments: The authors wish to thank M. J. Rutter and Dr. C. F. N. Cowan, Department of Electrical Engineering, University of Edinburgh, for their help and invaluable discussions. This work has been carried out with the support of the Procurement Executive, Ministry of Defence.

G. DUNCAN

Wolfson Microelectronics Institute
University of Edinburgh
Mayfield Road, Edinburgh EH9 3JL, Scotland

M. A. JACK

Department of Electrical Engineering
University of Edinburgh
Mayfield Road, Edinburgh EH9 3JL, Scotland

29th June 1983

References

1 JACK, M. A., and DUNCAN, G.: 'The helium speech effect and electronic techniques for enhancing intelligibility in a helium-oxygen environment', *Radio & Electron. Eng.*, 1982, **52**, pp. 211–223
2 LEVINSON, N.: 'The Weiner r.m.s.-error criterion in filter design and prediction', *J. Math. Phys.*, 1947, **25**, pp. 261–278
3 WEINER, N.: 'Generalized harmonic analysis', *Acta Mathematica*, 1930, **55**, pp. 117–258

Appendix B

HARDWARE AND SOFTWARE DESIGN OF THE
DIGITAL-TO-ANALOG SPEECH
CONVERTER INTERFACE.

APPENDIX B

HARDWARE AND SOFTWARE DESIGN OF THE DIGITAL-TO-ANALOG SPEECH CONVERTER INTERFACE.

In order to successfully simulate the effects upon helium speech intelligibility of various unscrambling techniques, it was necessary to devise an electronic digital to analog converter system which would allow analog speech output over a satisfactory range of simulated sample frequencies in either real time or an integer division of real time to permit subsequent (speeded) playback from a tape recorder at real time speed. The following report summarises both hardware and supporting software design of the converter system together with relevant details regarding the host device and its operating system to which the data-to-speech conversion device is interfaced, namely the Zilog PDS microprocessor development system.

B.1 INTRODUCTION

Analog helium speech data for analysis is digitised on a single-user PDP11/40 computer whose geographical location is far removed from the location of the computer on which the system simulations will be effected. The digitised speech is sampled to 12-bit resolution and stored as 16-bit integers on an RK05 disk. Transfer of digitised speech data to the destination Vax 750 computer is achieved by first of all transferring the data, stored on the RK05 disk in 'RT11' system format, to tape using an intermediate PDP11/60 machine with disk-to-tape

transfer facilities. The resulting tape is then transported to the Vax 750 machine where the data is downloaded into the Vax, converted to 'Unix' system format, then written out to a second tape for long term storage. This digitised speech data is subsequently retrieved and furnishes the data for the simulation, in floating-point software, of various helium speech unscrambler systems.

The Vax computer is, however, time-shared in this case, and it would therefore be difficult to output reasonable streams of uninterrupted speech data in real time in order to permit subjective evaluation as to improvements in speech intelligibility afforded by any single unscrambler technique. Possible solutions to this encumbrance include denying system access to other users at certain times and employing an analog/digital tape recorder in a start/stop mode synchronised to the bursts of analog/digital speech data. Neither of the above solutions, however, are very satisfactory, and it was therefore decided to transfer the digital data to a dedicated long term storage system capable of outputting uninterrupted streams of digitised speech of long duration to a D/A conversion device, which the dedicated system should also be in a position to control.

The system selected for the task was a Zilog Z80 PDS microprocessor system with floppy disk long-term storage, and the data-to-speech conversion unit is interfaced to this machine, supported by controlling software which is first loaded into the Zilog system.

The unit offers effective real time output sample rates from 4.5kHz to 9.12kHz and, with a variable-speed tape recorder, a wide range of sample rates can be simulated. Provision has been made in the controlling software to allow for two Zilog disk drives to be used, thus permitting (theoretically) an infinite amount of data output dependent only on the

number of floppy disks available, although at the present time, only one disk drive is linked to the Zilog system.

In the discussion which follows, firstly certain aspects of the Zilog disk operating system are investigated. These details are then used as a basis for the design of the data-to-speech conversion unit both in terms of hardware and controlling software. Note that certain system variables are expressed here from time to time as hexadecimal integers, denoted by xxh; eg. 20h = 32_{10}.

B.2 THE ZILOG DISK OPERATING SYSTEM

In order to achieve an uninterrupted data flow, it will be necessary (1) to be able to command at will the disk drive head in both a 'read' and 'step' capacity; (2) data output ports will have to be identified to allow digital speech data and status data to be passed to the external D/A conversion circuit; (3) data input ports must be identified to allow input of circuit status to the Zilog controlling software; (4) if the data-to-speech converter is to simulate several different sample rates, then it will need to have its own on-board clock in order to clock data bytes out, preferably derived from existing Zilog signals; (5) if the speed at which the disk is read is to be synchronised to the conversion rate of the converter unit, then the unit will require on-board buffer storage whose condition must somehow be monitored.

B.2.1 DISK DRIVE HEAD MOVEMENT CONTROL

In order to explore the manner in which the floppy disk is driven, it was necessary to disassemble the resident operating system software

to identify disk commands and protocol, since little published data concerning this could be found. Relevant details are as follows.

There are 78 tracks per disk, and 32 sectors per track; each sector holds 128 bytes of data.

Disks are in hard-sector format.

The disk rotation speed is 360r.p.m (1 cycle every 167mS).

Time to step drive head in either direction (in or out) is 10mS, +10mS settling time.

Disk data port address is 0CFh.

Disk status port address is 0D1h.

Disk control port address is 0D0h.

The bits in the disk status byte are:-

bits 0,1,2 form together a b.c.d. representation of the requested drive (0—7);

bit 3 indicates whether the requested drive is ON (bit 3='1') or OFF(='0');

bit 5 shows a 'start sector' pulse every time a hard sector marker is crossed;

bit 7 shows write protect status of the disk;

bits 4 and 6 are unused.

The bits in the disk control byte are:-

bit 0:-step direction ('1'=OUT towards track 0);

bit 1:-STEP command ('1'=STEP);

bits 2,3,4:-indicate Read/Write mode;

bit 5:-disk ready? (low true);

bit 6:-track 0? (low true);

bit 7:-CRC error? (high true).

B.2.2 DISK DATA FORMAT

In the operating system, disk sector read/write operations are effected by interrupt routines generated by interrupts from a clock/timer circuit every time a hard sector marker goes by. Note in particular that although each disk has 33 index holes (giving 33 inter-hole areas)

there are only 32 sectors. The last sector is written on either side of the sector 0 index hole marker (3 holes very close to each other). Since interrupts are disabled in the system while carrying out read/write operations to disk, and since the CPU clock is synchronised to the disk data clock, then in writing/reading several sectors, as a sector index hole goes by, the clock simply stops for the duration of its passing, and writing/reading continues on the other side.

The sector preamble contains all-zero bytes, and "start of data" is signalled by bit 7 of the sector address byte being set (sector addresses therefore have values from (80h+00) to (80h+1Fh)).

The order of data in each sector is (1) sector address (2) track address (3) 128 bytes of data (4) linkage (not required here).

In order to move the disk head, it is necessary to send first the command 'STEP' followed by the direction. This will step the head by a distance corresponding to one physical track only.

B.2.3 OUTPUT TIME FROM ZILOG MEMORY TO OUTPUT PORT

Once loaded into memory, it will be necessary to output the data to the external device. The fastest time available when this data is emitted as a stream of 32x128 bytes was observed to be 42mS.

B.2.4 CONCLUSIONS

From the above data, it was decided that the option providing fastest read-time was to read off directly one whole track at a time and output this for storage to the external circuit. Data should also be written beforehand to disk in a sequential manner either from track 0 inwards

to track 77 or vice-versa.

The overall time required per track from starting to read to completion of output is then 167mS (time to read one track) +42mS (time to output 4096 bytes to an external port) =209mS. This fixes the maximum theoretical data rates as 19.598×10^3 bytes/second (assuming that the time to step the drive head to the next track + settling time occurs within the 42mS allocated for data output).

B.3 INITIAL CIRCUIT DESIGN SPECIFICATIONS

B.3.1 ON-CIRCUIT TRANSITIONAL DATA STORAGE

If the external circuit, as specified earlier, will itself simulate the required sample rate, then there must be some means of buffering between Zilog system disk-read speed and data-to-speech conversion speed. This implies at least two on-circuit memory banks with independent addressing, each having 4096 byte-wide storage. One memory bank will be available for data-to-speech conversion while the other is available to receive fresh data from the Zilog processor. The memory units chosen were the Mostek 2048x8 bit static RAMs, to avoid having to provide refresh logic and timing arrangements. Thus, two 12-bit counters will be required, one to address the memory bank storing fresh Zilog data (input address counter (IAC)) and the other to address the memory bank outputting data for analog conversion (output address counter (OAC)).

