THE VOCAL TRACT AREAGRAM: A NEW TECHNIQUE FOR DISPLAYING SPEECH

by D.G. NICHOL B.Sc.(Hons), Ph.D.(Tas)

Department of Electrical Engineering
University of Adelaide

December 1982

A Thesis submitted for the Degree of Master of Engineering Science

This thesis embodies the results of supervised project work making up all of the work for the degree.

TABLE OF CONTENTS

SUMMARY

The possibility of making speech visible has attracted research workers since the scientific investigation of speech began. The reasons for this are not hard to find and are basically related to the synoptic view which is obtained by having a visual 'hard copy' of an utterance. For many years now speech spectrograms, also known as 'Sonagrams' or 'Voice Prints', have been extensively used as the primary method of visually displaying speech. These have found wide acceptance in fields as diverse as phonetics and linguistics, medical diagnosis and, more recently, jurisprudence. Over the last decade the application of the techniques of digital signal processing to the modelling of the voice production process has revolutionised speech research. From a synthesis viewpoint reasonably realistic speech can now be produced from a programmable micro 'chip' and linear prediction type analysis has led to a significant improvement in the compressibility of speech. The present study arises from the fact that, almost as a by-product of linear prediction analysis, it is possible to obtain an estimate of vocal tract shape from the raw speech data. Efficient algorithms exist to do this in near real time on general purpose computers. Accordingly a new method to display speech is possible and is the subject of this study. The proposed method produces an 'areagram' which displays, in a format similar to the spectrogram, the varying shape of the vocal tract as a function of time. Instead of frequency as the vertical axis and spectral density as the grey level the areagram has distance along the vocal tract and cross-sectional areas respectively.

The areagram is proposed as a complementary display to the spectrogram and not as a replacement. Features of the spectrogram such as formants and loci are obviously extremely significant in speech analysis but they are difficult to relate to articulatory processes. The areagram will provide just this information and as it can be plotted on the same scale as the spectrogram a direct comparison is very easy to make.

In the following study various aspects of areagram production and display are examined. These include the choice of time series window length and shape, how to handle random noises and breakdown of the all-pole model, the use of picture processing techniques to interpolate data, to enhance structure and how to assign grey levels or colour to the images. These latter, picture processing techniques, are shown to apply equally well to the spectrogram. Finally a comparison of the spectrogram and areagram from the viewpoint of their usefulness for visual recognition is made.

STATEMENT OF ORIGINALITY

I declare that this thesis does not contain any material which has been accepted for the award of any other degree or diploma of any University, and, that to the best of my knowledge, this thesis does not contain any material published or written by another person except where due reference is made in the body of the thesis.

(D.G. NICHOL)

20 December 1982

Parts of this work has been published or presented in the following papers:

1.	D.G. Nichol and R.E. Bogner	"A New Technique of Speech Display: The Vocal Tract Area Function Picture", presented to IREE 16th International Convention, Melbourne, Australia, 1977.
2.	D.G. Nichol and R.E. Bogner	"Quasi-Periodic Instability in a Linear Prediction Analysis of Voiced Speech", IEEE Trans. Acoust. Speech Signal Processing, Vol.ASSP-26, Page 210-216, June 1978
3.	D.G. Nichol	"Displaying Speech as Vocal Tract Area Function Pictures", Paper GGG8, Presented to Joint Meeting

of Acoustical Society of America and Acoustical Society of Japan, Honolulu, USA, 1978

4. D.G. Nichol and R.E. Bogner — "A Comparison of Spectrograms and Vocal Tract Area Functions for Displaying Speech", Paper A1-9.3, presented to Tenth International Congress on Acoustics, Sydney, Australia, 1980

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor, Professor R.E. Bogner, for his help and encouragement at all stages of this study.

I would also like to to thank my employer,the Department of Defence Support, for allowing me to persue the study and for providing access to the computer facilities necessary for its completion.

CHAPTER 1

INTRODUCTION AND THESIS OUTLINE

Being a temporal phenomenon speech is essentially sequential in both transmission and reception. From the viewpoint of a student of speech the sequential nature of reception and analysis is very limiting. For example, it is difficult to compare in detail many different repetitions of the same utterance, due to the limitations of the human aural memory. The concept of 'making speech visible' arouse because the semi-parallel nature of human visual processing (potentially) enables detailed comparison of speech segments. Another potential advantage of visible speech is that photographic images are much more convenient for display and publication than are, for example, magnetic tapes.

Having decided upon the desirability of displaying speech visually various techniques for achieving this had to be assessed. Plotting sound pressure as a function of time is an obvious method and easily achieved with a microphone, amplifier and pen-recorder. Whilst useful for comparing temporal structure and utterance duration it was soon realized that the same utterance from different speakers could appear very dissimilar in this time series display. A frequency domain display was found to be much more consistent in this regard and commercial systems to produce short term spectral analysis became common in the 1950's. Rather than display plots of the short term spectral estimates these devices used intensity modulation to show spectral power as a grey scale with time as the horizontal axis and frequency as the vertical. Thus was born the speech spectrogram, also known as a 'Sonagram' or 'Voice Print'[1]. Although beginning to be replaced by digital technology, using the FFT (Fast Fourier Transform) algorithm, many of these analog machines are still in use. The spectrogram, whether digital or analog, is still the universally used visible speech display.

The possibility of deriving the shape of the vocal tract directly from the received speech was due to two simultaneous developments. These were the development of an acoustic model of the vocal tract as a series of tubes of constant cross-sectional area [2,3] and the parallel development of a linear prediction model for speech prediction [4,5]. Tying these two approaches together enabled the direct estimation of vocal tract areas. The proposal to be examined in the present study is to use one such formulation (due to Wakita [6]) to derive vocal tract areas and to display these as 'pictures' with the area plotted as a grey scale. The other parameters, time and distance along the tract, are plotted on the x and y axes respectively. The resulting picture is an 'areagram'.

It is not intended that the areagram replace the spectrogram as the standard form of visible speech. Rather the areagram will provide a complement to the spectrogram and the two should be used together. The spectrogram will provide information on signal intensity and vocal tract resonances ('formants') whilst the areagram will provide simultaneous information on the variation of the vocal tract.

There are many areas of the study of speech that have a requirement for such a type of display. These include phonetic and linguistic studies, speech analysis and synthesis research, medical diagnosis of speech problems and the teaching of the deaf.

The thesis contents are now outlined. In chapter 2 the concept of the areagram is presented and the problems associated with their production are briefly discussed. An outline of the production process is then given. More details are given in the appendix and the relevant sections of subsequent chapters. In chapters 3 and 4 some of the deficencies of the model used to obtain area functions are addressed. Effects considered in chapter 3 include window lengths and shapes and the interaction of these with glottal waveform periodicities. The effects of random noises,

silences and non-voiced speech are considered in chapter 4. The all pole model used in linear prediction breaks down for non-voiced speech and techniques to detect and compensate for this are discussed. In chapter 5 techniques for enhancing the visual display of the articulatory data are described. These include interpolation, various two dimensional filters and techniques for grey level and colour assignment. To test the effectiveness of areagrams in visual speech classification an experiment using a database of test utterances and human observers is described in chapter 6. The areagram and spectrogram, both grey-level and colour, are compared from this viewpoint. Finally, in chapter 7, the overall usefulness of the areagram is assessed and suggestions for further work made.

CHAPTER 2

GENERATING AREAGRAMS FROM THE SPEECH TIME SERIES

To produce estimates of vocal tract shape the model of Wakita [6] is used. This is a so-called autocorrelation non-pitch synchronous algorithm derived by using an inverse filter model of the vocal tract. Due to the fact that the autocorrelation matrix is Toeplitz a fast inversion algorithm for it exists [7]. It is this aspect, coupled with the fact that it is fairly insensitive to pitch impulse position (but see chapter 3), that makes computation of the vocal tract area function (VTAF) by Wakita's method so fast. The inverse digital filter is all-zero and the inverse filtering technique (essentially) is to adjust the position of the zeros until the output of this filter, with the received speech signal as input, has a flat spectrum. This flat spectrum is posited (e.g.[8]) to be a good approximation to the glottal excitation, either impulse train or white noise. This inverse all-zero filter is equivalent to an all-pole filtering of the glottal waveform. In turn this is equivalent to linear prediction analysis [9,10] and it is usual to refer to all these classes of digital models as 'all-pole' or 'linear prediction' irrespective of their actual formulation.

2.1 THE BASIC AREAGRAM

Suppose the speech waveform is sampled at f_s points per second. After a buffer or 'window' of length N_w samples is obtained a VTAF estimate can be made. The cross-sectional area is estimated at N_f+1 equi-spaced points on the vocal tract where N_f is the number of filter coefficients used in the model. This is approximately 16 for a sampling frequency f_s =16 kHz. It should be noted that an assumption has to be made about either the glottal or the labial areas in linear prediction models. In this study a constant

glottal area is assumed. At intervals of N_c samples a new estimate of the VTAF is made. Usually $N_c < N_w$ so window overlap occurs. Thus a sequence of T VTAF estimates is made as a function of time. These may be thought of as a matrix $\underset{\sim}{A}$ given by:

$$\underset{\sim}{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1T} \\ a_{21} & a_{22} & & & \\ a_{31} & & & & \\ \vdots & & & & \\ a_{N_f 1} & & & & a_{N_f T} \end{bmatrix} \qquad (2.1)$$

where a_{ij} is the area estimated at position i for the jth window of data.

The proposal considered in the present study is to plot $\underset{\sim}{A}$ as a grey-level (or colour) image. The element a_{ij} is represented as a grey dot of intensity proportional to the value of a_{ij} at position (j,i) on a cathode ray screen. The glottis is plotted at the bottom of each subpicture and the lips at the top.

Figure 2.1 shows the results of plotting a matrix A obtained from analysis of the spoken digits, 'one' and 'two' etc. There are a number of obvious deficiences in this display. These include:

(1) Image is too small and thin to see much structure.

(2) 'Pulsating' areas can be seen when the speech is nearly stationary (e.g. final phoneme in 'two').

(3) Very wide lip areas are seen where they should not occur. Examples are the /t/ of 'two' and for some points between the utterances.

(4) grey scale is too dark.

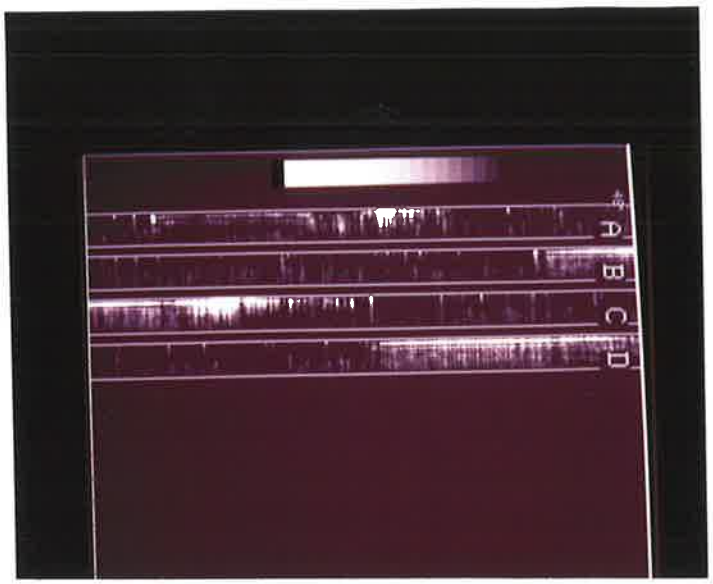


FIG.2.1 Raw areagram of spoken digits 'one - two - three'. Each subpicture (A,B,C,D) is of 1 second duration. Lips are at top.

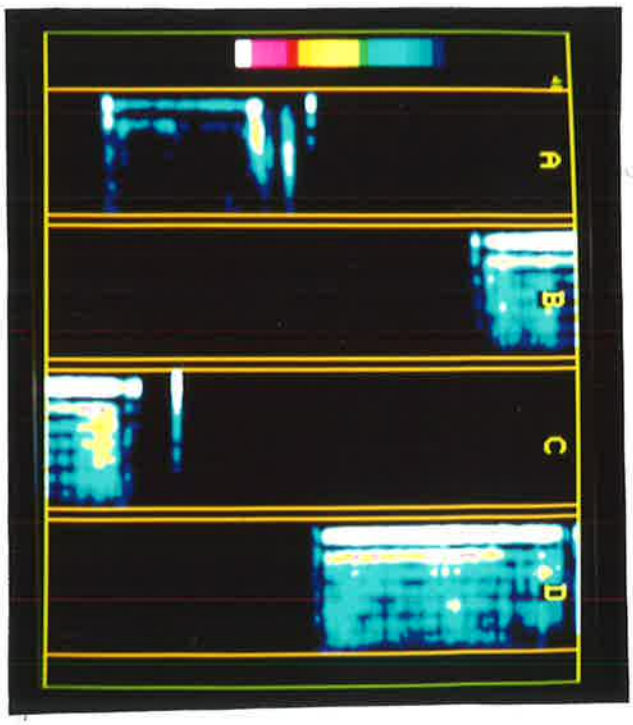


FIG.2.2 Same data as figure 2.1 but reprocessed in colour, with interpolation, fricative compensation and adequate windowing.

Figure 2.2 shows the same data reprocessed and shown as a colour image. Much of the rest of this thesis is concerned with seeing how to transform figure 2.1 to figure 2.2. The points raised above, and others ,will all be examined in detail in subsequent chapters.

Interpretation of figure 2.2 in articulatory terms is quite straightforward. For example the steady fairly long segments seen in subpictures C and D are the vowel parts of 'two' and 'three' respectively. For the phoneme /u/ of 'two' the central part of the vocal tract is is open with the upper (labial) and lower (velar) parts constricted. This corresponds to the known articulation of this vowel. The phoneme /i/ of 'three' is quite long, stable and closed in the alveolar region. This is also as expected. It seems then that this display method enables articulatory information, at least for vowels, to be easily observed.

Figure 2.3 gives an overall view of the software package produced for this study. This was initially written in FORTRAN to run on a mini-computer system but was rewritten and greatly expanded in PL/I when processing was transfered to a mainframe system. Most aspects of the package shown in figure 2.3 are covered in the relevant chapters. The data acquisition and time series editing modules are described in Appendix A.