B.3.2 DATA OUTPUT ADDRESSING

As specified earlier, the output address counter clock should, if

possible, be derived from the Zilog system. This is done by a simple division process on the Zilog system clock. Down counters are used, being first loaded with a count byte, and every time the counter counts down to zero, a time-out pulse occurs, forming the output address clock and at the same time causing its own count code to be reloaded. Time-out pulses therefore occur at divisions of the Zilog system clock rate (2.3256MHz), so e.g. a count of 119 corresponds to a data throughput of 19.543×10^3 bytes/sec, which is the closest to the theoretical maximum byte output rate that could be handled by the Zilog system. Assuming that 4-bit counters are used (74193 series), then for a down-count of 119 at least two counters are required. For two counters, the maximum down-count is 255, corresponding to a minimum data throughput of 9.120×10^3 bytes/sec, therefore the max/min ratio is 2.143. This ratio is convenient since the tape recorder used has speed division ratios forming a $1/2^n$ series, thus it should be simple to simulate higher sample rates by subsequent high-speed playback.

B.3.3 SIGNAL INTERFACING TO THE ZILOG SYSTEM

An inspection of the Zilog circuitry showed that only one PIO device (2 ports) was available, whereas so far the data required to/from the external circuit is:-

(a) count code corresponding to the desired output sample rate;
(b) state of output address counters in order to synchronise disk-read speed with external circuit memory output;
(c) raw digital speech data for external storage/conversion.

It was decided to use a dedicated PIO port (port A) for raw digital speech data input to the data-to-speech converter. This arrangement will also facilitate the derivation of the input address counter clock

from PIO command signals for port A.

The other PIO port (port B) must now (a) send the output address clock count code and (b) read the output address counter states, or at least 8 bits of this 12-bit address. As will be demonstrated later, this is used by the controlling software to determine when a memory bank is free for data refresh input.

It was further decided to send the output address clock count code only once, at the start of each conversion session. Therefore on-board latches will be required to store this value, since the 74193 series counters chosen have no on-chip latching facilities. Thus, it should be possible, using combinational logic and with a knowledge of the command signals sent by the Zilog system to PIO port B, to derive signals to reset the external circuit into a known state, latch the specified clock code, and also derive 'run' signals to release the circuit from the reset condition.

The circuit control signals to be used then are:-

(a) count code to PIO port B; (reset circuit)

(b) change PIO port B mode to "input data"; (latch clock code) (Note circuit is still reset)

(c) read status of output address counters through PIO port B; (release circuit to run).

B.3.4 D/A DATA CONVERSION

A further specification can be identified from the fact that the data to be output is uncompacted 12-bit data. Two bytes are therefore required before a complete 12-bits is output. By choice, odd addresses will contain the least significant (l.s.) bytes and even addresses the most significant (m.s.) bytes. Thus latching arrangements will be

required at the output, and the simulated sample rate will be exactly 0.5xthe byte throughput rate (output address clock frequency). Therefore a 12-bit D/A converter capable of operating comfortably at nominally 10kHz is required. The device chosen is the Analog Devices AD7521.

B.4 CIRCUIT OPERATION AND SIGNALS

The data-to-speech converter circuit is shown in Fig. B.1, with the board layout in Fig. B.2. Component values, device identification and functional groupings are shown in Fig. B.3.

Counters 1C and 2C cascaded together form the 8-bit output address counter clock whose output is 2C/13 or 1N/3. The clock pulse, via 1N/10 and gated by the system clock ϕz at 1N/9, reloads automatically the count-down code as held by latches 1L and 2L. The "enter code" (latch follow) signal and circuit reset command is derived from a combination of the PIO command signals from the Zilog system, namely $\overline{CE}$, C/D, $\overline{RD}$ and A/B. (see Fig. B.4(a) for waveforms).

B.4.1 CIRCUIT INITIALIZATION

With reference to Fig. B.1, the circuit is specified to be in the 'reset' state when the 0AC count code is sent to the circuit via PIO port B. Specifying the 'reset' signal as coming from a Nand bistable, then a '1' pulse must be derived from the relevant Zilog signals. Gates 4N/3, 4N/6, 9N/3, 9N/6, 3N/3, 3N/11, 3N/8 form the combinational logic required to generate the 'reset' pulse into the 'reset' bistable formed by 2N/6 and 2N/8. Notice that R1 and C1 form an integrating

Continued on pg.308

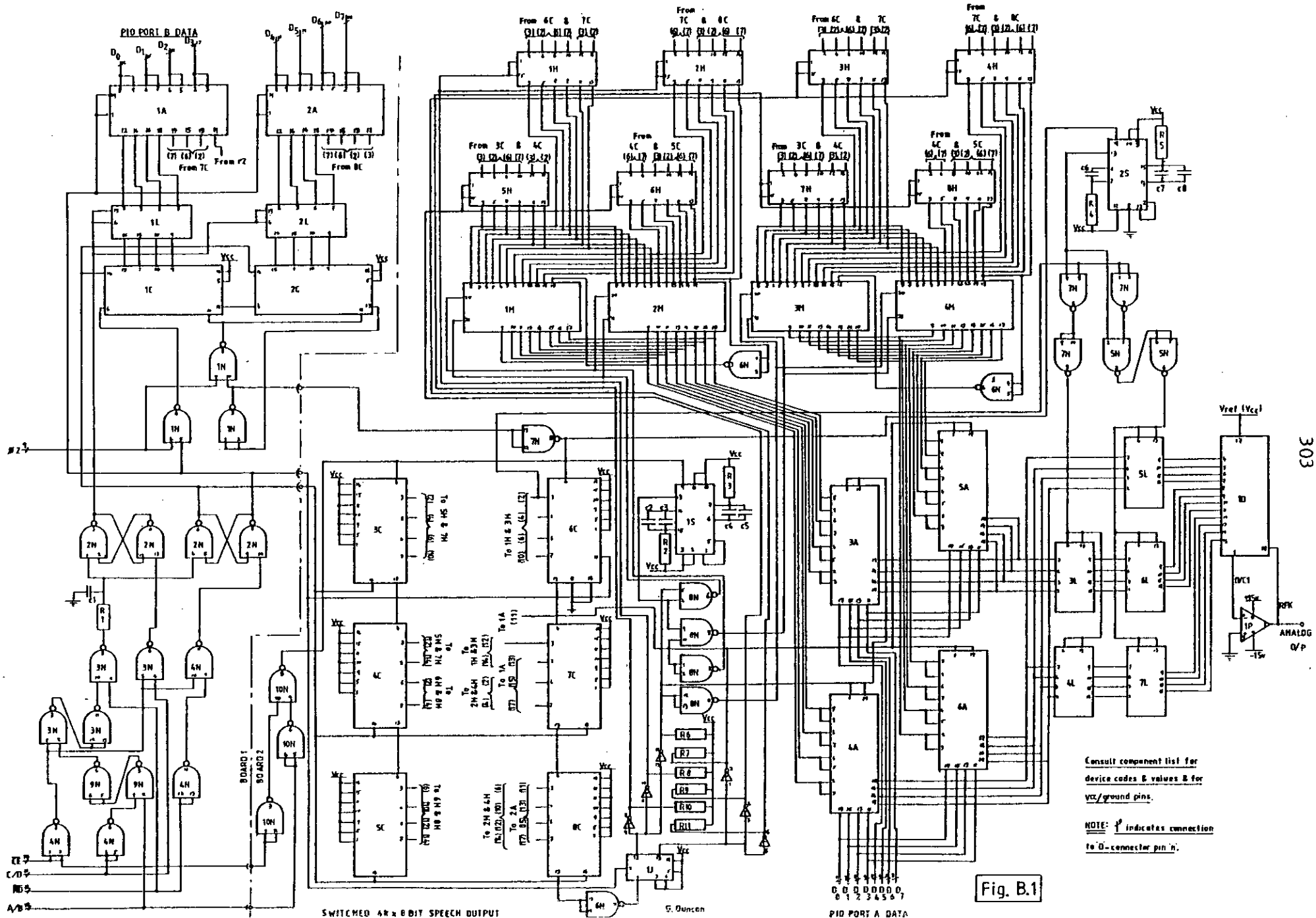
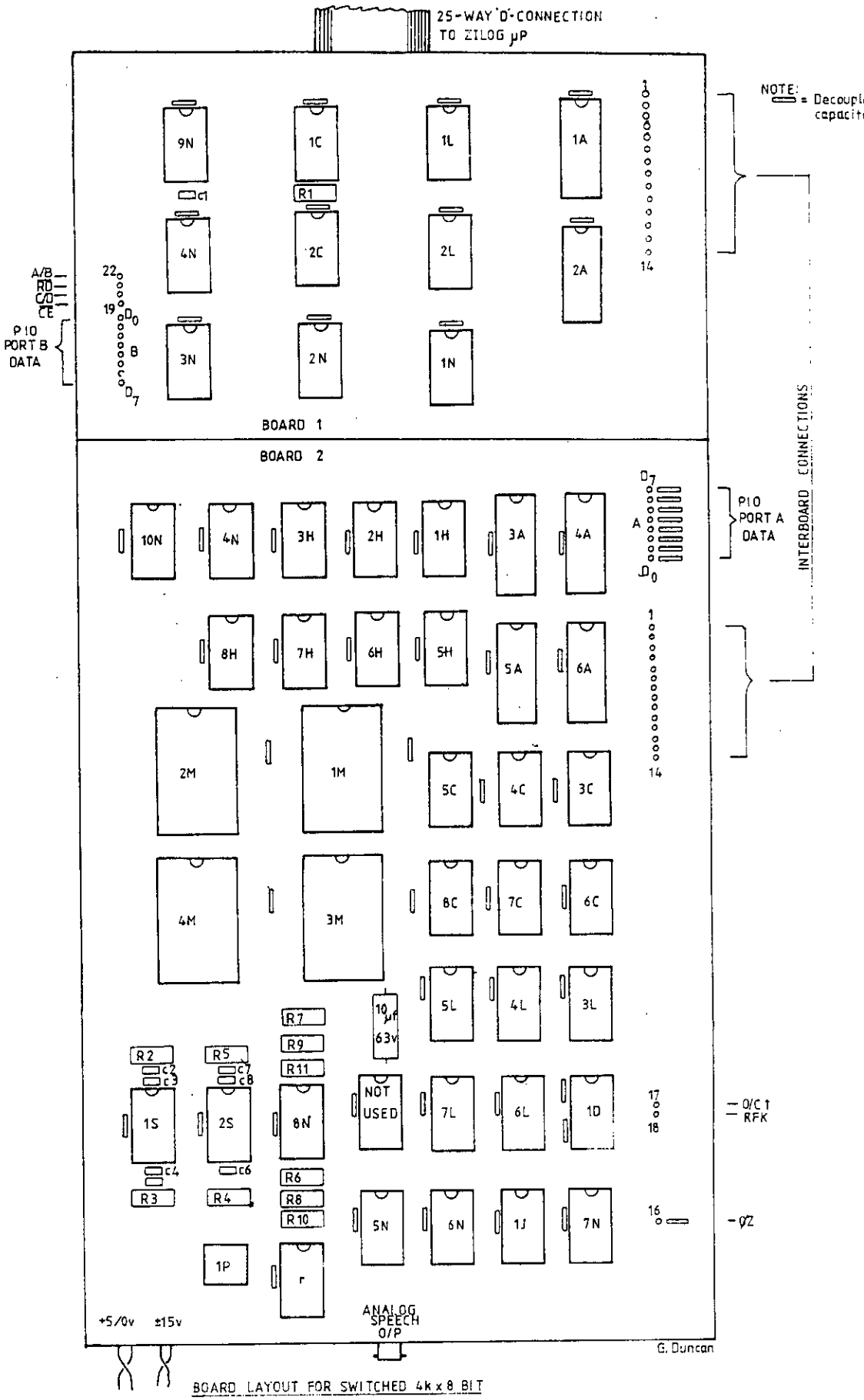