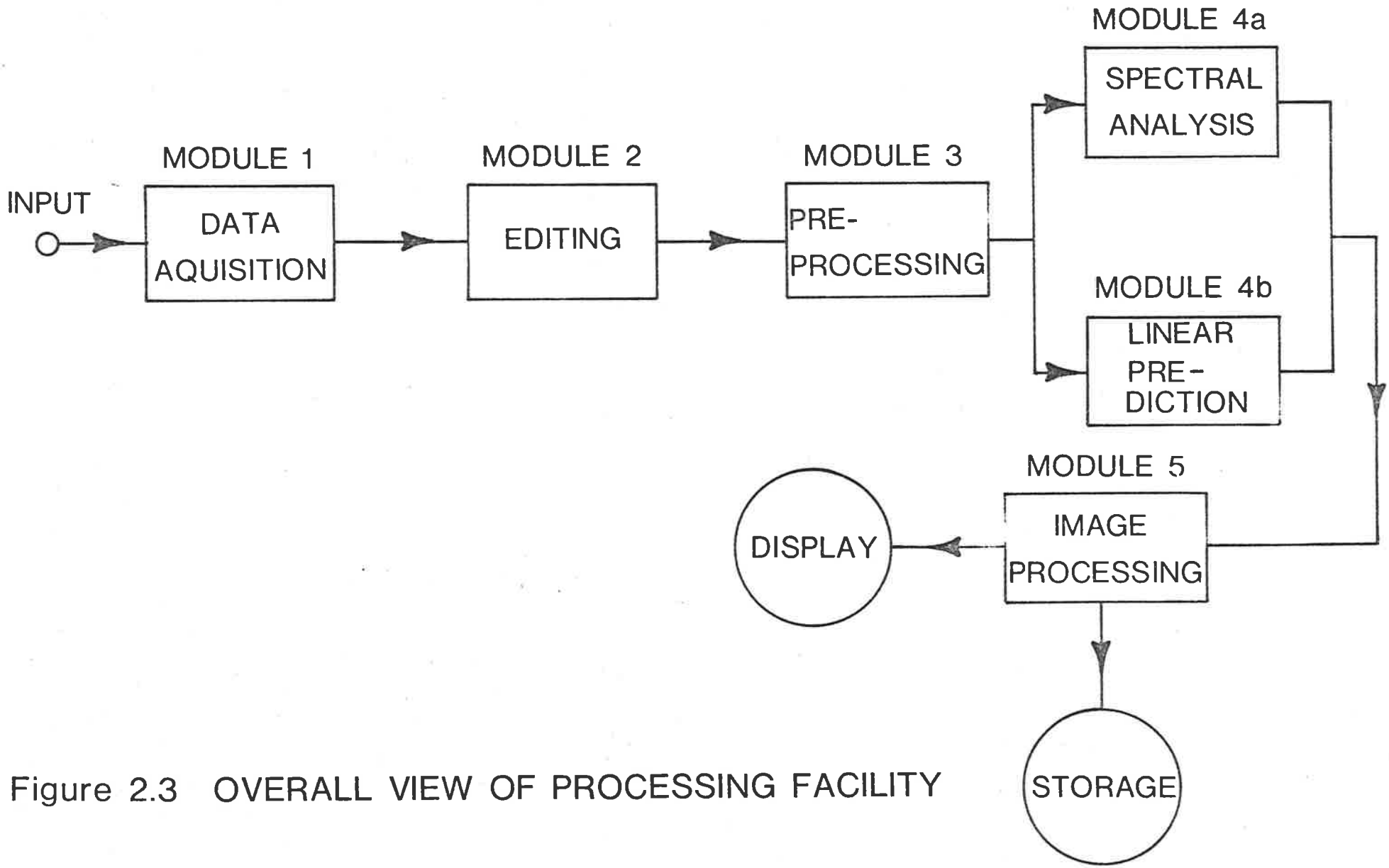


Figure 2.3 OVERALL VIEW OF PROCESSING FACILITY

CHAPTER 3

CORRECTING THE AREAGRAM FOR TIME SERIES 'WINDOW' EFFECTS

The phenomenon discussed in this chapter illustrates the usefulness of the areagram in providing a synoptic view of speech; in this case enabling the detection of a pathological condition caused by a 'beating' between the time series window and the computation interval. This effect can cause very serious errors in vocal tract estimation but can be easily suppressed as discussed below. It is interesting to note that this error was first noticed in an areagram of a female speaker and had not been noticed in earlier areagrams of the same utterance by male speakers. This was simply due to the near coincidence of the male pitch period and the sampling interval. However gross errors were present in the male areagrams but had not been noticed due to the longer period of pulsation for male speech. This observation has some cautionary implications for those speech workers who continually avoid using female speakers

Much of this chapter is based on a paper "Quasi-Periodic Instability in a Linear Prediction Analysis of Voiced Speech" by Nichol and Bogner [11]. In this paper section III was written mainly by the second author and the remaining sections mainly by the first author. The original section III has been somewhat expanded below. A copy of the original paper is included as an insert.

3.1 EXPERIMENTAL RESULTS

As noted in figure 2.1 of the previous chapter, and apparent in figure 3.1 of this chapter, parts of areagrams can show a quasi-periodic fluctuation. These occur during the vowel segments of the utterance. However vowels segments are supposedly the most stable parts of speech and it seems

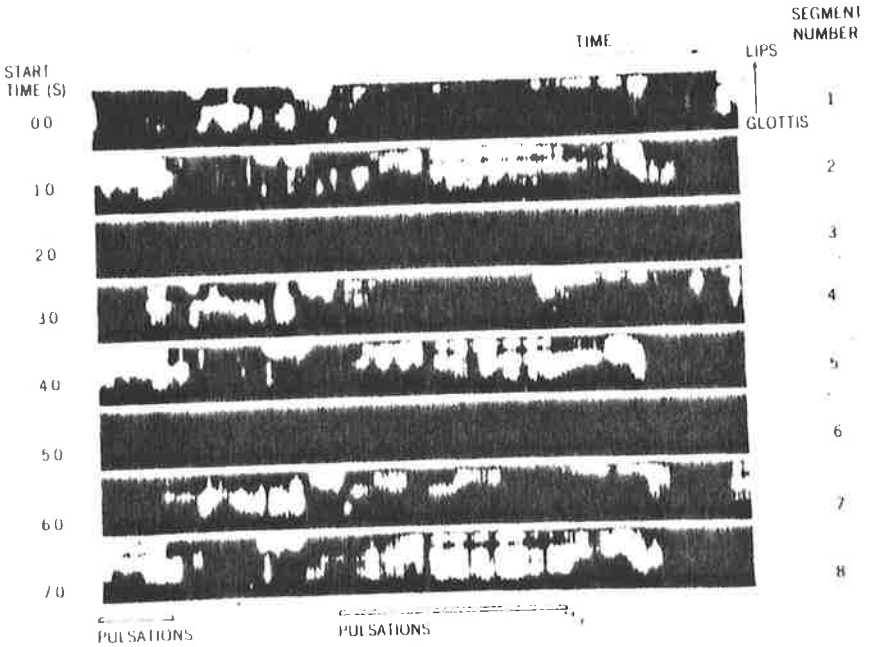


Fig. 1. Vocal tract area function for sentence "Speak to me now, bad kangaroo!"

FIG.3.1

unlikely that this pulsation reflects the actual behaviour of the vocal tract. Figure 3.2(a) and (b) support this viewpoint. These are the time series and waterfall plot of VTAF's for the vowel /ae/ from the word 'bad' in figure 3.1. The time series here, and for all other vowels of this speaker, is remarkably stationary. In figure 3.2(b) the areas vary by a ratio of approximately 10:1 (in waterfall plots the square root of the area is displayed).

A clue to the cause of these pulsations is given by comparing the pitch period N_p of the speaker with the computation interval N_c. For this example the parameters shown in table 3.1 apply.

TABLE 3.1

Sampling frequency	f_s	8192 Hz	
Time series window length	N_w	64 samples	(7.81 mS)
Computation interval	N_c	64 samples	(7.81 mS)
Number of inverse filter coefs.	N_f	8	
Pitch period (observed)	N_p	48 samples	(5.85 mS)
Pulsation period (observed)		254 samples	(31 mS)

The difference between the computational interval and the observed pitch period is such that samples of equal 'phase' will be separated by approximately 31.2 mS, which is very close to the observed period of pulsation. This suggests that the pulsations are an artifact arising from beating between the computation interval N_c and the pitch period N_p.

As mentioned in the introduction this phenomenon had not been observed on previous occasions when areagrams were produced by male speakers repeating the same utterance. This suggests that this is a pitch sensitive effect. As discussed in chapter 2 the Wakita formulation of linear prediction modelling is not pitch synchronous so whilst the total number of glottal

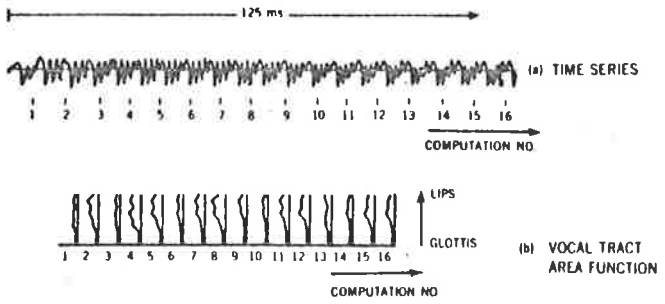


Fig. 2. (a) Speech waveform for /æ/. (b) Vocal tract area function for /æ/.

FIG.3.2

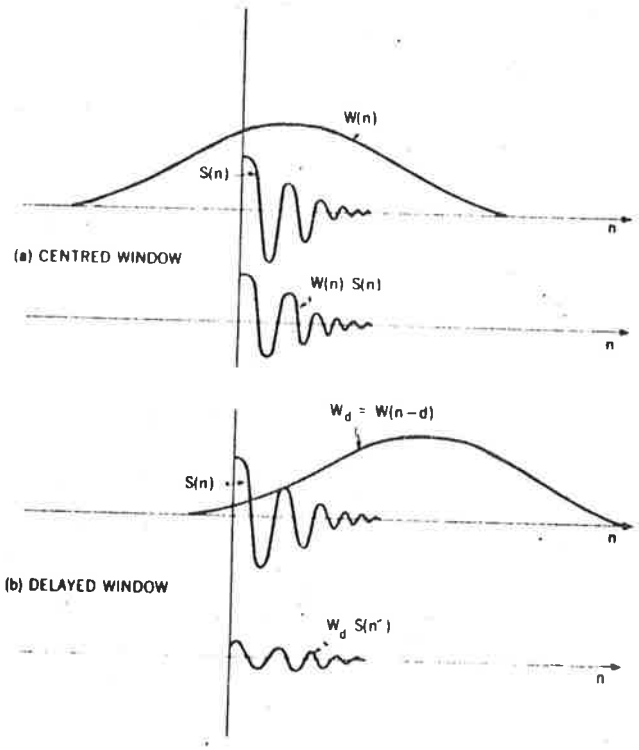


Fig. 3. Effect on windowed signal $s_w(n)$ of relative shift between signal and window. (a) Window centered on signal. (b) Window delayed with respect to signal.

FIG.3.3

impulses occurring in each data window is constant (for constant pitch period) the relative position of the impulses in each window will change unless, by accident, the computation interval is equal to the pitch periods. If the time series window is uniform this will not affect the autocorrelation or subsequent area function computation. However it is customary to use a Hanning window to weight the time-series, not only for VTAF calculation, but also for spectral analysis. Thus the relative position of the pitch impulse, or more correctly the impulse response, within a window may be significant. The effect of changing the relative positions of the impulses within a Hanning window will now be discussed.

3.3 ANALYSIS OF WINDOW POSITION EFFECTS

To develop an understanding of the effects of window position on VTAF estimates a second order (two-junction, three-section) model is used. This is clearly a grossly simplified model but it does facilitate an understanding of the effect and produce an estimate of error comparable to the observed one. The model is shown in figure 3.5. It will be shown below that, ignoring a phase term and a constant multiplier, the impulse response to a second order model is of the form:

$$h(n) = R^n \cos n \omega T, \quad n = 0,1,2.... \tag{3.1}$$

A Hanning window is of the form:

$$W(n) = 0.5 + 0.5 \cos \frac{2\pi(n-d)}{N_w},$$

$$n = \frac{-N_w}{2} \text{ to } \frac{N_w}{2} \tag{3.2}$$

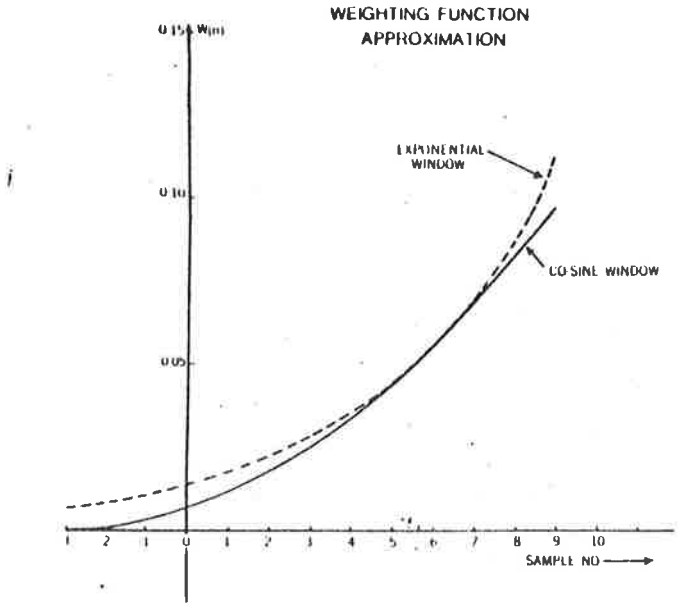


Fig. 4. Approximation to a delayed window $w_d(n) = 0.58\ [1 + \cos (2\pi/128)\ (n - 61)]$ by a rising exponential 0.014×1.26^n.

FIG. 3.4

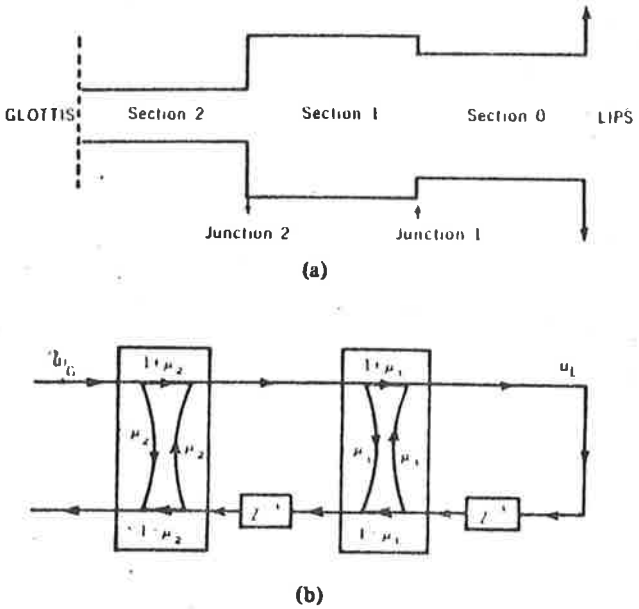


Fig. 5. (a) Physical tube model. (b) Signal flow model.

FIG. 3.5

where d is the delay or offset of the window. It is assumed that N_w is large compared with 1/(r-1) i.e. the effective decay time of the impulse response. Also, as usual in linear prediction analysis, no implicit glottal shaping is incorporated (see chapter 4). Thus the speech signal s(n) is identical to h(n).

In figure 3.3 the effects of the window position on s(n) are shown for different values of the delay d. For the centred case (figure 3.3a) the window has little effect but when the impulse occurs early in the window (figure 3.3b) the windowed speech is effectively undamped.

Figure 3.4 illustrates how the window shape can almost completely remove the speech damping. In this case the window is given by:

$$W_d(n) = 0.5 \left[1 + \cos \frac{2\pi}{128} (n-61) \right] \qquad (3.3)$$

and the exponential is $R\ r_1^{\,n}$ with R = 0.014 and r_1 = 1.26. The number R and r_1 produce a good approximation to the window. Thus for speech of the form

$$s(n) = r_s^{\,n} \cos n \omega T \qquad (3.4)$$

the window will produce the observed speech form (for r_s<1):

$$s_w(n) = s(n)\ W_d(n) \cong R(r_s r_1)^n \cos n \omega T \qquad (3.5)$$

This product $r_s r_1$ is the effective damping coefficient for the observed speech $s_w(n)$. For speech sampled at 10 kHz the value of r_s is likely to be in the range of 0.985 to 0.900 for formants with 50 to 300 Hz bandwidth. For the example given the effective damping coefficient will be greater than unity. Even if not greater than unity significant errors clearly can arise due to the window. The effects of this error on the area estimates

will now be discussed.

Figure 3.5a shows a two-junction acoustic tube model of the vocal tract. Denote the cross-sectional areas from the glottis to lips by a_2, a_1 and a_0 respectively. The volume velocity reflection coefficients are μ_m and related to the areas by [12]:

$$\mu_m = \frac{a_{m-1} - a_m}{a_{m-1} + a_m} \tag{3.6}$$

Total reflection ($\mu = 1$) is assumed at the lips; this is equivalent to a further tube of infinite area. For ease of analysis all delays in the model are associated with the backwards travelling wave in each section. These assumptions allow the acoustic model to be replaced by the signal flow graph of figure 3.5b. The glottal and lips flow are denoted by u_G and u_L respectively. Label the modes in figure 3.5b from 1 (upper left) to 8 (lower left) by proceeding clockwise around the graph and denote the mode value of the jth mode by W_j. Analysis of the graph then shows that the W's are related by:

$$\begin{aligned}
W_1(n) &= u_G(n) \\
W_2(n) &= (1+\mu_2)\, W_1(n) \\
W_3(n) &= W_2(n) \\
W_4(n) &= (1+\mu_1)\, W_3(n) + \mu_1\, W_5(n) \\
W_5(n) &= W_4(n-1) \\
W_6(n) &= -\mu_1\, W_3(n) + (1-\mu_1)\, W_5(n) \\
W_7(n) &= W_6(n-1) \\
W_8(n) &= -\mu_2\, W_1(n) + (1-\mu_2)\, W_7(n) \\
u_L'(n) &= W_4(n)
\end{aligned} \tag{3.7}$$

Solving these in the usual manner shows that, in z-transform notation, the transfer function H(z) is given by:

$$H(z) = U_L(z)/U_G(z)$$

$$H(z) = \frac{(1+\mu_1)\ (1+\mu_2)}{1 - z^{-1}\mu_1(1-\mu_2) - z^{-2}\mu_2} \qquad (3.8)$$

From (3.6) the equivalent expression in terms of cross-sectional areas is:

$$H(z) = \frac{4\ a_0\ a_1}{(a_0+a_1)\ (a_1+a_2)} \cdot \frac{1}{1 - z^{-1}\left(\frac{2\ a_2}{a_1+a_2}\right)\left(\frac{a_0-a_1}{a_0+a_1}\right) - z^{-2}\left(\frac{a_1-a_2}{a_1+a_2}\right)} \qquad (3.9)$$

The required impulse response can now be obtained.