PIO PORT B DATA
From 7C
From 8C
From 1H & 3H
From 2H & 4H
From 5H & 7H
From 6H & 8H
To 1A
To 2A
To 1H & 3H
To 2H & 4H
To 5H & 7H
To 6H & 8H
BOARD 1
BOARD 2
SWITCHED 4K x 8 BIT SPEECH OUTPUT
S. Duncan
PIO PORT A DATA
Vref (Vcc)
AHALOG O/P
Consult component list for device codes & values & for vcc/ground pins.
NOTE: # indicates connection to D-connector pin 'n'.

Fig. B.1



BOARD LAYOUT FOR SWITCHED 4k x 8 BIT SPEECH OUTPUT UNIT FOR ZILOG PDS.

Fig. B.2

Device	Code	Vcc pin	Ground pin	No. of pins	Remarks
74LS00	N	14	7	14	Quad 2-i/p Nand
74LS05	r	14	7	14	Hex O/C inverter
74LS75	L	5	12	16	Dual 2-bit latch
74LS78	J	4	11	14	Dual JK edge trgd flip-flop
74LS123	S	16	8	16	Dual Monostable
74LS193	C	16	8	16	Sync. 4-bit up/down counter
74LS241	A	20	10	20	Tri-state octal buffer
74LS367	H	16	8	16	Tri-state hex buffer
MK4802	M	24	12	24	2kx8 static RAM
AD7521	D	16	3	18	12-bit D/A converter
F136B	P	-	-	8	Dual Op-Amp.

Resistor and Capacitor Values

R1=1kΩ	R4=27kΩ	C1=270pF	C4,C6=68pF
R2=10kΩ	R5=5.6kΩ	C2=220pF	C5,C7=100pF
R3=33kΩ	R6-R11=1kΩ	C3=12pF	C8=15pF

Devices	Function
1A,2A	Sample rate code/output address selector
1L,2L	Latches for sample rate code
1C,2C	output address clock generator
1N	"reload sample rate code" signal
2N	'latch' & 'reset' bistables
3N,4N,9N	Combinational logic for 'reset','latch' & 'run'
10N	Input address clock decoder
3C,4C,5C	Input address counter(IAC)
6C,7C,8C	Output address counter(OAC)
1H,2H	OAC buffers for memory bank 1
5H,6H	IAC buffers for memory bank 1
3H,4H	OAC buffers for memory bank 2
7H,8H	IAC buffers for memory bank 2
1M,2M	Memory bank 1
3M,4M	Memory bank 2
1J	Memory bank switch signal
r	Switch signal line driver
1S	Memory bank write enable pulse generator
3A,4A	Data output/input selector for memory bank 1
5A,6A	Data output/input selector for memory bank 2
2S	'latch output byte' pulse generator
3L,4L	Low byte output latch
5L,6L,7L	12-bit word latch
1D,1P	D/A converter & analog output buffer

Fig. B.3 Device functional groupings and component values for the Speech Converter Interface.