Putting $c = (1+\mu_1)(1+\mu_2)$, $d = \mu_1(1-\mu_2)$ and $e = \mu_2$ into (3.8) gives:

$$H(z) = \frac{c\ z^2}{z^2 - dz - e} \qquad (3.10)$$

If the roots of the denominator of (3.10) are α and β then equivalently

$$H(z) = \frac{c\ z^2}{(z - \alpha)(z - \beta)} \qquad (3.11)$$

where $\alpha = d/2 - i\sqrt{-e-d^2/4}$

and $\beta = d/2 + i\sqrt{-e-d^2/4}$ } (3.12)

Expanding (3.11) by partial fractions gives

$$H(z) = c\left[\frac{\alpha}{\alpha-\beta}\frac{z}{z-\alpha} - \frac{\beta}{\alpha-\beta}\frac{z}{z-\beta}\right] \qquad (3.13)$$

The impulse response is given, as usual, by comparing coefficients of h(n)

and the series expansion of the right hand side of (3.13). This yields:

$$h(n) = \frac{c}{\alpha-\beta}\left[\alpha^{n+1} - \beta^{n+1}\right] \qquad (3.14)$$

$$n \geqq 0.$$

From (3.12) it is clear that either α and β are real or else complex conjugates. Supposing the latter and putting $\alpha = r\, e^{i\omega T}$ into (3.14) gives:

$$h(n) = c\, r^n \frac{\sin (n+1)\omega T}{\sin \omega T} \qquad (3.15)$$

where T is twice the time each way delay in each section. From (3.12) and replacing a, b and c gives:

$$h(n) = h(0)\, r^n \frac{\sin (n+1)\omega T}{\sin \omega T} \qquad (3.16)$$

where

$$h(0) = (1+\mu_1)(1+\mu_2) = \frac{4\, a_0\, a_1}{(a_0+a_1)(a_1+a_2)} \qquad (3.17)$$

$$r^2 = -\mu_2 = \frac{a_2-a_1}{a_2+a_1} \qquad (3.18)$$

and

$$\cos \omega T = \frac{-1}{2r} \mu_1(1+\mu_2)$$

$$= \frac{1}{2r} \left(\frac{2a_2}{a_1+a_2}\right)\left(\frac{a_o-a_1}{a_o+a_1}\right) \qquad (3.19)$$

Apart from a constant multiplier and a phase term the impulse response of (3.15) is of the same form as (3.1). It is clear from 3.18 that the damping coefficient r is dependent only on the reflection coefficient at the junction closest to the glottis. It should be noted that if $\mu_2<0$ then the roots α and β of (3.12) are real and the response is not oscillatory. This case is thus not of interest in the present study.

To apply this model to a realistic sized vocal tract each section is put equal to half the length of a vocal tract i.e. about 9 cm. For a sampling rate of 10 kHz this is equivalent to an each way delay of 3×10^{-4} s. The delay T in the second order model described above thus becomes 6×10^{-4} s. The delay terms in figure 3.5b become z^{-6} and the second order model becomes a twelth order one. Thus, in the transfer function expression given by (3.8), z^{-1} should be replaced by z^{-6}. The denominator in the equivalent expression to (3.11) thus has factors resulting in 6 poles. However the base pole of each factor is still $z = r\, e^{i\omega T}$.

From (3.18)

$$a_2 = \frac{1 + r^2}{1 - r^2} a_1 \qquad (3.20)$$

From this

$$\frac{1}{a_2}\left.\frac{\partial a_2}{\partial r}\right|_{a_1} = \frac{4r}{1-r^4} \tag{3.21}$$

and

$$\frac{1}{a_1}\left.\frac{\partial a_1}{\partial r}\right|_{a_2} = \frac{-4r}{1-r^4} \tag{3.22}$$

Thus the proportional variation of a_2 or a_1, when the other is held constant, becomes very great as $r \to 1$. As seen above the effect of windowing can be to cause the apparent value of r to approach and even exceed unity. Letting r move from 0.90 to 0.99 will, from 3.20, produce a variation of a_2/a_1 from 9.53 to 99.5. Clearly this effect will cause area variations as large as those noted in section 3.2.

To complete this section note that ωT is not affected by the window phenomenon and thus $d(\cos \omega T) = 0$ when (3.19) is differentiated. Hence

$$a_0 = a_1 \frac{r}{1+r^2} \cos \omega T \tag{3.23}$$

and

$$\frac{1}{a_0}\left.\frac{\partial a_0}{\partial r}\right|_{a_1} = \frac{1-r^2}{r(1+r^2)} \tag{3.24}$$

The second junction area ratios are thus not strongly influenced by r. For more complicated vocal tract shapes the effects are more complex. However this model appears to adequately explain the observed effect, quantitatively as well as qualitatively. A previous study [13] showed that under moderate variation of the pole dampings in a five pole signal the resultant VTAF retained its gross features, but underwent a gradual smooth change.

3.3 EXPERIMENTAL DIAGNOSIS

To show, under more controlled conditions, the experimentally and theoretically demonstrated sensitivity of the linear predicted VTAF to impulse position, synthetic vowels were generated using an 8 pole (four-formant) model. The pole positions and bandwidths were those measured by Fant [14] for Russian vowels. These are shown in table 3.2.

TABLE 3.2

Vowel	First formant		Second formant		Third formant		Fourth form.	
	Frequency	Bandwidth	F.	B.	F.	B.	F.	B.
/ae/	616	57	1072	72	2430	130	3410	176
/e/	432	39	1959	95	2722	170	3500	325
/I/	222	60	2244	75	3140	240	3700	230
/ /	510	54	900	65	2400	100	3220	135
/Λ/	231	69	615	50	2375	110	3320	115

Figure 3.6 shows plots of the VTAF's obtained for the vowel /ae/ when the pitch period N_p is equal to the computation interval N_c. In each column of figure 6 the impulse occurs in a different portion of the Hanning window ($N_w = N_c$). Remembering that this waterfall plot shows the square root of area it is clear that the variation in area observed is very similar to that for real speech. In this case however no variation occurs with time as the 'phase' of the impulse is constant and consequently so is the (apparent) value of the damping coefficient r. When the nexus between pitch period and computation interval is broken the 'phase' changes with time and variations occur with time for all vowels as shown in figure 3.7. Not all vowels are as susceptible to this effect as /ae/, but this is to be expected in view of their different damping coefficients. This has also been noted in real speech. In figure 3.7 $N_p = 0.8\ N_c$ and the pulsation period is similar to that observed in the female speaker of figure 3.1.

For male speakers N_p is much closer to the N_c used in the present study; hence the pulsation effect is not nearly as noticable as only one complete

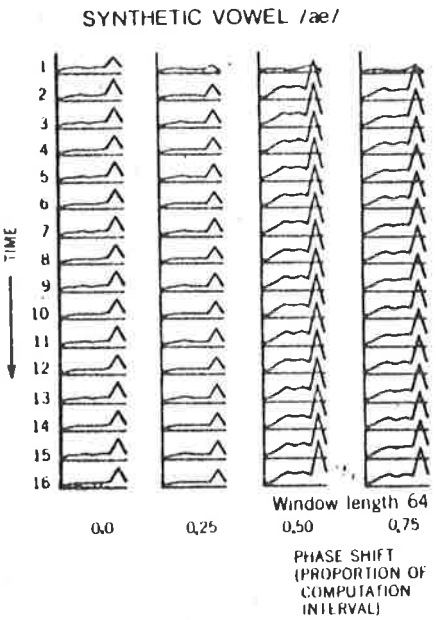


FIG.3.6

Fig. 6. Pitch period and computation interval are equal but with different relative positions in each column.

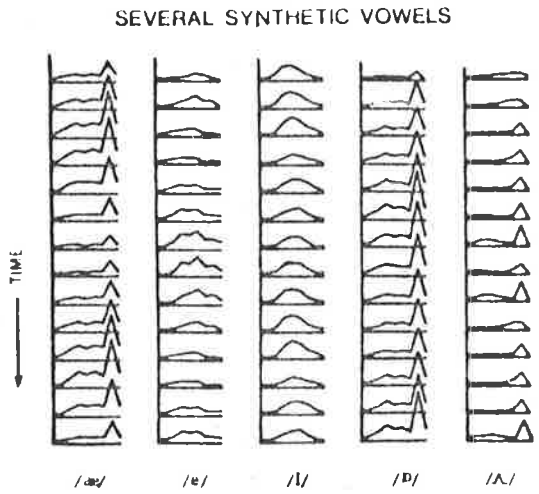


FIG.3.7

Fig. 7. Changing position of pitch pulse within window leads to pulsations when $n_c = n_w$.

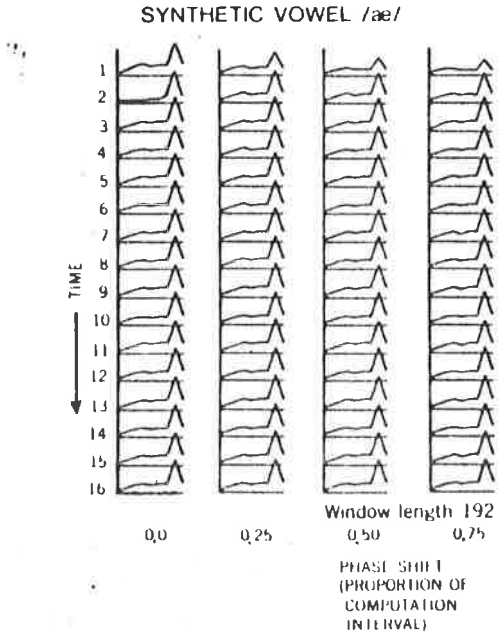


FIG.3.8

Fig. 8. Increasing n_w to 3.5 n_p suppresses the fluctuations for synthetic vowel /æ/.

cycle may occur in a given vowel segment. Errors of the same magnitude in r (and hence area) will still be occurring but may not be noticable. This is undoubtably the reason why this effect was first noticed in female speech. However by increasing N_c, the pulsation effect can be induced in male speech.

3.4 SUPPRESSING THE FLUCTUATIONS

Two obvious 'cures' for this problem are, firstly, using a uniform window and, secondly, increasing the window length. Both do alleviate the problem but both have their penalties. The first will tend to cause fluctuations in the autocorrelation function estimates r(j). Essentially this is because only a few sample points are used for the longest shifts and thus estimates of r(j) for large j are unreliable. Apart from this effect the estimates of r(j) are independent of impulse position and the pulsations are suppressed. In the second method several impulses are included in each window and the overall influence of any one impulse is correspondingly reduced. The cost is a decrease in time resolution. Figure 3.8 shows the effect of increasing window length (to 3.5 W_p) on the same synthetic speech used in figure 3.6. Clearly the effect of 'phase' of the impulse is suppressed by this technique.

Figure 3.9 shows the original areagram data reprocessed with a window length of 3.5 N_p. Most pulsation effects have been removed. After considerable experimentation the second approach was adopted as standard for the areagram presentations. Basically this is because windows longer than N_p are still needed for the first approach in order to suppress the fluctuations due to increased errors in r(j). This means that time resolution is also reduced with this technique. A Hanning window is preferred to maintain commonality with the spectral analysis window.

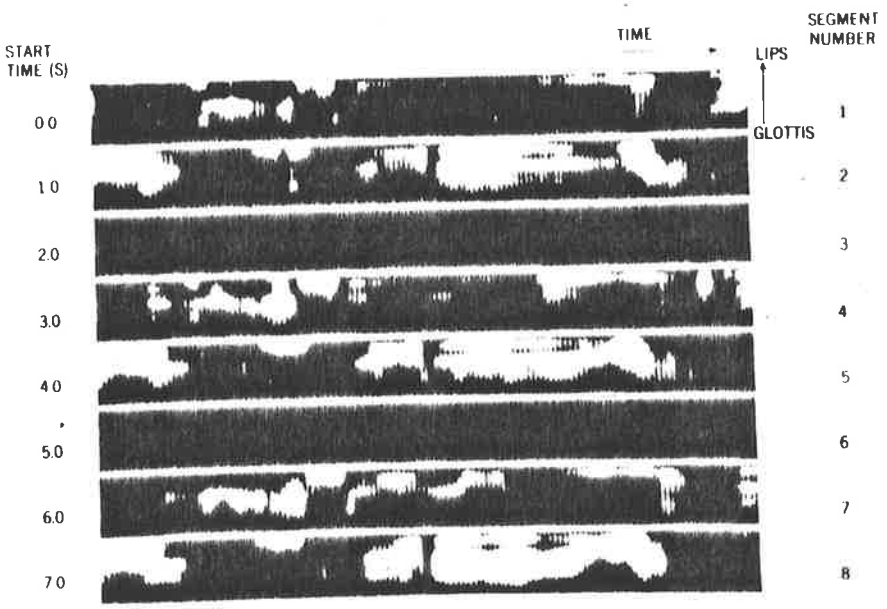


Fig. 9. Vocal tract area function for "Speak to me now, bad kangaroo!" for window n_w = 3.5 times pitch period n_p.

FIG.3.9

CHAPTER 4

SUPPRESSING ANAMOLOUS TRACT SHAPES DUE TO RANDOM NOISES AND SPECTRAL ZEROS

It was noted in Chapter 2 that unrealistically high labio-dental areas were sometimes obtained using the Wakita algorithm. In this chapter it is shown that two classes of sources produce these anomolous areas. The first of these are random noises which may inadvertantly occur during the 'silences' between utterances. The second class is produced by non-glottal speech sounds, such as fricatives, which have zeros in their spectra. As the first class are non-articulatory sounds there is no point in computing their VTAF's. Fortunately these sounds can be easily detected by computing a running measure of signal energy and the area function computation can be stopped when this measure is below a threshold. The second class is a little more difficult as they consist of real speech sounds such as fricatives and stops which the (finite) all-pole model cannot accommodate as their spectra contains zeros. Clearly a complete solution to this problem requires the development of a more general speech model and is thus outside the scope of the present study. To alleviate this problem an algorithm to detect this class of sounds is described. However rather than, as for the random noises, not computing the areas at all during such segments it is suggested that the speech pre-emphasis (i.e. pre-whitening) be modified before VTAF computation. This leads to more realistic areas and improves the areagram appearance. Of course it is not suggested that these areas are now exact (this would require that the zeros be modelled as well as the poles) but that they are significantly better estimates than the pre-emphasised values.

Before describing the algorithms and examining their effectiveness some discussion of the observed VTAF's and their corresponding spectra is given.

4.1 OBSERVED SPECTRA AND AREAS

Figure 4.2 shows waterfall plots of the areas calculated from the time series shown in figure 4.1. The utterance shown is the word 'beat' spoken by a male subject. In figure 4.2 asterisks show some of the anomolous VTAF estimates. Clearly those occurring before the utterance do not correspond to real articulations but seem (after listening to the recordings) to arise from background noise, tape hiss, microphone clicks etc. This class will be referred to as 'random noises'. In figure 4.2 a second region of anomolous areas can be seen. These correspond to the /t/ consonant in this utterance. These VTAF estimates are clearly anomolous because they imply lip to glottis area ratios of around 100:1 (waterfall plots show the square root of area) and the growth is almost monotonic. This is contrary to the observed vocal tract shape for this stop. Very similar, but equally anomolous, VTAF's are seen for nearly the whole range of fricative and stops. Thus, besides /t/, the other stop consonants /d/ and /g/, the voiced fricatives /v/, /ð/, /ʒ/ and /z/ and voiceless fricatives /f/, /θ/, /s/ and /ʃ/ all produce anomolous areas. The voiceless fricative /h/ does not appear to produce this effect and the voiced fricatives sometimes appear as an intermediate form. That is the anomolous lip area is superposed on a reasonably shaped vocal tract. These observations strongly suggest that it is the position of the source which is producing the anomoly [15]. This second class of sources will be referred to as the 'fricative class'.

It should be noted that the areas produced by the classes are somewhat different. The random noises produce a monotonically increasing VTAF but the fricatives tend to be more irregular and have a final (lip) area which is less than the previous one. It should also be noted that some of the first class of areas appear to be almost perfect exponential horns. Some speculation on this is given below. These results are summarised in table 4.1.

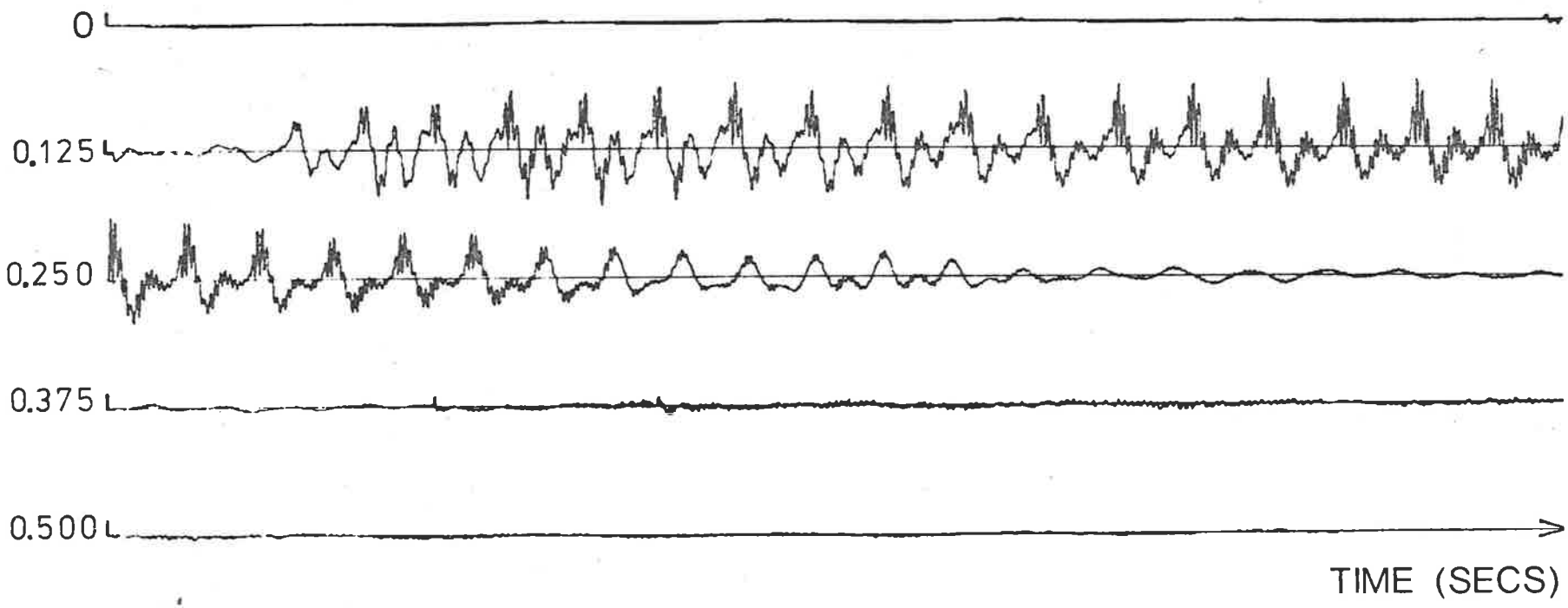


Figure 4.1 TIME SERIES OF 'BEAT'

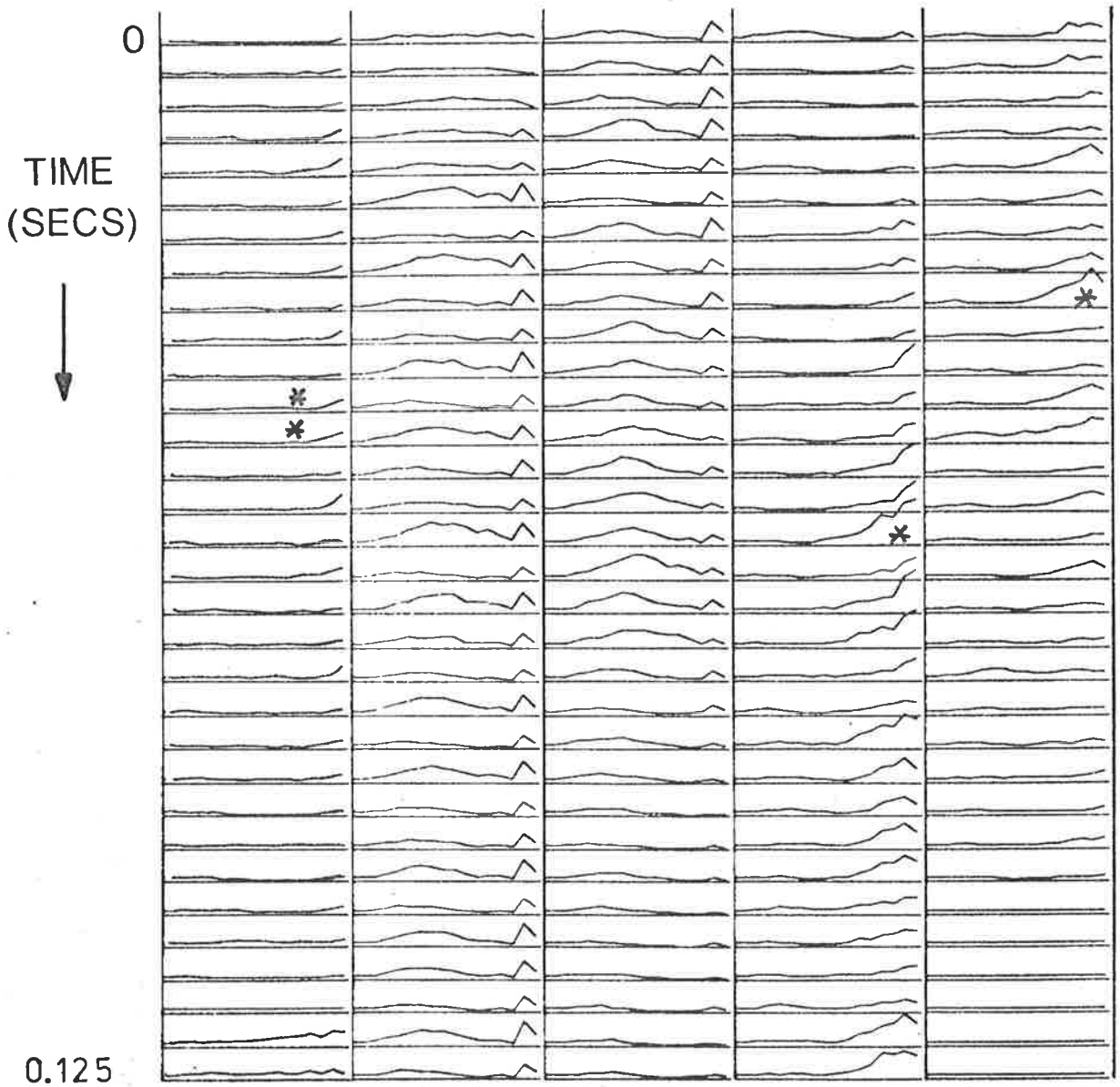


Figure 4.2 VOCAL TRACT AREA FUNCTION ESTIMATES OF 'BEAT', (GLOTTIS TO LEFT) : γ = 1.0

TABLE 4.1

Class	Source	Area Function	Spectrum
Type 1	Random, non speech	$a_1>a_2>a_3>....> a_m$	High range enhancement
Type 2	Fricatives, stops	$a_1<a_2>a_3>....> a_m$	mid and high range enhancement

Also shown in table 4.1 are the observed character of the spectra. This character may be seen in figure 4.3 which shows the short term spectral estimates for the waveform shown in figure 4.1. As usual in computing speech spectra the time series was differenced to emphasise the higher frequencies [16]. The same practice is employed before computing the autocorrelation function in the VTAF calculation. Thus the all-pole modelling has to produce spectra similar to those shown in figure 4.3. The reasons and desirability of this pre-emphasis are discussed below in section 4.3. Before discussing this some speculation on the origin of these anomolous shapes is given.

4.2 EXPONENTIAL HORN

It was noted above that the class 1 (random noises) anomolies were sometimes exponential and it is interesting to speculate why this should be. Some insight into this can be gained by considering the case of a wave propagating down an acoustic exponential horn.

For the case of a correctly terminated horn (and thus with propagation only in the positive direction) the pressure P and volume velocity V are given as a function of distance (x) and time (t) by [17]:

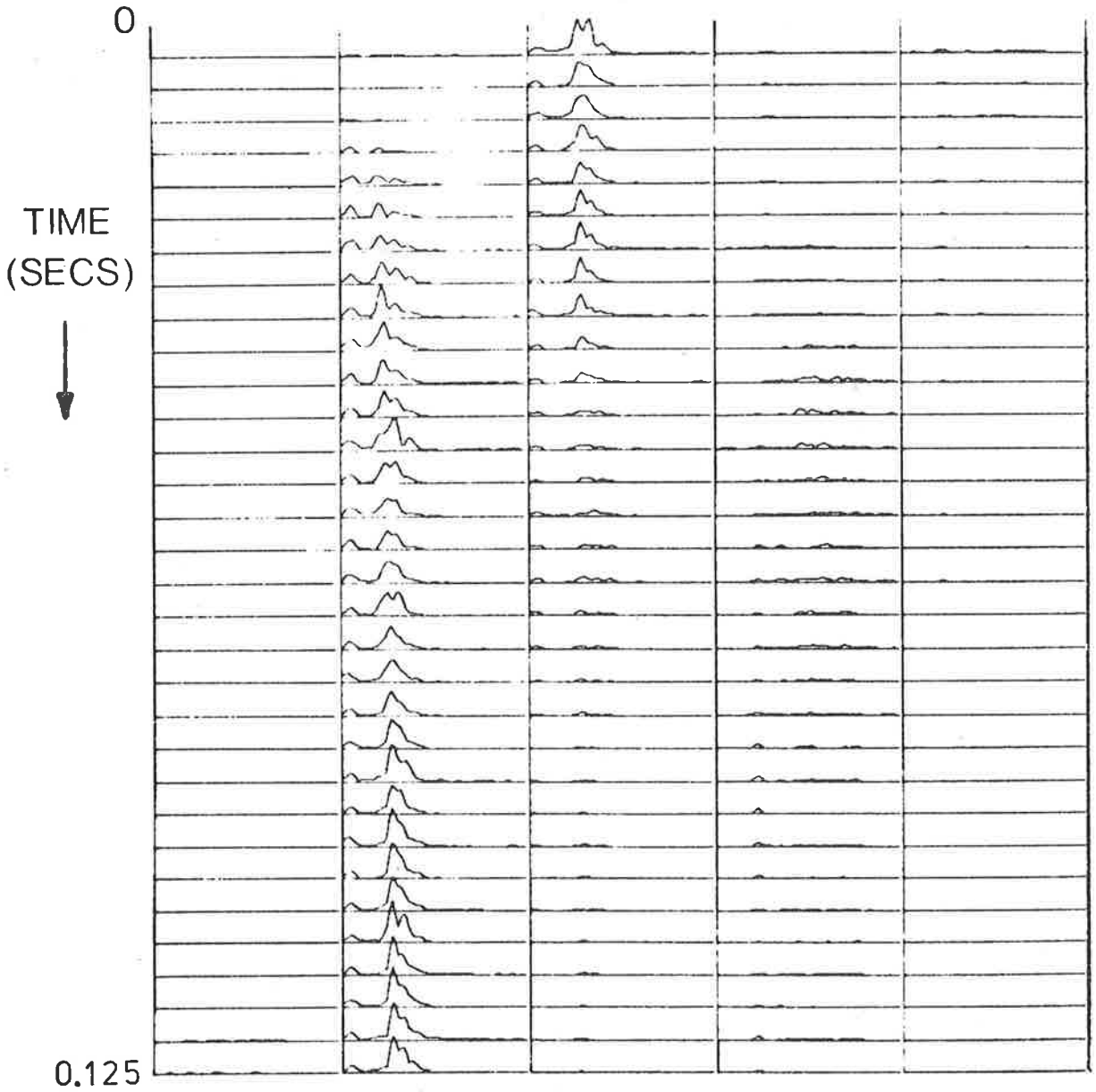


Figure 4.3 SHORT TERM SPECTRA OF 'BEAT'
0-8192 Hz

$$P \cong -(i\rho\omega/\sqrt{So})\ C_+ \exp[i\ x \sqrt{\left(\frac{\omega}{c}\right)^2 - \left(\frac{1}{h}\right)^2}\ i\ \omega\ t - \frac{x}{h}] \qquad (4.1)$$

and

$$V \cong \frac{-i}{\sqrt{So}} \sqrt{\left(\frac{\omega}{c}\right)^2 - \left(\frac{1}{h}\right)^2}\ C_+ \exp\ [i\ x\sqrt{\left(\frac{\omega}{c}\right)^2 - \left(\frac{1}{h}\right)^2}\ -i\ \omega\ t - \frac{x}{h}] \qquad (4.2)$$

where $\omega = 2\pi f$ is the angular wave frequency, c the velocity of sound in air, ρ the density, C_+ the wave amplitude and S_o and S are tube areas at the origin and distance x respectively. These areas are related by an exponential expression:

$$S = S_o \exp 2x/h \qquad (4.3)$$

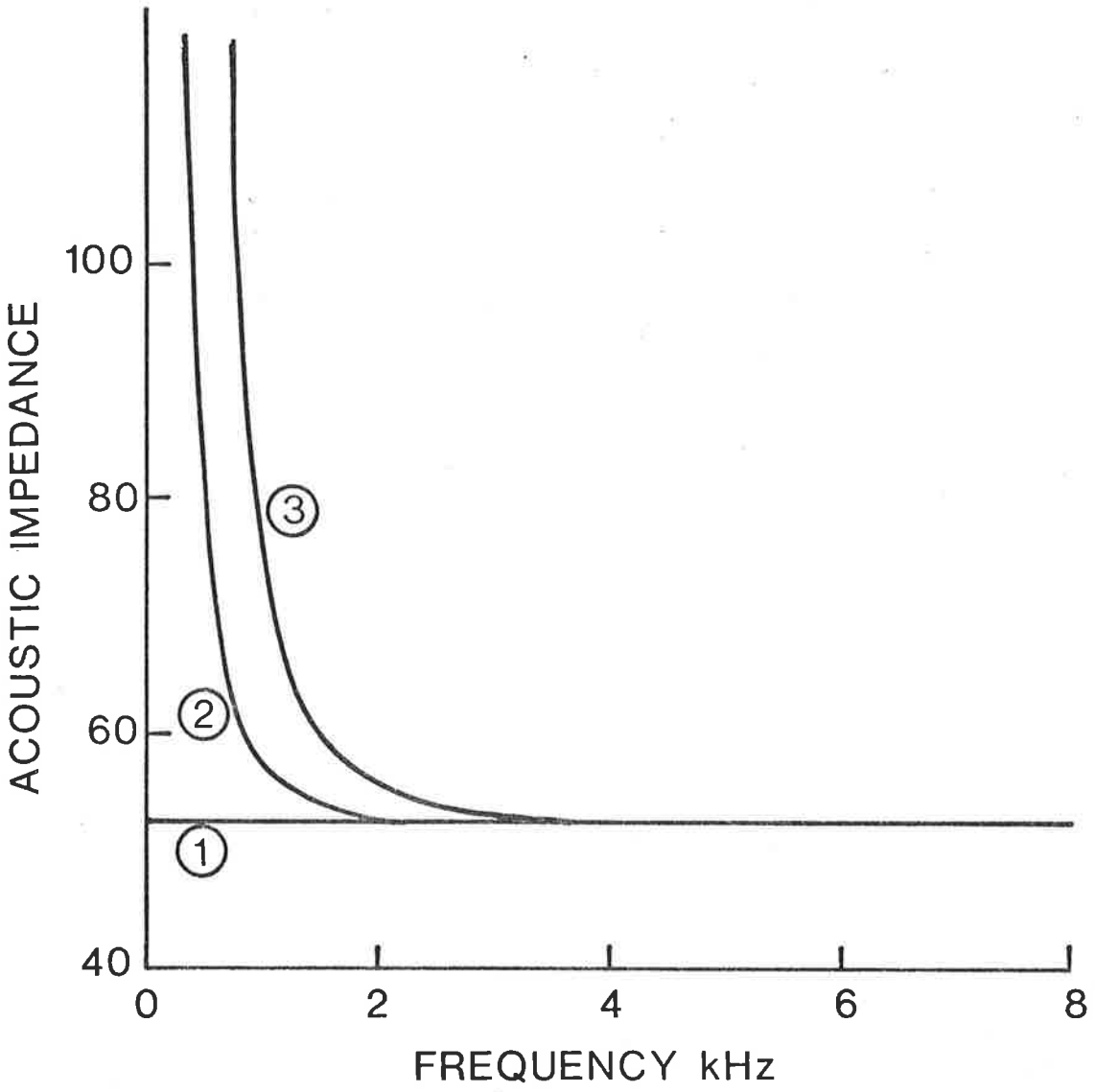
The acoustic impedance is thus given by:

$$Z_A = P/V$$

$$= \frac{\rho\omega}{\sqrt{\omega^2 - \left(\frac{c}{h}\right)^2}} \qquad (4.4)$$

Figure 4.4 shows the variation of Z_A with frequency for various values of h. For $h \to \infty$ the impedance is constant with frequency, but for finite h the impedance decreases with frequency (above the characteristic frequency).