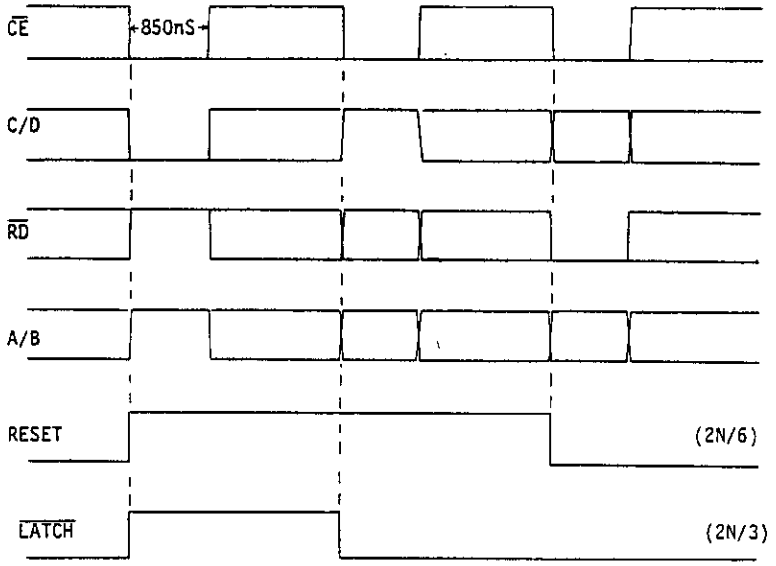


Fig. 8.4(a). Circuit RESET/LATCH Signals

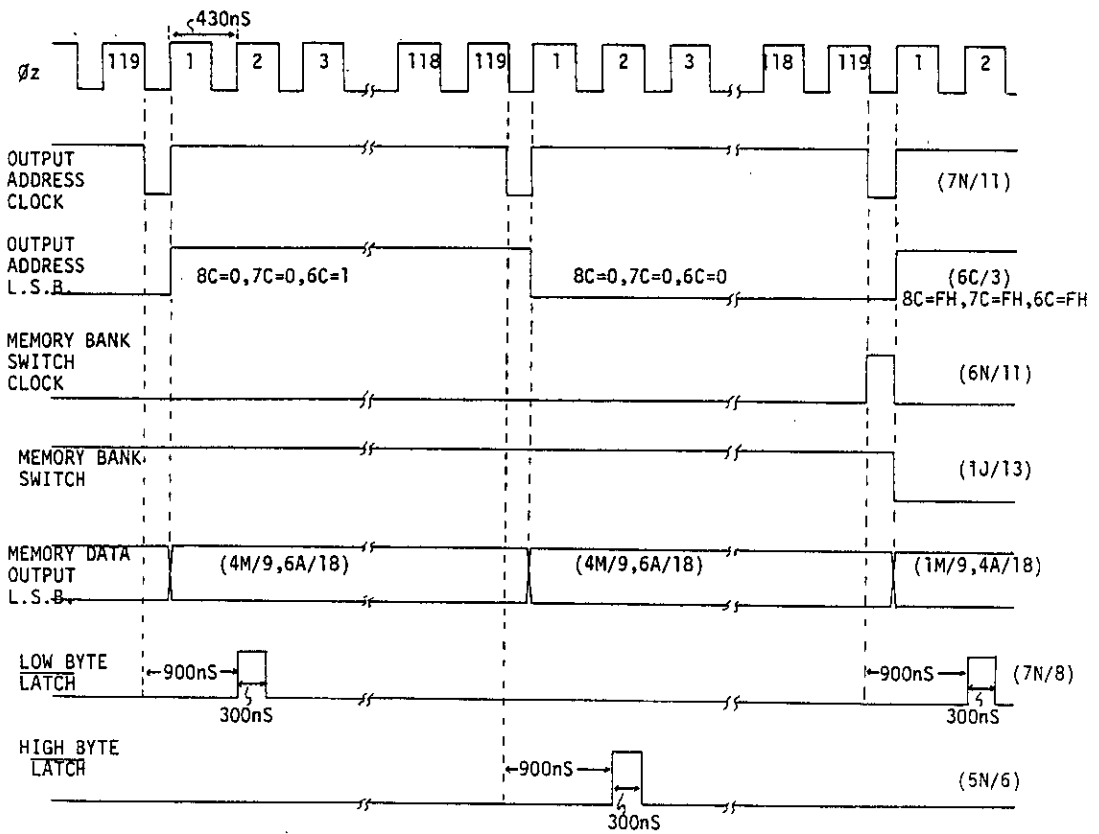


Fig. 8.4(b). Data Output Signals.

Fig 8.4 Data-to-Speech Converter circuit signals.

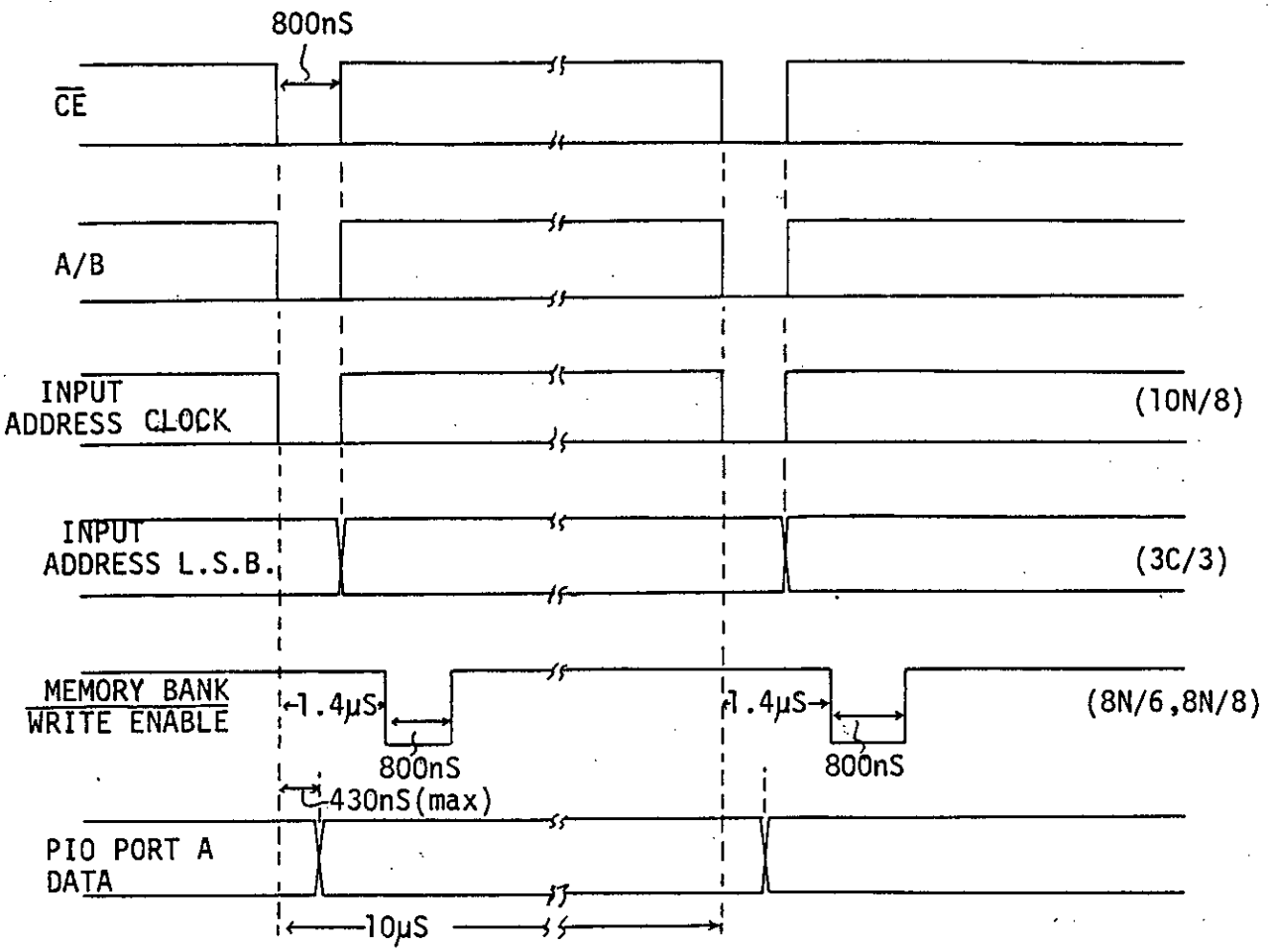


Fig. 8.4(c) Data Input Signals.

Fig. 8.4 Data-to-Speech Converter circuit signals.

network. This is necessary to obviate a small static hazard spike which seems to occur under certain PIO signal conditions. Notice too that this same signal activates the 'latch code' bistable formed by 2N/3 and 2N/11.

The next signal to occur should be that which latches the OAC count code. This is derived from gates 4N/3 and 3N/6, and should be decoded from a command word sent to either port A or B to set them to the data output/input mode respectively.

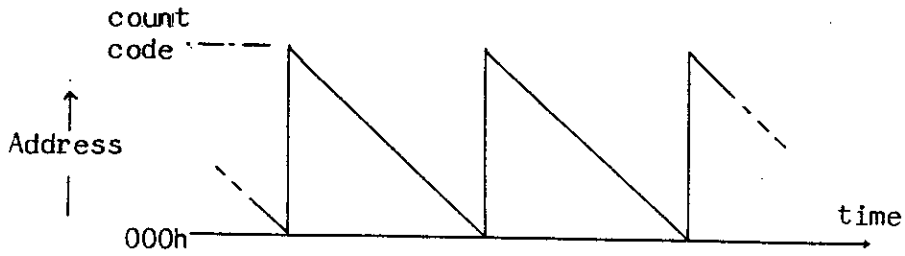
The last signal to occur in this series should be the 'run circuit' pulse into bistable 2N/6, 2N/8. This is derived from gates 4N/11 and 4N/8 and should be decoded from a 'read' command from PIO port B.

Thus, on immediately releasing the circuit to 'run', the following conditions should apply:-

Components	Condition
1L,2L	Latched with output address counter code.
2N/3,2N/6	Outputs in '0' state.
3C,4C,5C	Count down from 000h.
6C,7C,8C	Count down from 00Fh.
1J	Set (1J/13='1').
1M,2M	Write enabled.
3M,4M	Read enabled.
1H,2H,7H,8H	Tri-stated.
3H,4H,5H,6H	Active.
3A,4A	Select PIO port A data.
5A,6A	Select memory data ouput.

B.4.2 SYNCHRONISATION OF DISK READ SPEED TO CIRCUIT OUTPUT RATE

Since the output address counters are in a down-count mode, the address sequence ressembles a saw-tooth waveform:-



It should thus be possible to detect the circuit switching from one memory bank to another by observing the output address waveform. Notice that although in theory only the most significant bit (bit 11) is required to achieve this, bits 5-11 are in fact read. This will facilitate an early knowledge of circuit malfunction (e.g. if the output address is static due to power failure or clock failure, then several successive readings of the counter states will produce the same number, which can be detected by the software.). The memory bank switching waveform is generated by flip-flop 13 whose clock signal is derived from the 'borrow out' signal of counter 8C. Notice that with the output address clock as derived through 1N/3 and 7N/11, the 'borrow out' from 8C will not occur immediately state 000 is reached, but rather just before the rising clock edge causing the counters to change state to FFFh, thereby allowing the last data byte from the memory bank to be successfully latched at the output.

A selection of signals pertaining to the output addressing section of the circuit is shown in Fig. B.4(b), and signals pertaining to the input of fresh data to the on-board memories are shown in Fig. B.4(c).