Thus the observed shape of the vocal tract (h~0.07) appears to arise because the model is trying to match the (observed) rising spectrum. The more complex shape of class 2 anomolies no doubt reflects their more complex spectrum (mid-range as well as high frequency enhancement).



	AREA RATIO (at 0.17 m)	HORN PARAMETER (H)
CURVE ①	1	0.0738
CURVE ②	10	0.1477
CURVE ③	100	∞

Figure 4.4 PLOTS OF ACOUSTIC IMPEDANCE (IN ARBITARY UNITS) FOR VARIOUSLY SHAPED EXPONENTIAL HORNS

An examination is now made of the rationale for pre-emphasis in VTAF computation.

4.3 SPECTRAL PRE-EMPHASIS

The model of the vocal tract used in linear prediction analysis does not explicitly incorporate glottal shaping or lip radiation impedance. It does these implicitly by the following strategem [18]. Denote the glottal excitation by E(z), the glottal shaping by G(z), vocal tract transfer function by V(z) and lip radiation model by L(z). The received speech is then given (in z-transform notation) by:

$$S(z) = E(z)\, G(z)\, V(z)\, L(z) \tag{4.5}$$

G(z) is approximated by a two pole filter of the form:

$$G(z) = \frac{1}{(1 - e^{-cT} z^{-1})^2} \tag{4.6}$$

and the lip radiation is assumed to be approximately 6dB/octave and given by:

$$L(z) = 1 - z^{-1} \tag{4.7}$$

In equation 4.6 the product c T, where c is the speed of sound and T is the sampling interval, is typically of the order 0.03 and so $e^{-c\,T}$ is approximately unity. Substituting (4.6) and (4.7) into (4.1) thus gives the approximate solution:

$$S(z) = E(z)\, V(z)\, (1 - z^{-1})^{-1} \tag{4.8}$$

The usual technique is to define a 'pre-emphasis' filter to be applied to the sampled speech and given by:

$$P(z) = (1-z^{-1}) \tag{4.9}$$

and hence:

$$P(z)\ S(z) = E(z)\ V(z) \tag{4.10}$$

Thus if the received speech S(z) is pre-emphasised by P(z) then the white noise model of chapter 2 (the right hand side of (4.10)) will hold.

The pre-emphasis is accomplished by differencing the speech time series. That is by computing:

$$x(n) = s(n)-s(n-1) \tag{4.11}$$

The present chapter is concerned with cases where the implicit assumptions leading to pre-emphasis do not hold. Thus the resultant area anomolies may be partially ascribed to the inclusion of glottal shaping G(z) for class 2 sounds when this is not in fact appropriate. A technique to automatically recognise and partially correct for these effects is described. However, as mentioned above, a complete solution will have to await the development of a more comprehensive vocal tract model.

4.4 VARIABLE PRE-EMPHASIS AND RANDOM NOISE DETECTION

From the above discussion it is clear that pre-emphasis is not only unnecessary but also deliterious for non-glottal sources. Accordingly, at the very least, pre-emphasis should be removed for fricatives and stops. In the case of random background noise no computation of VTAF's should be made at all. What is required is thus both variable pre-emphasis and

'random noise detection'. The first of these requirements has been suggested by other workers [19] but with somewhat different emphasis. Their major interest was linear prediction spectral analysis and they were concerned with the spectral shapes resulting from pre-emphasis. Their basic conclusions and those of the present study are in agreement however. As random noise is not a problem for spectral analysis (the total energy in background noise is small) a different approach has been developed for the present vocal tract shape study.

The following algorithm is included in module 3 to detect and classify both fricatives and random noises. The flow chart for this program is shown in figure 4.5.

(1) Each 'window' of time series data TDATA is sent for testing before evaluating the vocal tract area function. The data is unweighted and non pre-emphasised.

(2) Two pattern recognition 'features' are extracted from the samples in TDATA. These are the variance X_v and a measure of high frequency signal X_H. These are calculated over a subset of TDATA; viz. the N_c data points s(n) centred on the midpoint of TDATA. The idea of this is that the results of this analysis will be centre weighted and thus fairly independent of window lengths. This can be important for long windows which, as seen in chapter 3, can cover several compution intervals. Thus:

$$X_v = \sum_{n=n'-N_{c/2}}^{n'+N_{c/2}-1} (s(n) - m_{n'})^2 \qquad (4.12)$$

and

$$X_H = \sum_{n=n'-N_{c/2}}^{n'+N_{c/2}-1} (s(n) - s(n-1))^2 \qquad (4.13)$$

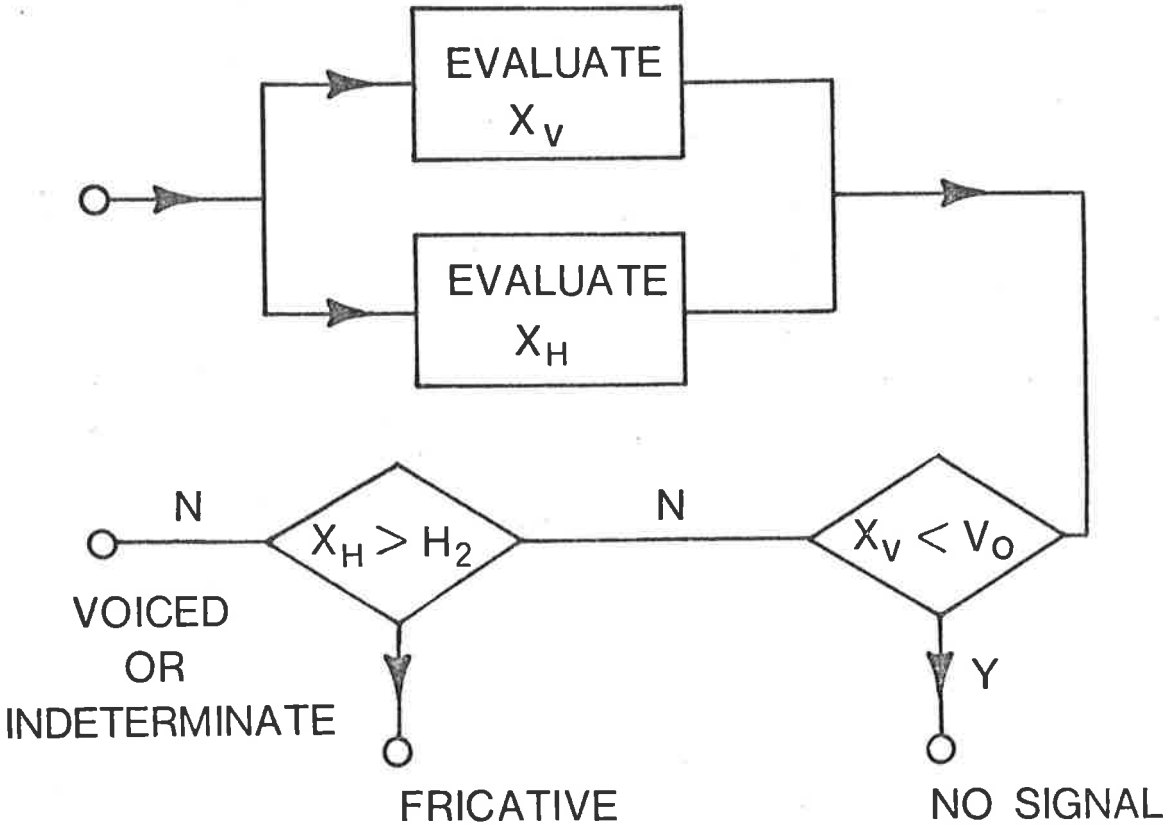


Figure 4.5 PRE-PROCESSING CLASSIFICATION

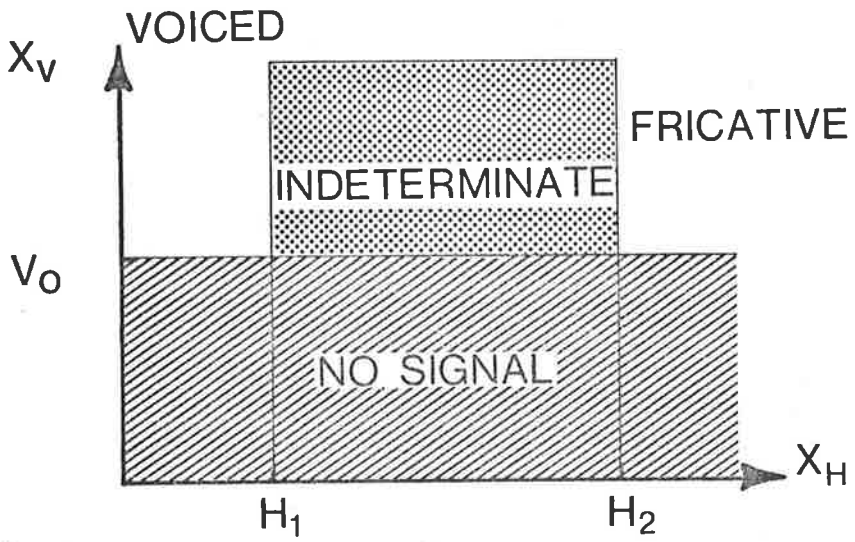


Figure 4.6 FRICATIVE FEATURE SPACE

where n' is the centre of the current window and $m_{n'}$ is the mean of this window. Obviously X_H is a measure of high frequency content as it is the same form as a pre-emphasis filter.

Conceptually these features can be plotted in a two dimensional feature space as shown in figure 4.6. Each new block of data generates a point in this space.

(3) Discrimination is based on the following principles. Thresholds are defined for both features. These are V_o for X_v and H_1 and H_2 for X_H. As seen in figure 4.6 $V_o>0$ and $H_2>H_1>0$.

Various sub-regions of feature space are determined by these thresholds and the current speech segment is classified according into which sub-region it falls. The classes are shown in table 4.2

TABLE 4.2

TYPE	TEST	ACTION
A. No signal	$X_v<V_o$	Suppress VTAF compution
B. Fricative	$(X_v>V_o)\wedge(X_H>H_2)$	Minimum pre-emphasis
C. Non-voiced speech	$(X_v>V_o)\wedge(X_H<H_1)$	Maximum pre-emphasis
D. Indeterminate	$(X_v>V_o)\wedge(H_1\leqslant X_H\leqslant H_2)$	Variable pre-emphasis

For classes B, C and D the pre-emphasis is implemented by computing the new time series:

$$x(n) = s(n) - \gamma\, s(n-1) \qquad (4.14)$$

where $0\leqq\gamma_1<\gamma<\gamma_2\leqq 1$.

For the case D γ is given by:

$$\gamma = \gamma_1 + (X_H - H_1) \cdot (\gamma_2 - \gamma_1)/(H_2 - H_1) \qquad (4.15)$$

For cases B and C values of γ are γ_1 and γ_2 respectively.

Many different values for the thresholds, and for the pre-emphasis limits γ_1 and γ_2, were tried. Figure 4.7 shows the results of using various values for γ (that is $\gamma_1 = \gamma_2 = \gamma$) on the data shown in figure 4.1. It is interesting to note that for $\gamma = 0$, that is no equalisation, the exponential shapes of figure 4.2 change into nearly uniform tubes. As seen in figure 4.7 the non pre-emphasised spectra for these segments is then almost flat. This is what would be expected from equation (4.4) because with $h = \infty$ the impedance of an exponential tube is constant with frequency. After much experimentation the values shown in table 4.3 were used for the remainder of this study. Obviously these thresholds depends on the particular equipment used.

TABLE 4.3

V_o	H_1	H_2	γ_1	γ_2
1.0	.05	.07	.3	.9

Little change in these values was found to be required for the present experimental set-up when new data was added to the data base. However a facility to test the new data by evaluating X_H and X_v for sequential blocks was included in module 4 and routinely used before analysis.

Figure 4.8 shows the same data as figure 4.2 processed using the values shown in table 4.4. Nearly all evidence of anomoly has been removed. Figure 4.9 shows an areagram representation of this same utterance with the upper two frames showing the effect of non-variable pre-emphasis and the lower two the results of variable pre-emphasis. The straight forward techniques described in this chapter appear adequate to produce satisfactory areagrams.