B.4.3 CIRCUIT MEMORY ADDRESSING

As far as memory addressing is concerned, each memory bank consists to two 2048x8-bit chips,and each individual memory chip only requires

bits 0-10 of the input/output address counters. To achieve full 4096-byte access however, address bit 11 is used to address the chip enable ($\overline{CE}$) pin on each chip (see 6N/3 and 6N/6). Address selection is achieved via hex buffers 1H to 8H. At any time, 1H, 2H, 7H and 8H are in the same mode (tri-stated or active), being the inverse mode to that of buffers 3H, 4H, 5H and 6H.

B.4.4 MEMORY WRITING

The write enable ($\overline{WE}$) signal for each memory bank is derived from monostable 1S. The -ve-going edge of the input address clock, which is derived from PIO signals via gates 10N/3, 10N/6, 10N/8, triggers the first monostable whose -ve-going output edge then triggers the second monostable whose output (1S/13) forms the write enable pulse, which is vectored to the correct memory bank via gates 8N/6 and 8N/8 according to the state of 1J/13, 1J/12 (1J/13='1' puts 1M and 2M into the 'write enable' mode).

Memory bank data input/output is selected via octal buffers 3A-6A, and with 1J/13='1', PIO port A data is selected via 3A and 4A.

B.4.5 MEMORY OUTPUT AND D/A CONVERSION

The memory bank data output is passed to latches 3L-7L. Latch control is achieved via monostable 2S. The -ve-going edge of the output address clock (7N/11) triggers the first monostable, whose output pulse duration should at least equal the output address clock -ve pulse width plus the output address counter settling time + transfer time through e.g. buffers 1H and 2H + data retrieval time for e.g. memory 1M. The second

monostable is subsequently triggered to provide a "follow data" pulse, vectored by the state of bit 0 (6C/3) of the output address counter to either 3L and 4L (bit 0='1') or 5L, 6L and 7L (bit 0='0'). Latches 3L and 4L latch the least significant (l.s.) byte of a 12-bit data word (from odd memory addresses) and latches 5L, 6L and 7L subsequently latch the l.s. byte as held in 3L and 4L and the lower 4 bits of the m.s. byte (from even memory addresses). All 12 bits enter D/A converter 1D whose analog output is buffered via operational amplifier, 1P.

B.5 DATA-TO-SPEECH CONVERTER CONTROLLING SOFTWARE

The following section describes the Zilog program SPEAK, which consists of the following nine modules:- RESET, INSERT, GOONE, GOTWO, BACKUP, UTILTY, LIBARY, ERRORS, NITRTN. Flow charts for the first six are shown in Figs. B.5 to B.9, and a brief description of the salient features of the remaining routines is given overleaf. This set of subroutines was implemented in Z80 Assembler language.

The software assumes that the speech data written on disk is located on sequential tracks and from sector 0 to sector 31. If only one disk drive is in use (DR0), the disk head will read each disk from track 0 inwards to track 77; if two drives are in use (DR0 and DR1) then the reading sequence is:-

Disk No.	from track		to track	Drive
0	0	-in-	77	DR0
1	0	-in-	77	DR1
2	77	-out-	0	DR0
3	77	-out-	0	DR1
4	0	-in-	77	DR0 etc.

Continued on pg.

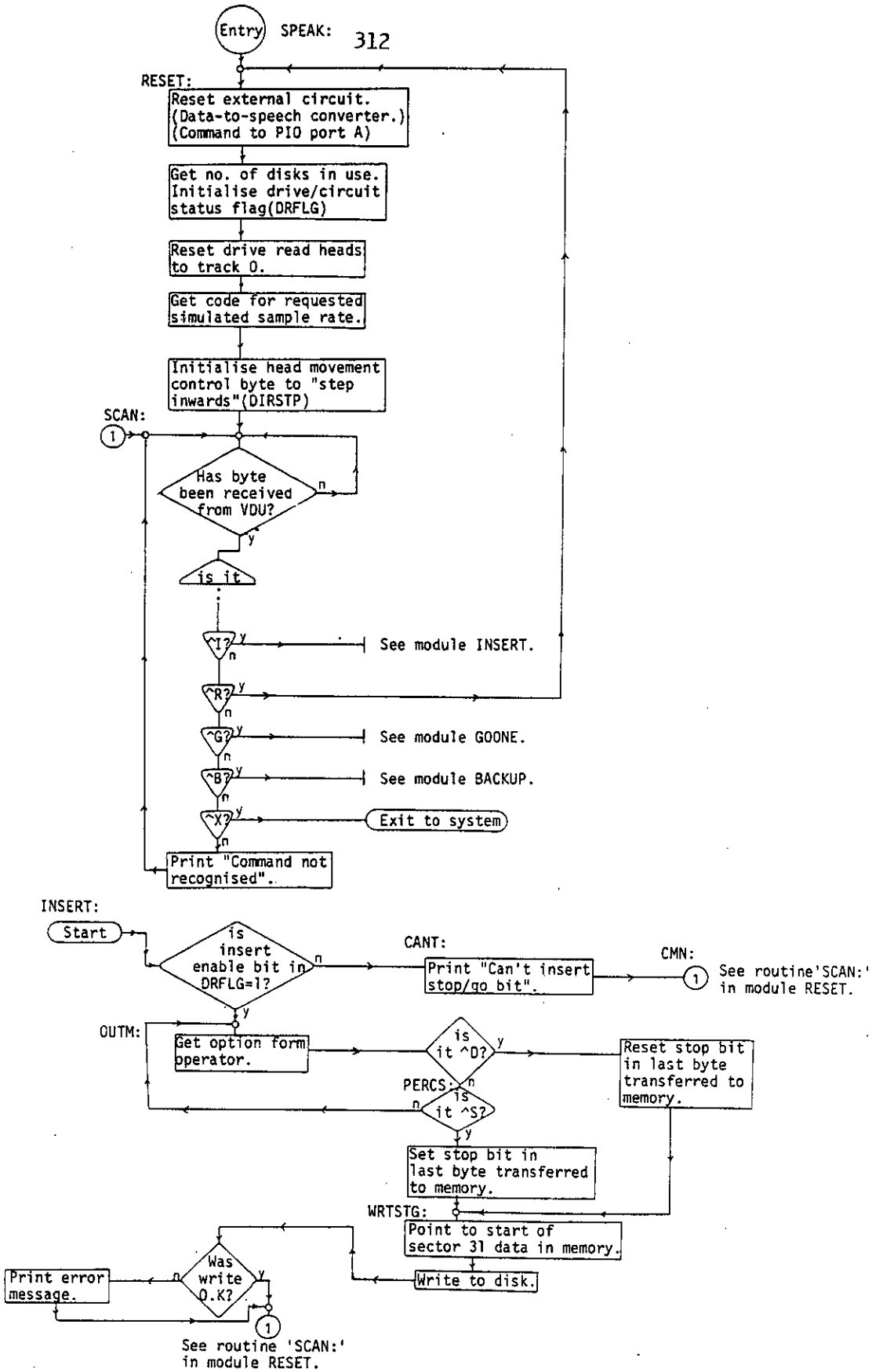


Fig.B.5 Flow Charts for modules RESET and INSERT for the program SPEAK.

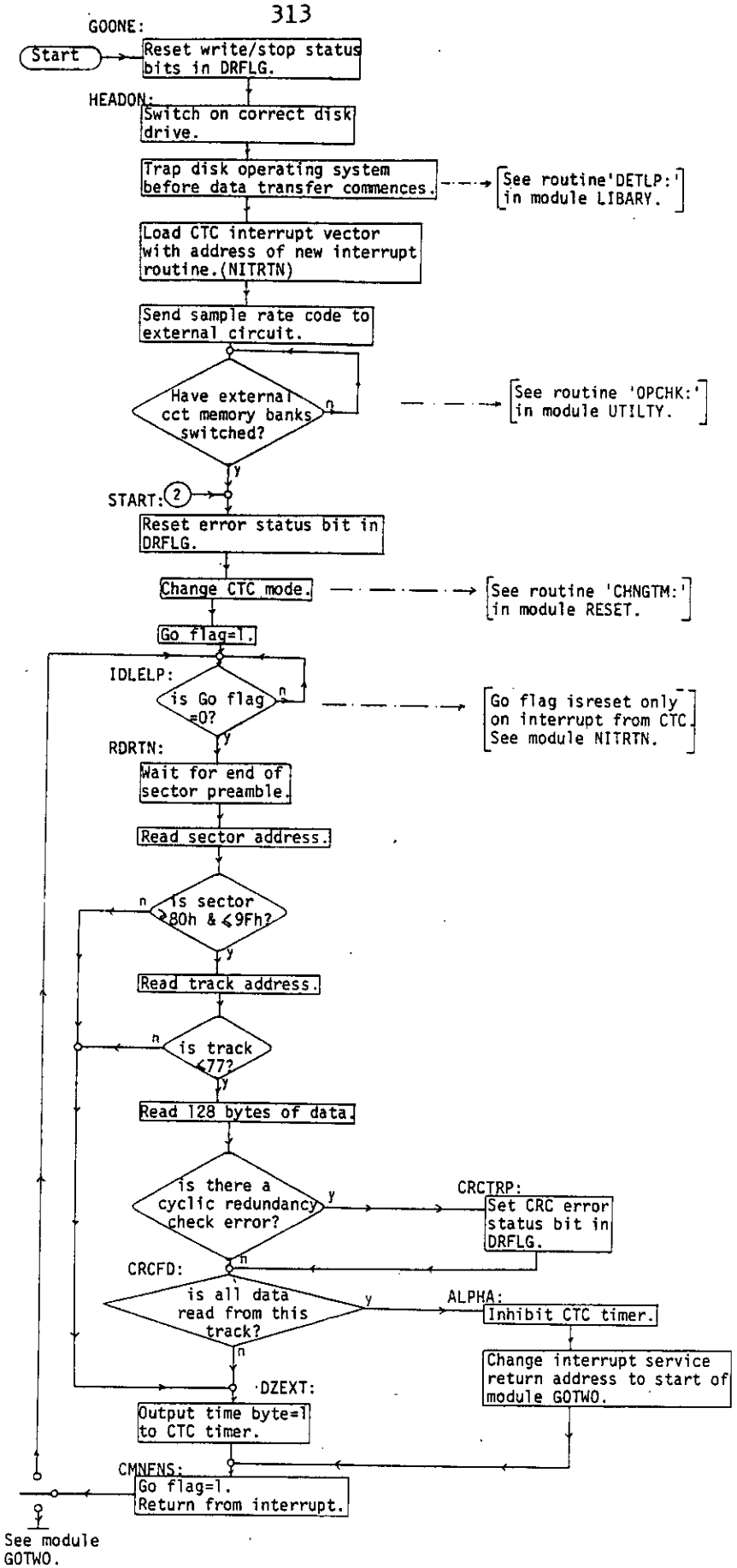


Fig.B.6 Flow Chart for module GOONE
for the program SPEAK.

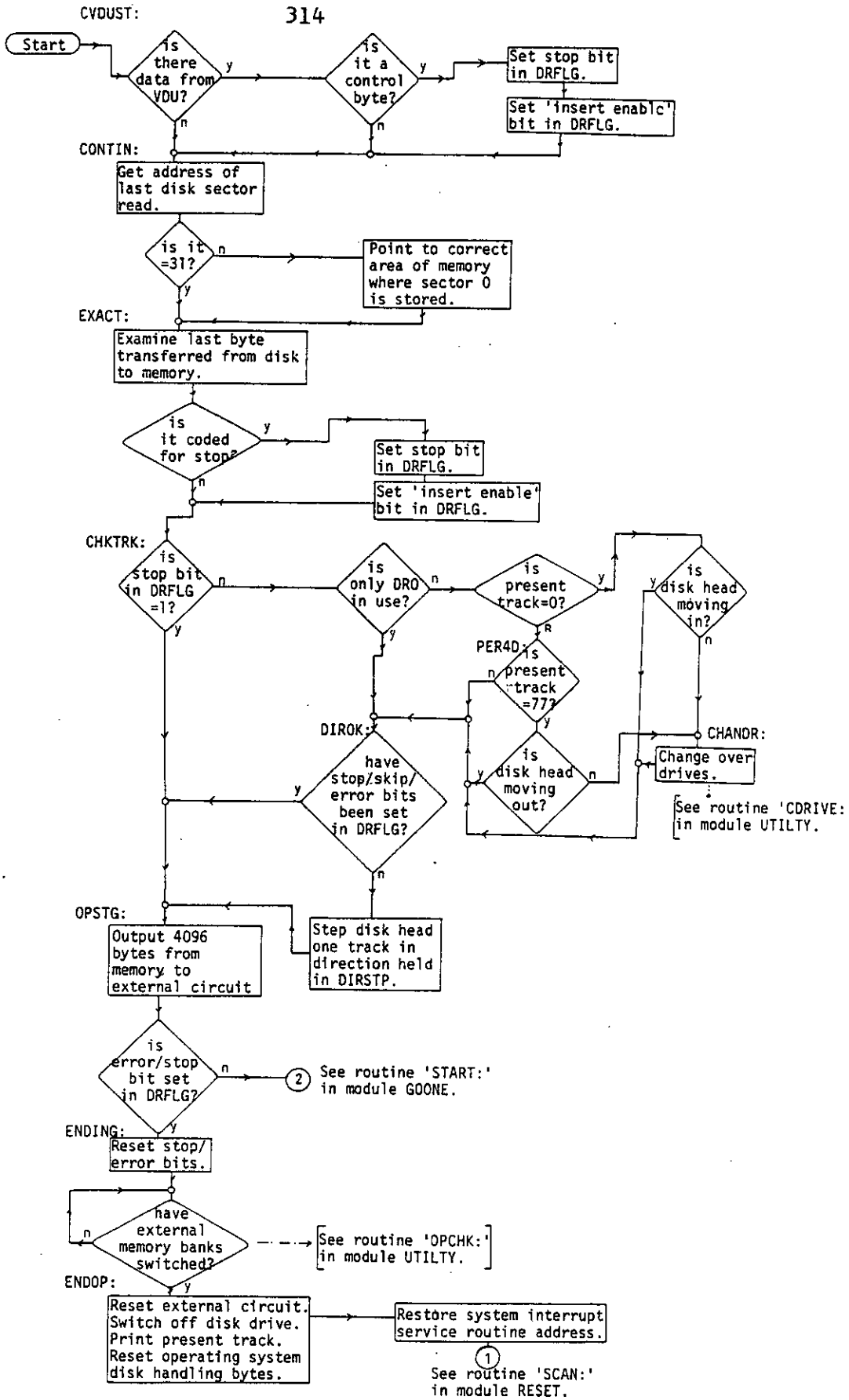


Fig.8.7 Flowchart for module GOTWO
for program SPEAK.

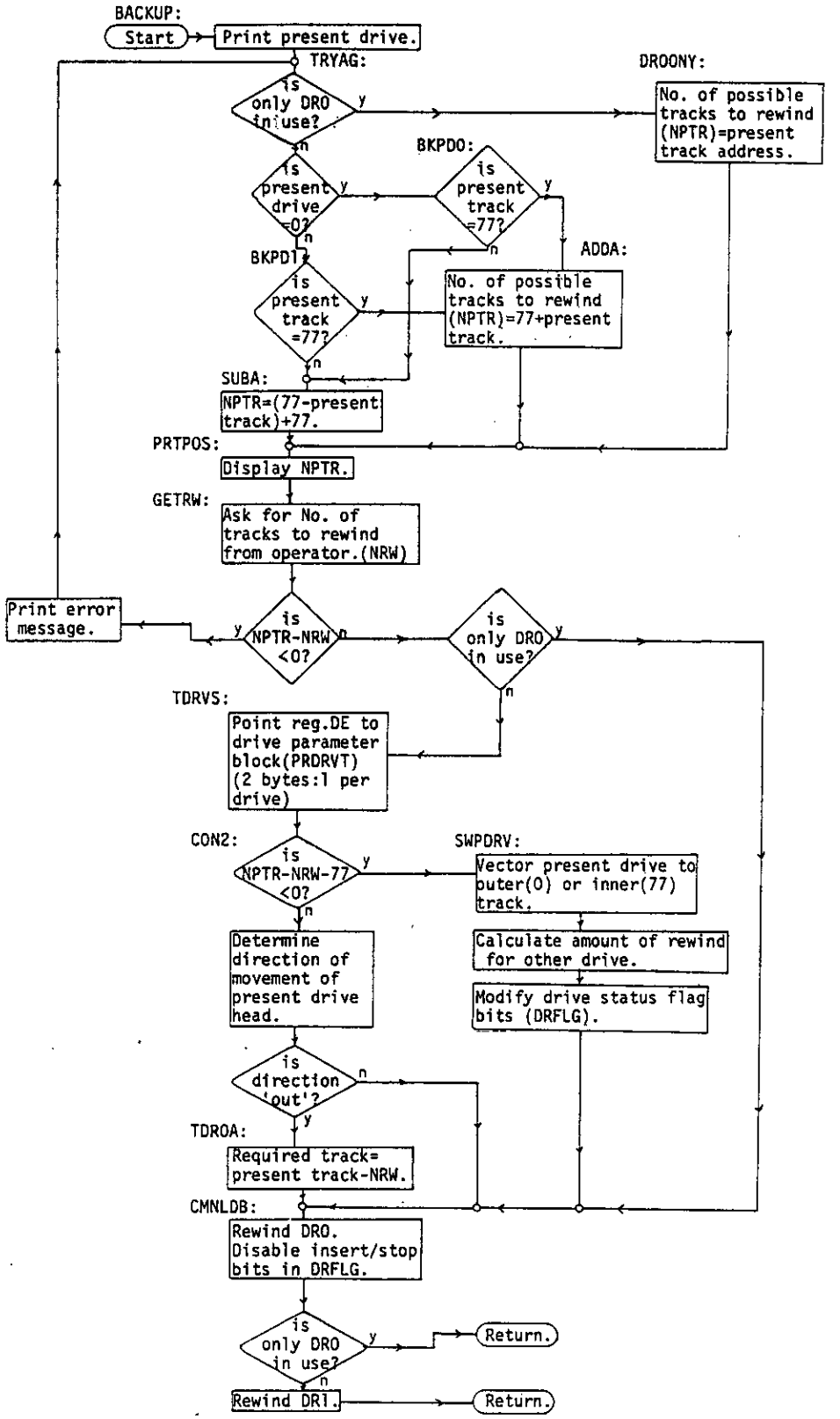


Fig.B.8 Flow Chart for module BACKUP
for the program SPEAK.