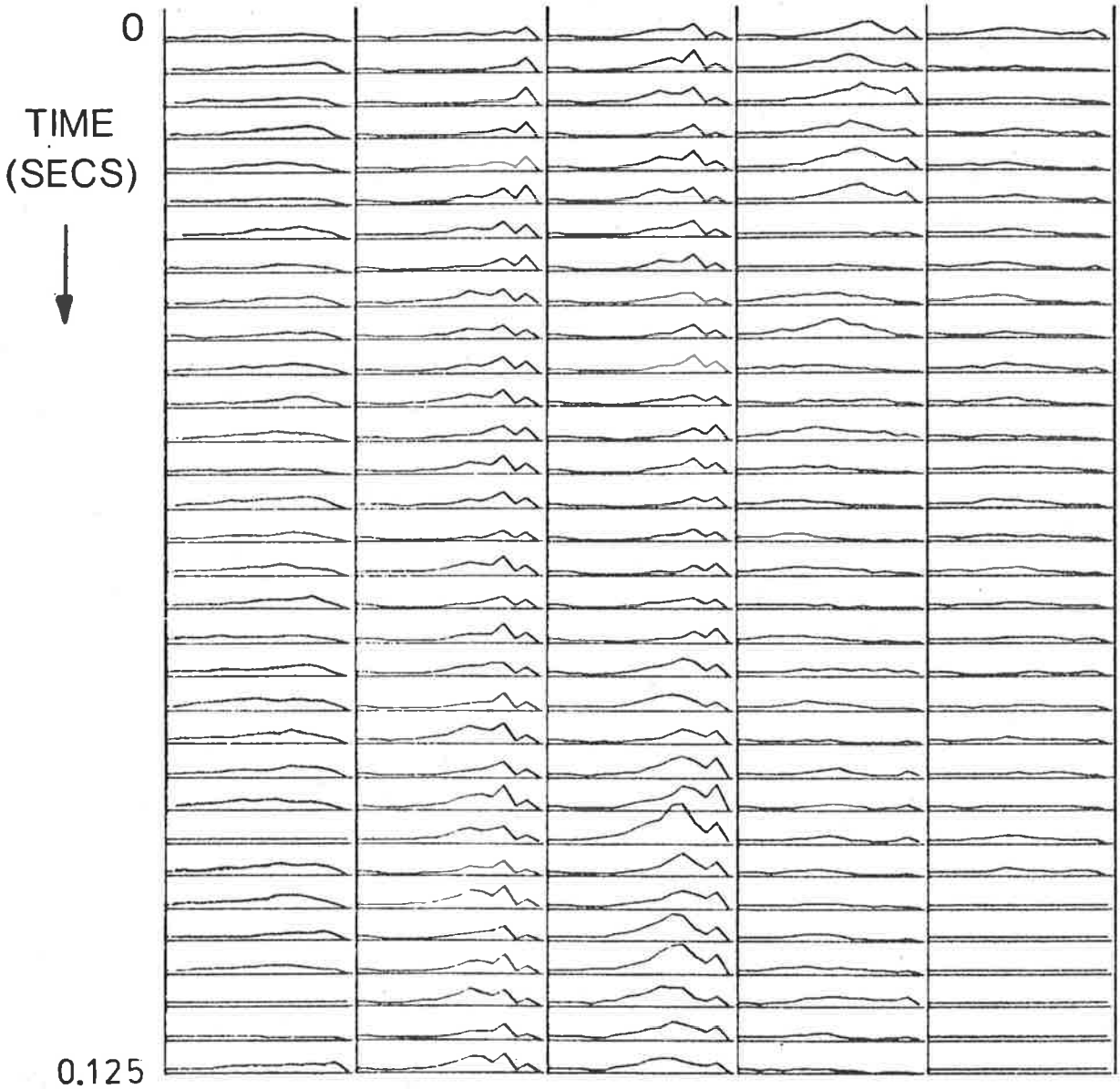


Figure 4.7a VOCAL TRACT AREA FUNCTION ESTIMATES OF 'BEAT' : γ = 0.0

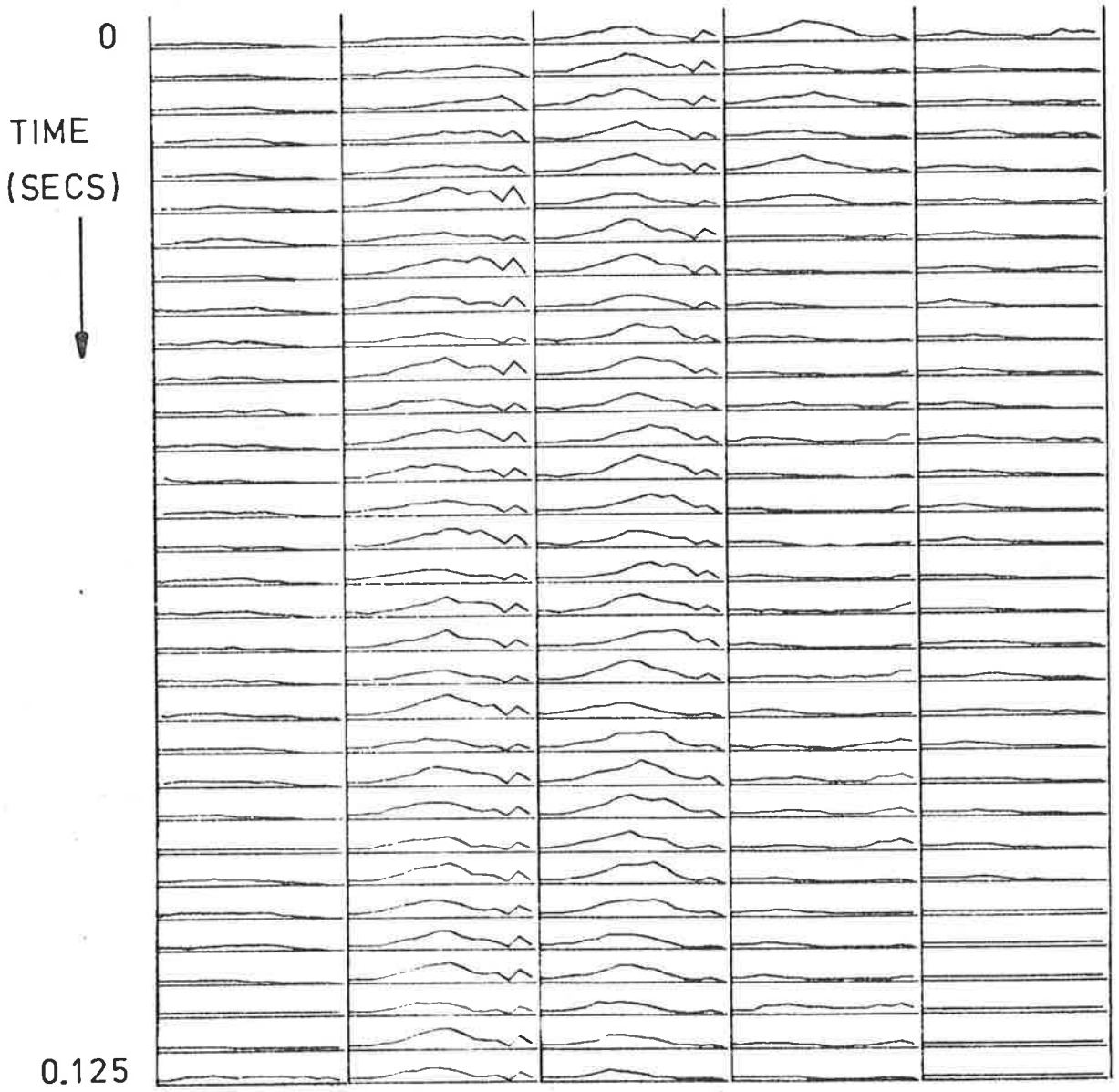


Figure 4.7b VOCAL TRACT AREA FUNCTION ESTIMATES OF 'BEAT' : $\gamma = 0.8$

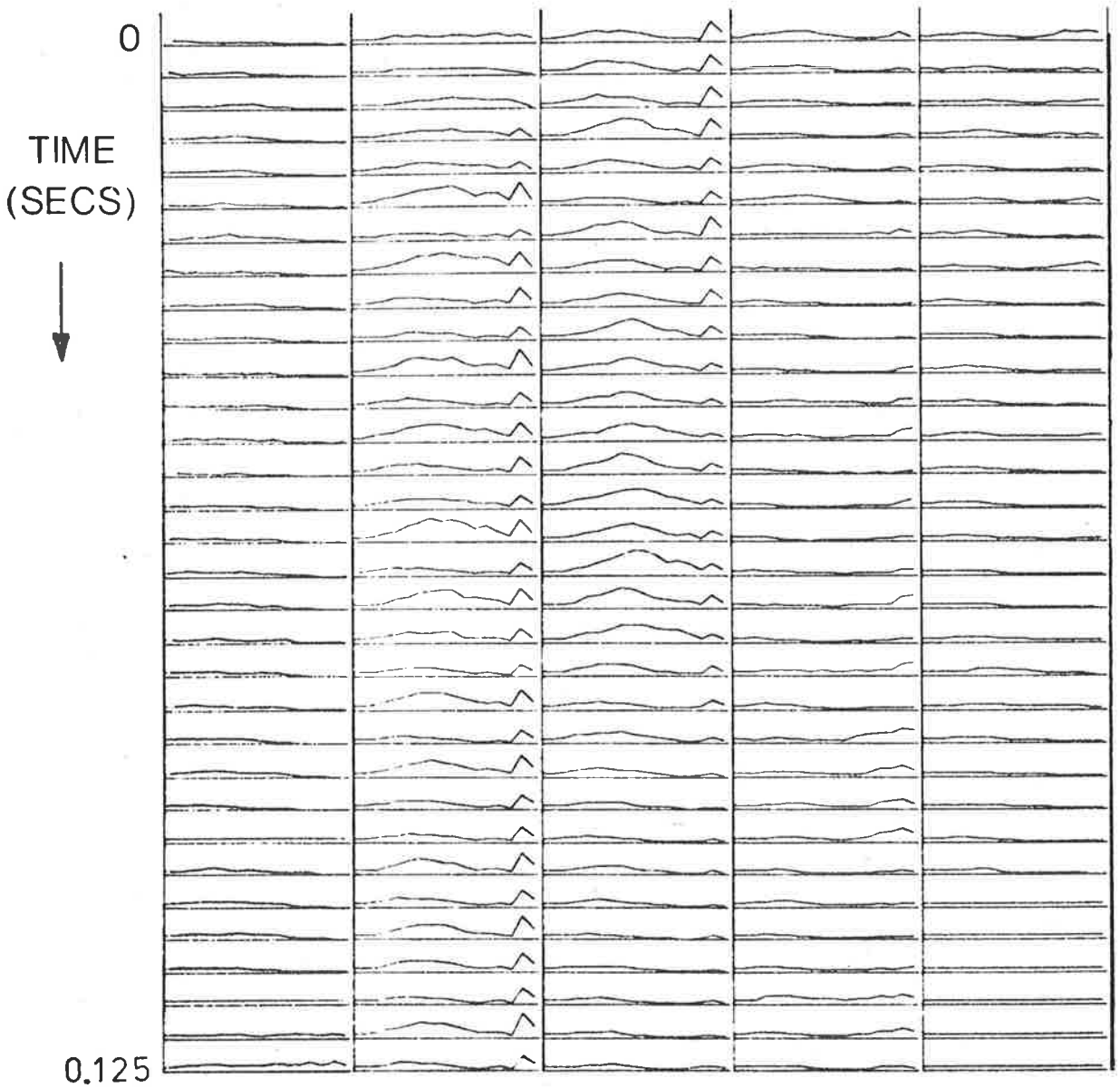


Figure 4.8 VOCAL TRACT AREA FUNCTION ESTIMATES OF 'BEAT' : VARIABLE γ

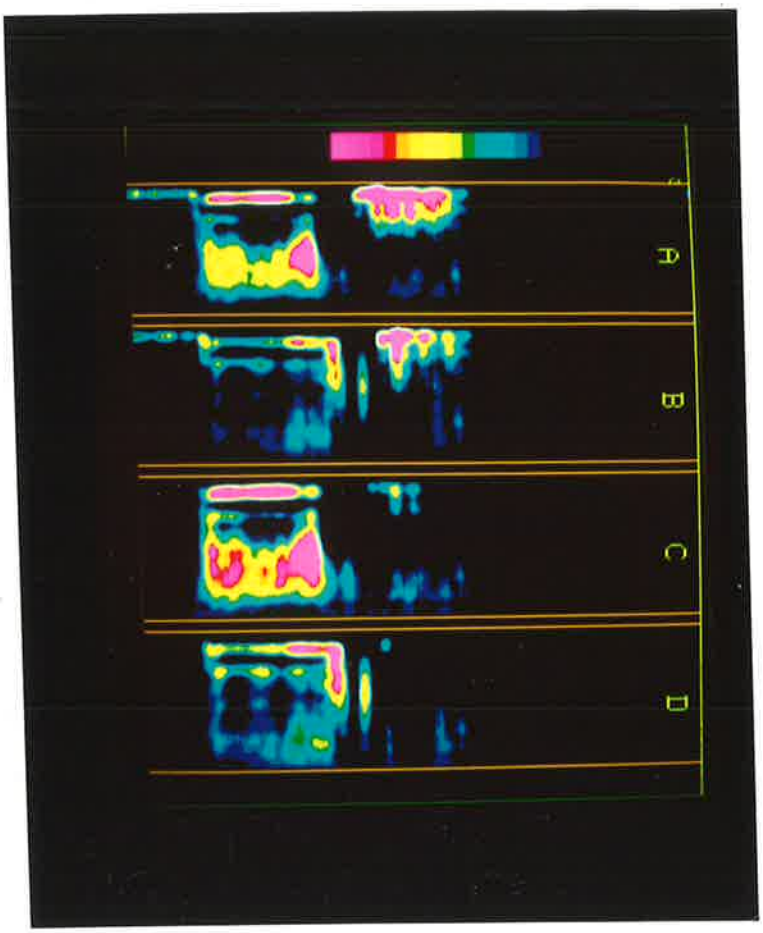


FIG.4.9 Subpictures A and B have constant pre-emphasis (γ=1). Subpictures C and D have been reprocessed with variable pre-emphasis.

CHAPTER 5

PROCESSING AREAGRAMS AS DIGITAL PICTURES

In the previous two chapters the types of processing described have treated speech in the conventional way, that is as a one-dimensional time series. However once VTAF estimation has been performed the speech is coded as a two dimensional function of time and vocal-tract position. This information is, of course, what is displayed as an areagram and the present chapter describes the effects of various image processing techniques on the areagram. Thus if the previous work is regarded as a pre-processing this can be regarded as a post-processing.

The types of processing to be discussed include image interpolation, linear filtering, peak and trough enhancement and grey level/colour assignment. Interpolation is necessary to provide a smooth looking display and linear filtering will enhance various spatial frequency bands of the image. Peaks and troughs have articulatory significance and enhancement of them is often desirable. Assigning grey levels or colours is very important in optimising information transfer to the observer. However, in order to minimise processing, it is desirable to do this as autonomously as possible.

Of course the spectrogram is also a digital image and many of the above techniques can be directly applied to it. Examples of this are also given.

Figure 5.1 shows the flow chart of the various stages of module 5, the picture processing portion, of the software. These stages, which may be applied to either the spectrogram or areagram, are performed in the following order:

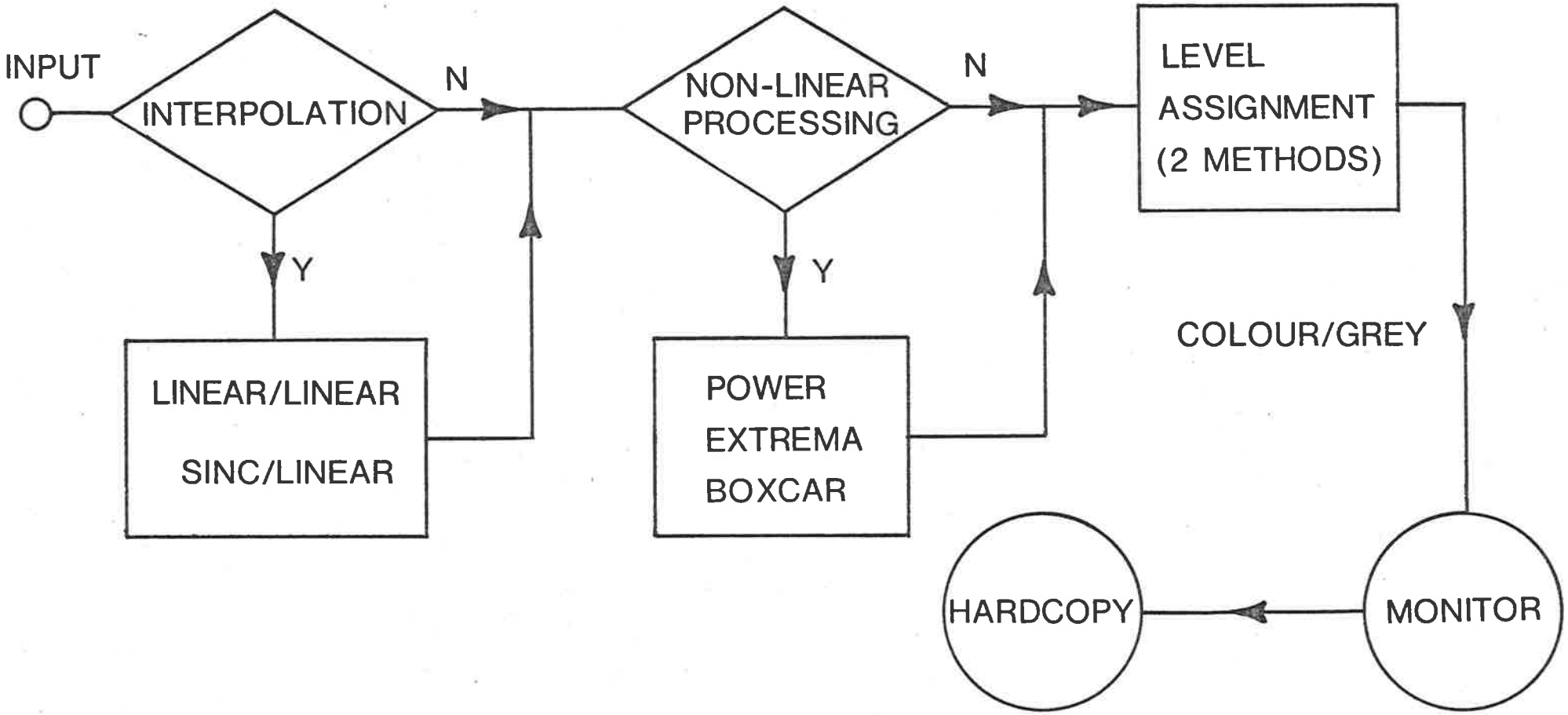


Figure 5.1 IMAGE PROCESSING AND DISPLAY

1. Interpolation
2. Non-linear processing
3. Linear filtering
4. Grey level/colour assignment

Each stage operates on the output image from the previous stage. Thus if peak and trough enhancement is performed in stage 2 linear filtering may be applied to this enhanced image in stage 3. However none of the first three stages are obligatory, and all may be bypassed. These stages will now be discussed in the above order.