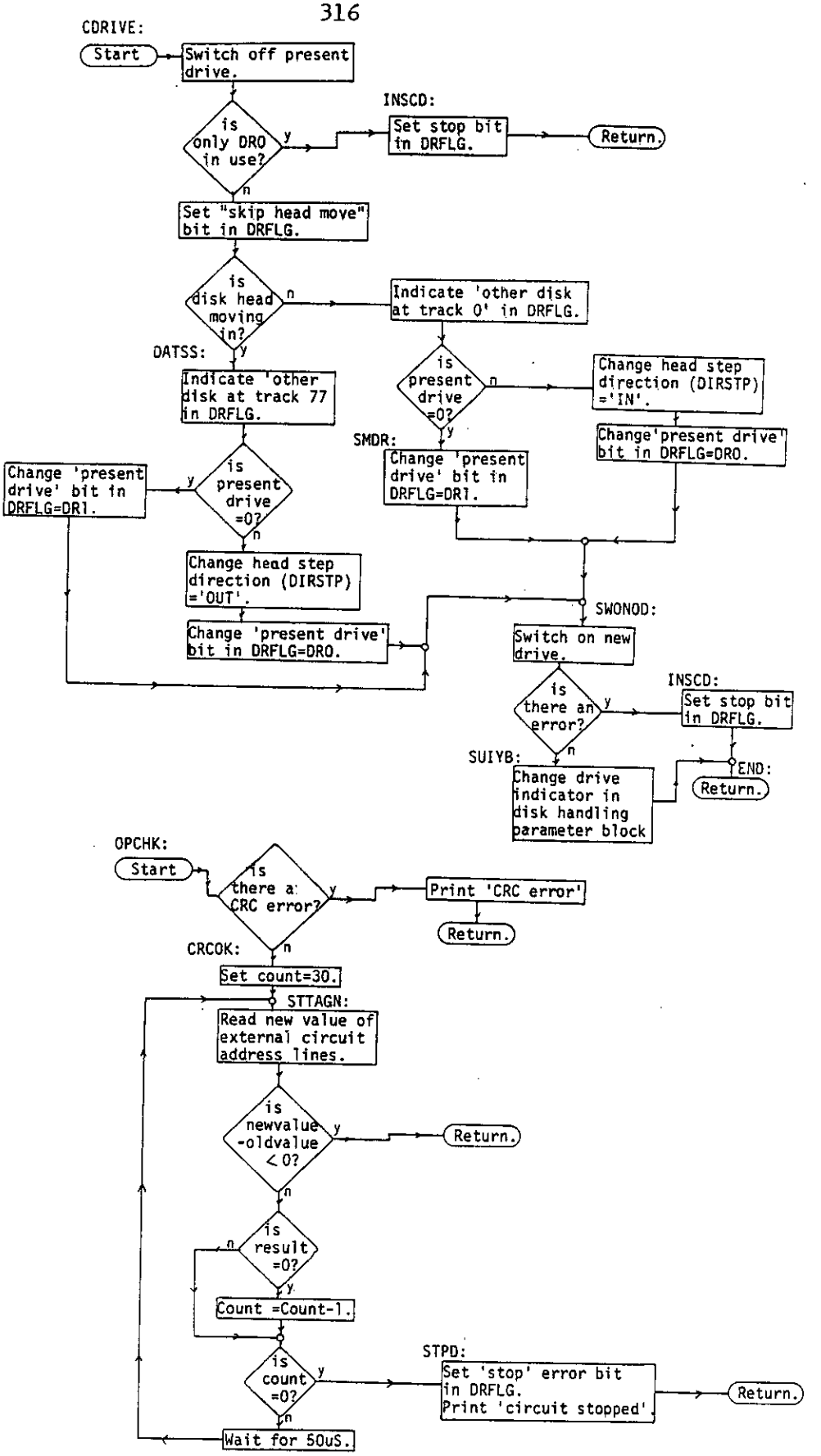


Fig.B.9 Flow Charts for routines 'CDRIVE:' and 'OPCHK:' in module UTILTY for the program SPEAK.

This strategy allows the operator to successfully load alternate drives with new disks with a minimum reload time of 17 seconds.

If the bit 7 of the last byte transferred from any track is set='1', then a software 'stop' is decoded, and output will cease.

Certain disk operating system locations are also used. These are:-

- 0BDFh - system disk handling routine start address;
- 13D7h - head location for drive 0;
- 13D8h - head location for drive 1;
- 13DFh - present (active) drive head location;
- 12ABh - address of present track of active drive;
- 12ADh - address of present sector of active drive;
- 13E0h - contains address for start of interrupt vector routine;
- 12AFh - contains address, accessed by vector routine, pointing to start of service routine.

In addition, the following system ports are used:-

Port	Description
CFh	Disk data
D0h	Disk control
D1h	Disk status
D4h	CTC Channel 0
D8h	PIO A data
D9h	PIO B data
DAh	PIO A command
DBh	PIO B command
DEh	VDU data
DFh	VDU status

Several user-defined library functions are used, which are not detailed here but are breifly described in relation to Figs. B.5-B.9.

Name	Function
IFCP:	input text from VDU to program. Convert control characters to visible format;
DRVOFF:	switch off all presently active disk drives;
DETLP:	detect and inhibit disk operating system as it is about to transfer data from disk to memory;

DECHEX: convert ASCII decimal from VDU into hexadecimal;
HEXDEC: convert hex. to ASCII decimal string;

All the above routines are contained in the module LIBARY:.

CHNGTM: change CTC mode from 'time' to 'count';

This routine is to be found in the module RESET:.

ERRORS: (module) decode system disk error code;
NITRTN: (module) vector execution to interrupt service routine.

The general program strategy is that firstly, in module RESET:, the external circuit is put into the 'reset' mode. The number of disk drives in use is also requested, together with the sample rate code for this session. Various options are then open to the user according to the control option typed at the consol. A ^G will start speech conversion from the present track (see modules GOONE and GOTWO); a ^I will either insert a 'stop' marker or delete an existing one (see module INSERT); a ^B will rewind the drive heads by x tracks (see module BACKUP); a ^R will reset the external circuit and drive heads ready for another session (see module RESET); a ^X will return the user to the Zilog system.

All of these modules check at some time or other the status bits of a user-defined byte called DRFLG. This reflects information about the disk drive/external converter circuit status. The bits are as follows:-

Bit	Meaning
0	head direction, 0='in', 1='out';
1	present active drive, 0=DR0, 1=DR1;
2	head location of other drive, 0=track 0, 1=track 77;
3	insert/delete stop marker enable? 0=disabled, 1=enabled;
4	software stop? 0=continue, 1=stop;
5	not used
6	CRC error/Circuit error? 0=o.k., 1=error;
7	only DR0 active? 0=false, 1=true;

The software synchronises the speed at which the disks are read to the external circuit simulated sample rate. This is achieved by detecting when memory banks switch over and, only when this happens, allowing the track to be read then transferred. (See routine OPCHK: in module UTILTY, Fig. B.9).

The maximum data rate achieved so far has been 18.6kbytes/sec at the output of the external memory banks, as opposed to the earlier theoretical value of 19.598kbytes/sec. The difference occurs since the latter figure is calculated assuming that, every time the software is ready to read a track, a sector index marker is just present. This will not always be the case, and it is possible that the sector marker may just be missed (time to traverse one sector=5.2mS). In particular, it is possible that the sector 0 index hole is the first marker to present itself which, as explained earlier, occurs in the middle of the sector 31 data stream and must be ignored if detected, in which case the delay to start of data transfer is $5.2\times10^{-3} + (5.2/2)\times10^{-3}$ sec, assuming this index hole occurs in the centre of a normal sector span.

Thus, the maximum simulated sample rate now possible is given by $4096/(167\times10^{-3} + 42\times10^{-3} + 7.8\times10^{-3}) = 18.89$kbytes/sec, which is very close to the maximum rate observed to date. A list of sample rate codes for real-time and slowed speech output is contained in Fig. B.10.