5.1 IMAGE INTERPOLATION

For convenience the notation p(u,v) will be used to denote the non-interpolated raw image (areagram or spectra). The values u and v are discrete and refer to the horizontal and the vertical axes of the display. Thus for both types of pictures u is time. For VTAF's v is distance from the glottis and, for spectra, it is frequency. It was intended to display these images on a Lexidata 3400 image display system having a memory capability of 640 (horizontal) by 512 (vertical) picture elements or 'pixels'. It is necessary to decide how long a segment of speech to display on this system. Several different sizes were tried experimentally until a final display of 4 seconds of data arranged in four one second subpictures was chosen. Each subpicture is thus comparable to one standard Sonagraph sheet [20]. The number of points to be plotted in each subpicture depends on sampling frequency f_s, computation interval N_c, number of inverse filter coefficients N_f (for the VTAF), and window length N_w (for the spectrum). Table 5.1 shows the size of each subpicture for the usual values of these parameters.

TABLE 5.1

f_s	N_c	N_f	N_w	Areagram	Spectrogram	Subpicture
16384 Hz	64	15	128	256 x 16	256 x 64	512 x 128

The last column of this table refers to the number of pixels available by splitting the Lexidata display into four subregions. This was most conveniently done in practice by plotting time in the vertical direction and the distance/frequency in the horizontal. However as spectrograms (and areagrams) have time as the horizontal axis all images from the display are rotated by 90° for viewing. In the following discussion this rotation is presumed and references to the horizontal, or x, axis refer to time and the vertical, or y, axis to distance/frequency.

From table 5.1 it is clear that the interpolation for the areagram is more dense than for the spectrogram, at least for the vertical axis. That is a 16 to 128 point interpolation for the former as opposed to 64 to 128 for the latter. However it should be noted that because of the limited information in spectrograms above 6 kHz often only the lower 75% of the available frequency bins were used. The vertical interpolation was then from 48 to 128 points. The stronger interpolation required for the VTAF's can cause some display problems as discussed below. Two types of interpolation were investigated, a linear/linear scheme and a sinc function/linear scheme.

The first scheme is the most straightforward and uses linear interpolation in both y and x directions. Thus interpolation from the given function

p(u,v) to the display function g(x,y) is given by the following scheme:

(1) For each pixel (x,y) in the interpolated image evaluate:

$$u' = x\, N_u/N_x \tag{5.1}$$

and

$$v' = y\, N_v/N_y \tag{5.2}$$

where $N_u \times N_v$ is the size of the original subpicture and $N_x \times N_y$ the size of the interpolated image.

(2) Find the four nearest neighbours of the point (u',v'). Thus if u' and v' are non-integers then these can be found by truncating u' and v' to get the lower integers u_1 and v_1, and getting the upper integers u_2 and v_2 by adding 1 to these values. The four neighbours are thus at (u_1,v_1), (u_1,v_2), (u_2,v_1) and (u_2,v_2).

(3) Use the following formula to evaluate the interpolated value g(x,y):

$$\begin{aligned} g(x,y) = {} & p(u_1,v_1)(1-\delta_u)(1-\delta_v)+p(u_1,v_2)\,\delta_u(1-\delta_v) \\ & + p(u_2,v_1)\,\delta_v(1-\delta_u)+p(u_2,v_2)\,\delta_u\delta_v \end{aligned} \tag{5.3}$$

where

$$\delta_u = (u-u_1)$$

and

$$\delta_v = (v-v_1) \tag{5.4}$$

Figure 5.2 (A and B) show the results of this linear interpolation on the areagram of the utterance. 'We were away a year ago.' Figure 5.3 (A and B) show the same interpolation on the spectrogram of the same utterance.

This type of interpolation can be implemented in a quick and efficient

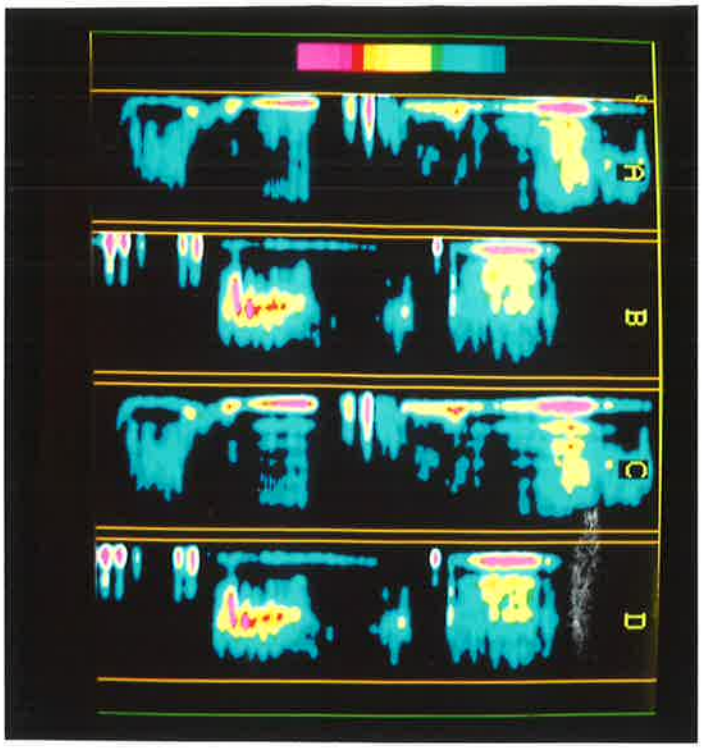


FIG.5.2 Areagram: A and B show linear/linear interpolation on the phrase 'We were away --'. C and D have been reprocessed with sinc/linear interpolation.

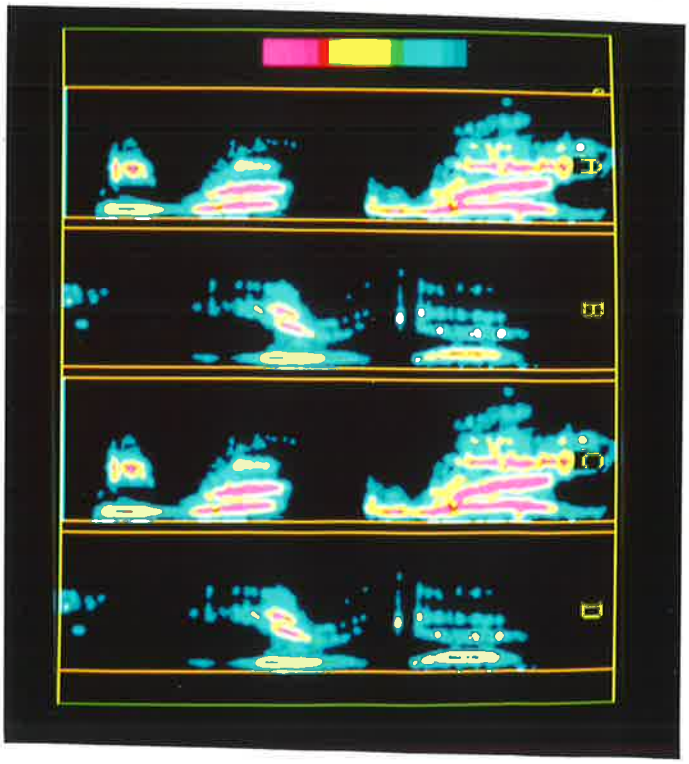


FIG.5.3 Spectrogram: A and B show linear/linear interpolation on the phrase 'We were away --'. C and D have been reprocessed with sinc/linear interpolation. Maximum frequency is 6kHz.

manner and is quite effective for most purposes. The process is equivalent to convolving the original picture with a triangular window, of width twice the sampling interval,in the u and then the v directions; thus obtaining a continuous function which is then resampled at the new pixel separation. This process is illustrated, in one dimension, in figure 5.4. One possible deficiency of linear interpolation is seen from examining this figure. This is the fact that the original sample points will always be the maximal values in the new picture. This can be seen more formally from equation (5.3). Thus if p(u',v') is the maximum of the four 'neighbours' of (u,v) then:

$$g(x,y) < p(u',v') \left[(1-\delta_u)(1-\delta_v) + \delta_u(1-\delta_v) + \delta_v(1-\delta_u) + \delta_u\delta_v\right]$$

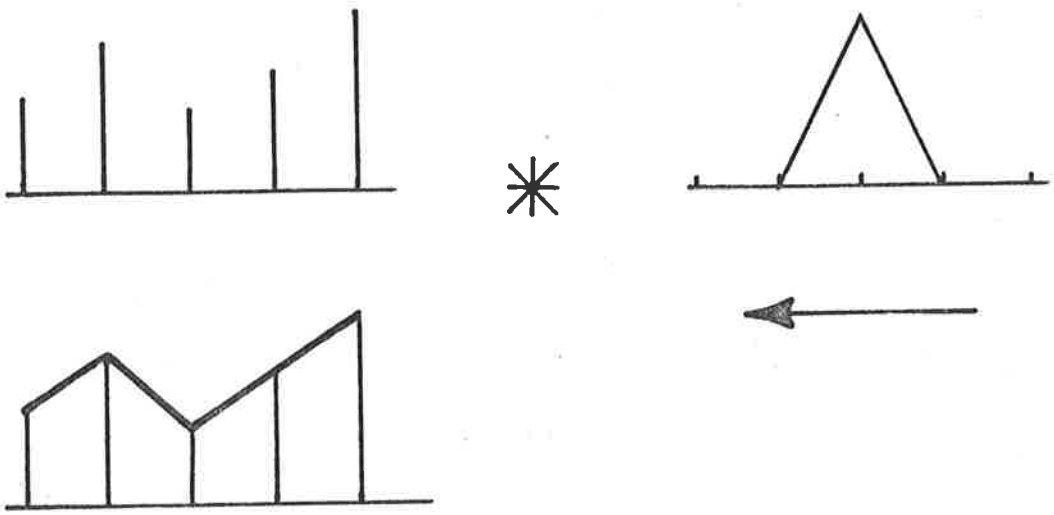
$$\equiv p(u',v') \qquad (5.5)$$

Similarly it can be that if p(u*,v*) is the lowest of the neighbours then:

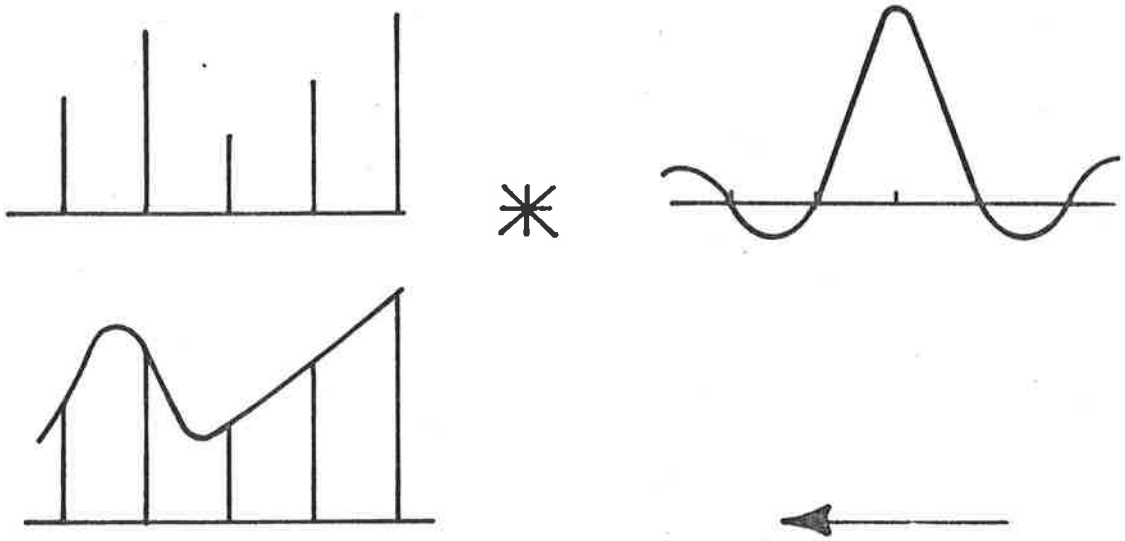
$$g(x,y) > p(u^*,v^*) \qquad (5.6)$$

These results mean that changes is local extrema positions will not appear smooth but rather as a series of discontinuous 'jumps'. Some evidence for this is seen in figure 5.2 but a clearer example will be discussed in the next section.

For spectra this is usually no real problem, as an interpolation ratio of only 1:2 in both directions is necessary. (However, even here, for a formant tracking requirement it may not be desirable to use this linear interpolation.) As an interpolation ratio of 1:8 is needed for the areagram in the y direction, then problems can clearly arise if local



LINEAR INTERPOLATION



SINC INTERPOLATION

Figure 5.4

extrema are to be emphasised or tracked. An alternative scheme is used for such cases.

As the areagram only needs a 1:2 interpolation in the x direction the linear scheme is usually quite satisfactory. In the y direction a sinc function interpolation based on Whittaker's 'cardinal function' interpolation [21] was investigated. Here, instead of convolution with a triangular function, convolution with a sinc function is performed.

Thus if the equivalent position v' is obtained from (5.2) then the interpolated value at y is [22]:

$$g(y) = \sum_{n_v=0}^{N_v-1} p(n_v) \frac{\sin \pi(v'-n_v)}{\pi(v'-n_v)} \qquad (5.7)$$

This can be shown to be [23] given by:

$$g(y) = \frac{\sin \pi v'}{\pi} \sum_{n_v=0}^{N_v-1} \frac{p(n_v)(-1)^{n_v}}{v'-n_v} \qquad (5.8)$$

Equation (5.8) should not be evaluated for $v' = n_v$ but rather the substution $g(n_v) = p(n_v)$ made.

This interpolation, especially in the form of (5.8), can be made fairly quickly but takes 2 to 3 times as long as the linear scheme. Figure 5.2 (C and D) show this interpolation used for an areagram and figure 5.3 (C and D) for a spectrogram. Little difference is apparent for the spectrogram but the movement of extrema is more realistic for the area function. This method was thus used as standard for the areagram interpolation.

It should be noted that sinc interpolation in both directions was investigated experimentally but was not found to give any obvious advantage. In view of the increased processing time it was, therefore, not used as the standard technique.

5.2 NON-LINEAR PROCESSING

The simplest type of non-linear processing is replacing each data point by a power of itself. That is g(x,y) is mapped onto $g_p(x,y)$ by:

$$g_p(x,y) = [g(x,y)]^p \tag{5.9}$$

For p>1 this increases the dynamic range of the data and compresses it for p<1. If p = 2 and g is the VTAF then g_p becomes the true area. Figure 5.5 shows the same utterance processed with p = 0.5(A), 1.0(B), 1.5(C) and 2.0(D). For p>1 data overflow can easily occur due to finite word length and must be guarded against. In general p = 1 was found to give the most pleasing display for areagrams. For spectrograms a higher value produces formant enhancement which may sometimes be desirable.

Considerable effort has been expended investigating techniques for enhancing local maxima and minima. This 'extrema enhancement' is important for various reasons. The case of formant analysis has already been mentioned and in this case the enhancement can be seen as a preliminary to formant extraction [24]. In the case of the areagram these extrema correspond to constrictions and openings of the vocal tract and are thus important in classifying speech. The tongue hump position is particularly important in this regard [25]. Another reason for the interest in extrema enhancement is the possibility of extrapolating the tongue hump position, for example, into regions where the linear prediction model does not hold. The idea being that if the articulatory parameters can be extrapolated into fricatives etc then it may be possible to recognise them from prior knowledge of their articulation.