When data is first read off the disk into Zilog memory, it is done so from the first good sector available, whose address is likely to be at random. Subsequent sectors are all read sequentially into memory, therefore only the last sector number transferred is known. (see variable SCTADD in module GOONE). However, providing the 4096 bytes from any track are always read into the same memory space, this is not a problem, and a simple calculation (see routine CONTIN: in module GOTWO) points

to the position of sector 0 data in the Zilog memory for any track. In the program SPEAK:, the data read from any track will lie between memory addresses 4000h and 4FFFh.

Sample Freq.	Nearest Code	Actual Freq.	Percnt Diff	Speed Rdctn	O/P Time
5000	233	4991	-0.2	1	32.0
5500	211	5511	0.2	1	29.0
6000	194	5994	-0.1	1	26.7
6500	179	6496	-0.1	1	24.6
7000	166	7005	0.1	1	22.8
7500	155	7502	0.0	1	21.3
8000	145	8019	0.2	1	19.9
8500	137	8488	-0.1	1	18.8
9000	129	9014	0.2	1	17.7
9500	245	9492	-0.1	2	16.8
10000	233	9981	-0.2	2	16.0
10500	221	10523	0.2	2	15.2
11000	211	11022	0.2	2	14.5
11500	202	11513	0.1	2	13.9
12000	194	11988	-0.1	2	13.3
12500	186	12503	0.0	2	12.8
13000	179	12992	-0.1	2	12.3
13500	172	13521	0.2	2	11.8
14000	166	14010	0.1	2	11.4
14500	160	14535	0.2	2	11.0
15000	155	15004	0.0	2	10.6
15500	150	15504	0.0	2	10.3
16000	145	16038	0.2	2	10.0
16500	141	16493	-0.0	2	9.7
17000	137	16975	-0.1	2	9.4
17500	133	17486	-0.1	2	9.1
18000	129	18028	0.2	2	8.9
18500	126	18457	-0.2	2	8.7
19000	245	18984	-0.1	4	8.4

Fig. B.10

Sample Freq.	Nearest Code	Actual Freq.	Percnt Diff	Speed Rdctn	O/P Time
19500	239	19461	-0.2	4	8.2
20000	233	19962	-0.2	4	8.0
20500	227	20490	-0.1	4	7.8
21000	221	21046	0.2	4	7.6
21500	216	21533	0.2	4	7.4
22000	211	22043	0.2	4	7.2
22500	207	22469	-0.1	4	7.1
23000	202	23026	0.1	4	6.9
23500	198	23491	-0.0	4	6.8
24000	194	23975	-0.1	4	6.7
24500	190	24480	-0.1	4	6.5
25000	186	25006	0.0	4	6.4
25500	182	25556	0.2	4	6.3
26000	179	25984	-0.1	4	6.1
26500	176	26427	-0.3	4	6.0
27000	172	27042	0.2	4	5.9
27500	169	27522	0.1	4	5.8
28000	166	28019	0.1	4	5.7
28500	163	28535	0.1	4	5.6
29000	160	29070	0.2	4	5.5
29500	158	29438	-0.2	4	5.4
30000	155	30008	0.0	4	5.3
30500	152	30600	0.3	4	5.2
31000	150	31008	0.0	4	5.2
31500	148	31427	-0.2	4	5.1
32000	145	32077	0.2	4	5.0
32500	143	32526	0.1	4	4.9
33000	141	32987	-0.0	4	4.8
33500	139	33462	-0.1	4	4.8

Sample Freq.	Nearest Code	Actual Freq.	Percnt Diff	Speed Rdctn	O/P Time
34000	137	33950	-0.1	4	4.7
34500	135	34453	-0.1	4	4.6
35000	133	34971	-0.1	4	4.6
35500	131	35505	0.0	4	4.5
36000	129	36056	0.2	4	4.4
36500	127	36623	0.3	4	4.4
37000	126	36914	-0.2	4	4.3
37500	248	37509	0.0	8	4.3
38000	245	37969	-0.1	8	4.2
38500	242	38439	-0.2	8	4.2
39000	239	38922	-0.2	8	4.1
39500	236	39417	-0.2	8	4.1
40000	233	39924	-0.2	8	4.0
40500	230	40445	-0.1	8	3.9
41000	227	40979	-0.1	8	3.9
41500	224	41528	0.1	8	3.8
42000	221	42092	0.2	8	3.8
42500	219	42476	-0.1	8	3.8
43000	216	43066	0.2	8	3.7
43500	214	43469	-0.1	8	3.7
44000	211	44087	0.2	8	3.6
44500	209	44509	0.0	8	3.6
45000	207	44939	-0.1	8	3.6
45500	204	45600	0.2	8	3.5
46000	202	46051	0.1	8	3.5
46500	200	46512	0.0	8	3.4
47000	198	46981	-0.0	8	3.4
47500	196	47461	-0.1	8	3.4
48000	194	47950	-0.1	8	3.3

Appendix C

Contents of
Demonstration Cassette.

Contents of
Demonstration Cassette.

Rec. A1 Referenced in section 1.4, pg.7.

Some examples of helium speech recorded during a working dive using heliox respiratory mixtures. The speech heard here is spoken at a depth of 300ft below sea level.

Rec. A2 Referenced in section 3.2.1, pg.68.

A2(a) List 4 of Fig.3.1 spoken by the subject in air at normal surface temperatures and pressures.

A2(b) List 4 of Fig.3.1 spoken by the subject at a depth of 100ft breathing a gas mixture of 91.6%He and 8.4%O_2 at a pressure of 4bar. See Fig.3.3 for further details.

Rec. A3 Referenced in section 3.5.1, pg.128.

An example of ambient noise from a carbon dioxide gas filter situated within the diving chamber.

Rec. A4 Referenced in section 4.1.3, pg.159.

The Rainbow Passage, which is list 5 of Fig.3.1, spoken by the subject in air at normal surface temperatures and pressures.

Rec. A5 Referenced in section 4.1.3, pg.159.

The Rainbow Passage spoken by the subject at a depth of 100ft breathing heliox.

Rec. A6 Referenced in section 4.1.3, pg.159.

A6(a) Part of the Rainbow Passage of rec. A5 unscrambled by digital computer simulation of the timedomain-based unscrambler system.

A6(b) Part of the Rainbow Passage of rec. A5 unscrambled by the hardware implementation of the timedomain-based unscrambler device.

Rec. A7 Referenced in section 4.2.6, pg.187.

The Rainbow Passage of rec. A5 unscrambled by the digital computer simulation of the unscrambler system based on the short-time Fourier transform.

Rec. A8 Referenced in section 4.2.6, pg.188.

The Rainbow Passage spoken by the subject at a depth of 300ft breathing a gas mixture of 96.7%He and 3.3%O_2 at a pressure of 10bar. This speech has been unscrambled by the digital computer simulation of the unscrambler system based on the short-time Fourier transform.

Rec. A9 Referenced in section 4.2.6, pg.189.

A signal synthesised by digital computer and representing the impulse excitation, at a rate of 150Hz, of a LTI filter system whose frequency response exhibits resonances at 1, 3 and 5kHz with bandwidths 80, 100 and 150Hz respectively and with relative peak power ratios of 0dB, -6dB and -9dB respectively.

Rec. A10 Referenced in section 4.2.6, pg.189.

A signal synthesised by digital computer and representing the impulse excitation, at a rate of 150Hz, of a LTI filter system whose frequency response is compressed by a factor of 2 compared to that of rec. A9. Note, however, that relative peak power ratios have been conserved.

Rec. A11 Referenced in section 4.2.6, pg.193.

The synthetic signal of rec. A9 processed by digital computer simulation of the short-time Fourier transform unscrambling technique in an attempt to reproduce the signal heard in rec. A10.

Rec. A12 Referenced in section 4.3, pg.196.

A12(a) The Rainbow Passage spoken by the subject at a depth of 100ft unscrambled by the simulated timedomain technique. (This is the same as rec. A6(a)).

A12(b) The Rainbow Passage spoken at 100ft unscrambled by the modified timedomain technique.

A12(c) The Rainbow Passage spoken at a depth of 300ft unscrambled by the modified timedomain technique.

Rec. A13 Referenced in section 5.2.3, pg.242.

A13(a) A signal synthesised by digital computer and representing the impulse excitation, at a rate of 150Hz, of a LTI filter system whose frequency response is compressed by a factor of 2 compared to that of rec. A9. (This is the same as rec. A10).

A13(b) The synthetic signal of rec. A9 processed by digital computer simulation of the residually excited linear predictive coding technique in an attempt to reproduce the signal heard in rec. A13(a).

Rec. A14 Referenced in section 5.2.5, pg.250.

The Rainbow Passage spoken at 100ft unscrambled by the digital computer simulation of the RELPC unscrambler system.