Many different schemes were tried to produce peak and trough (i.e. extrema) enhancement of the areagram and the results were usually disappointing. For example a multiplicative scheme which replaced each pixel by the product of itself and the maximum neighbour to either side in

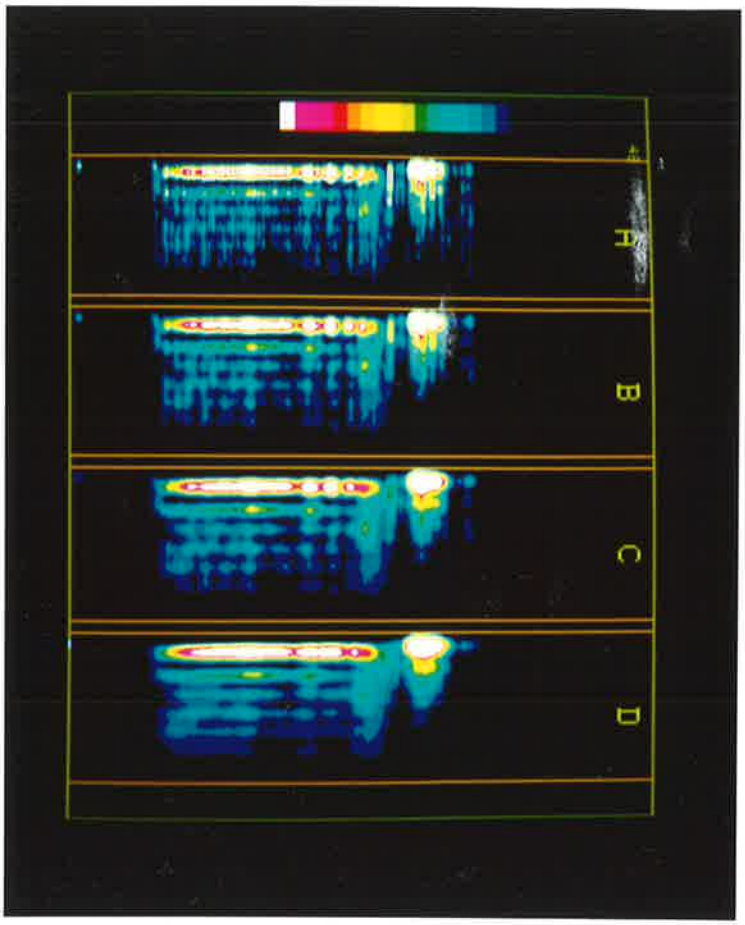


FIG.5.5 The values of area for the phrase 'beat' have been raised to the power 0.5(A), 1.0(B), 1.5(C), 2.0(D).

the x direction had been found to work very well on oceanographic data [26]. However for the areagram the results were very inconsistent. After considerable experimentation variations on a simple basic algorithm were developed and found to give satisfactory results.

The basic one provides peak enhancement/trough suppression as follows.

(1) Detect local maxima or minima within a window of size k_W in the y direction.

(2) If g(x,y) is a local maximum then g(x,y) → 2 g(x,y).
If g(x,y) is a local minimum then g(x,y) → g(x,y)/2.
Else g(x,y) is unchanged.

Figure 5.6 shows this algorithm applied to an areagram. The subpictures A and B show a linear/linear and C and D sinc/linear interpolation. The 'jumpy' artifact is quite clear in the first two.

The second method is a variation on the above. Thus after finding a local maxima at (x,y) then, as before, replace this by 2 p(x,y). However also replace p(x,y-1) and p(x,y+1) by 1.5 p(x,y-1) and 1.5 p(x,y+1) respectively. Similarly the adjacent pixels to a minima are replaced by 0.75 times their initial value. As seen in figure 5.7 (A and B) this leads to a broadening of the peaks and troughs. Figure 5.7 (C and D) shows another variation on this method. This time all non-extrema are replaced by zeros and the extrema, and adjacent pixels, are unchanged. Figure 5.8 shows the effects of similar processing on a spectrogram.

5.3 LINEAR 'BOXCAR' WEIGHTING

This technique is simply the convolution of the image with a uniform rectangular window. This was done to smooth out variations particularly in the time direction.

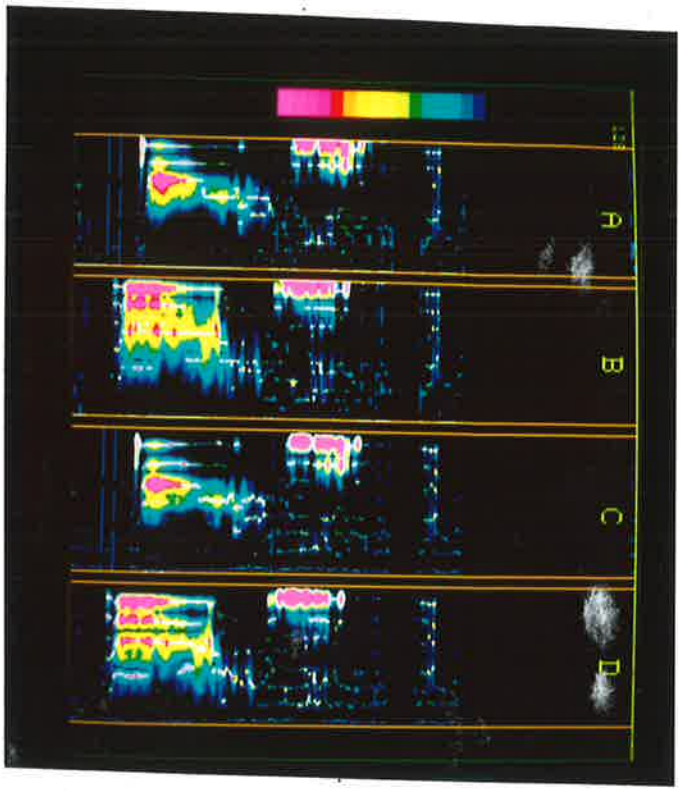


FIG.5.6 The utterance 'beat - boot' processed with the first extrema enhancement technique. A and B have been linearly interpolated and C and D by sinc/linear interpolation. The 'jumps' seen in A and B are much less marked in C and D.

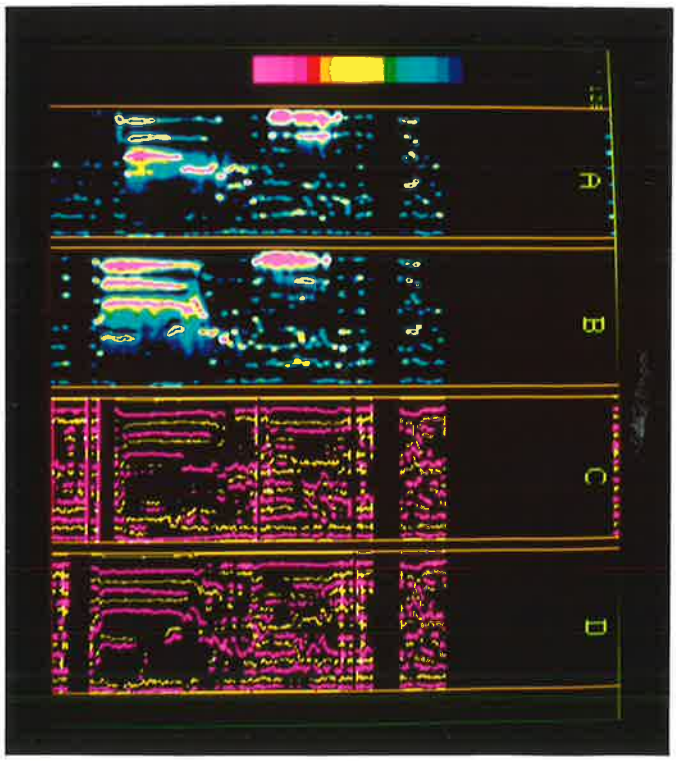


FIG.5.7 A and B show the results of the second extrema enhancement technique. C and D is a variation on this as described in the text.

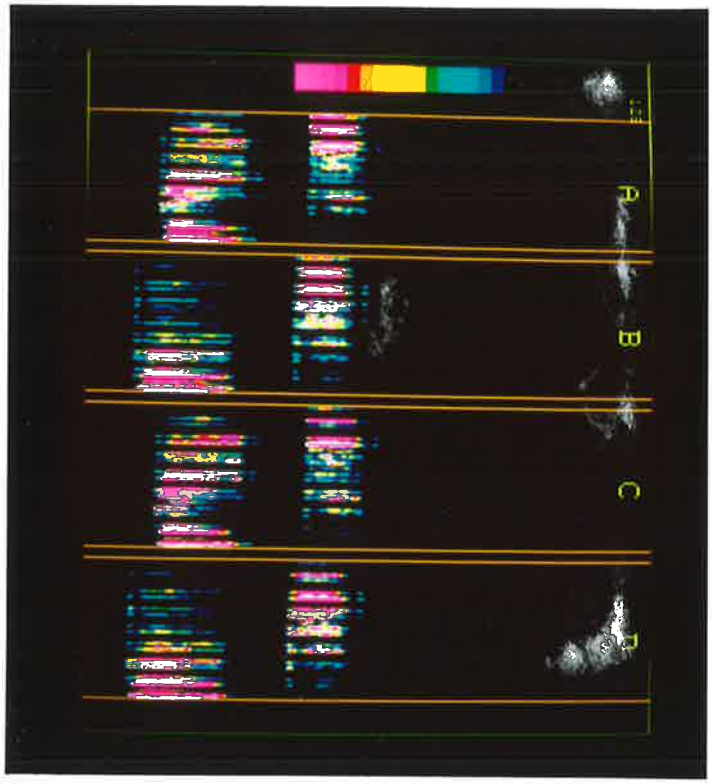


FIG.5.8 The spectrogram equivalent of figure 5.6. There is no obvious advantage in sinc interpolation.

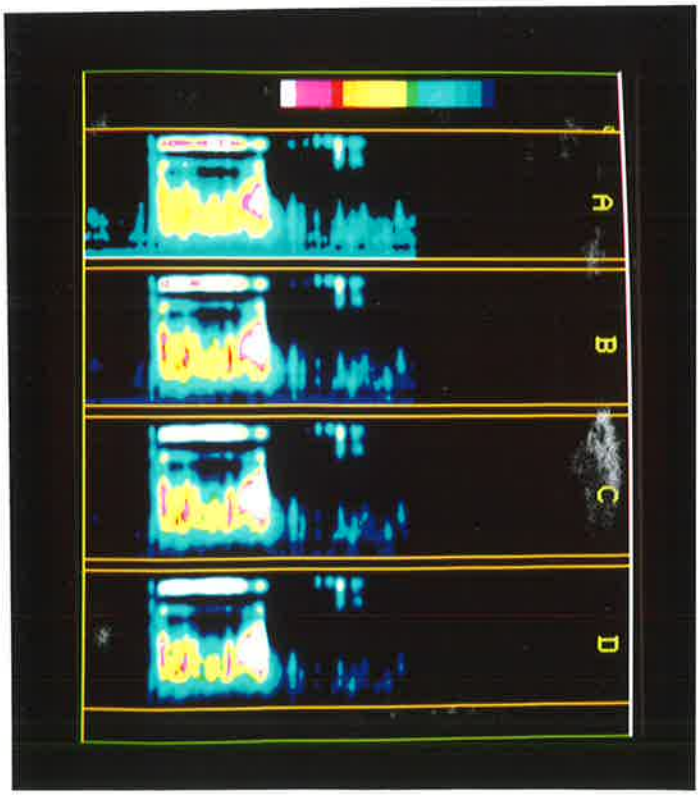


FIG.5.9 The utterance 'bart' processed with 4 different boxcar windows. These are A:1x1 B:3x1 C:5x3 D:9x5 .

Thus g(x,y) is replaced by $g_s(x,y)$ where:

$$g_s(x,y) = \sum_{y'=y-N_y}^{y+N_y} \sum_{x'=x-N_x}^{x+N_x} g(x',y')/N_s \qquad (5.10)$$

where the window is of size $2N_x+1$ by $2N_y+1$. N_s is the total number of pixels in this window. Equation (5.10) is equivalent to multiplying the spectrum of g_s by a two dimensional sinc function and is thus a crude form of low-pass filtering. Thus if g_s is subtracted from g a form of 'high pass' filtering is obtained. That is:

$$g_h(x,y) = g(x,y) - g_s(x,y) \qquad (5.11)$$

Figure 5.9 shows the effects on areagrams of four different sized windows. It is interesting to compare these with spectrograms smoothed by the similar windows. It appears that the spectral estimates are more stable than the VTAF estimates. This has a significance that will be discussed in the next chapter. Finally it should be noted that, high pass processing was found to be not very useful as the results were very noisy. This is unfortunate and contrary to the experience in other areas of image processing [27].

5.4 GREY LEVEL/COLOUR ASSIGNMENT

The dynamic range of areas or spectral levels can be high. However the human observer can only distinguish a comparatively small number of grey levels or colours [28]. Accordingly some way must be found to map the number of values of p(x,y) onto the 16 or so grey levels/colours which the human can assimilate. Two assignment schemes were investigated. This first was the 'minimum entropy loss' technique which had previously been found useful in enhancing thermal structure in infrared images of the ocean

taken by satellite [29]. This method works by equalizing the histogram of the displayed images grey levels. That is, it chooses thresholds so that (approximately) the same number of pixels fall into each grey level bin. A fast algorithm to implement this was developed as follows:

(1) Define an array KLEVEL of dimension the same as the number of levels in the (stored) picture data. Typically this would be 1024.

(2) Count the number of data points in each bin. This can be done quickly because, assuming all data falls in the range 0-1023, then a pixel of value K can be counted by simply incrementing the array element KLEVEL(K).

(3) It is known how many pixels L_B in all are to be displayed and therefore the required number in each bin of the displayed picture. For an M X N picture this is simply MN/K_D where K_D is the number of grey levels or colours available. Having determined this number the thresholds α_1, α_2 ..... α_{D-1} can be found by adding adjacent bins of KLEVEL (starting at K = 1) until L_B is reached. The first value of K where this occurs is then used as α_1. This process is then repeated for all other thresholds. There will not usually be exactly L_B pixels in each bin but this method leads to a good approximation.

Grey level or colour assignment is then given by the following algorithm.

$$g(x,y) \rightarrow \text{level } k \text{ if } \alpha_{k-1} < g(x,y) \leqq \alpha_k \qquad (5.12)$$

The reason for the name of this process ('minimum entropy loss') is that it can be easily shown that by equalizing the displayed histogram the first order entropy of the displayed image is maximised [29].

The second technique used was equi-level thresholding, which is simply

setting the thresholds α at equal intervals. Unlike the previous scheme this results in a histogram re sembling the input histogram. The two processes are illustrated in figure 5.10.

For the second scheme the thresholds are related ($1 < k < K_D$) by:

$$\alpha_k - \alpha_{k-1} = \varepsilon \text{ (a constant)}$$

Usually this kind of technique involves considerable operator intervention to obtain a satisfactory value of ε. However it was found that a satisfactory ε could be found autonomously for both kinds of data. This was done by computing the mean μ and the standard deviation σ of the picture. The silences between utterances were ignored in this computation. The interval ε can then be set such that the available grey levels/colours fall equally between the values μ and $\mu + 4\sigma$.

$$\begin{aligned} \text{Thus} \quad & \varepsilon = 4\sigma/K_D \\ \text{and} \quad & \alpha_k = \mu + (k-1)\varepsilon \end{aligned} \qquad (5.13)$$

Figure 5.11 shows the same areagram processed with both techniques. Both give satisfactory results but as the second is easier to implement this was nearly always used. The effectiveness of (5.13) suggests that the VTAF's and spectra are close to Gaussian distribution.

Finally it will have been noted that colour, rather than grey level, has usually been used to display the areagram in this study. This is because this results in a more pleasing display. Thus grey level areagrams appear blurred and lack definition. Some evidence that this is a significant and not just a cosmetic effect is given in the next chapter. Figures (5.12) and (5.13) show grey level areagrams of differing separation between grey levels. By dropping some grey levels (figure 5.13) a sharper, less 'blurred', picture results. The same effect does not occur with spectrograms, and colour does not appear to add a great deal to these.

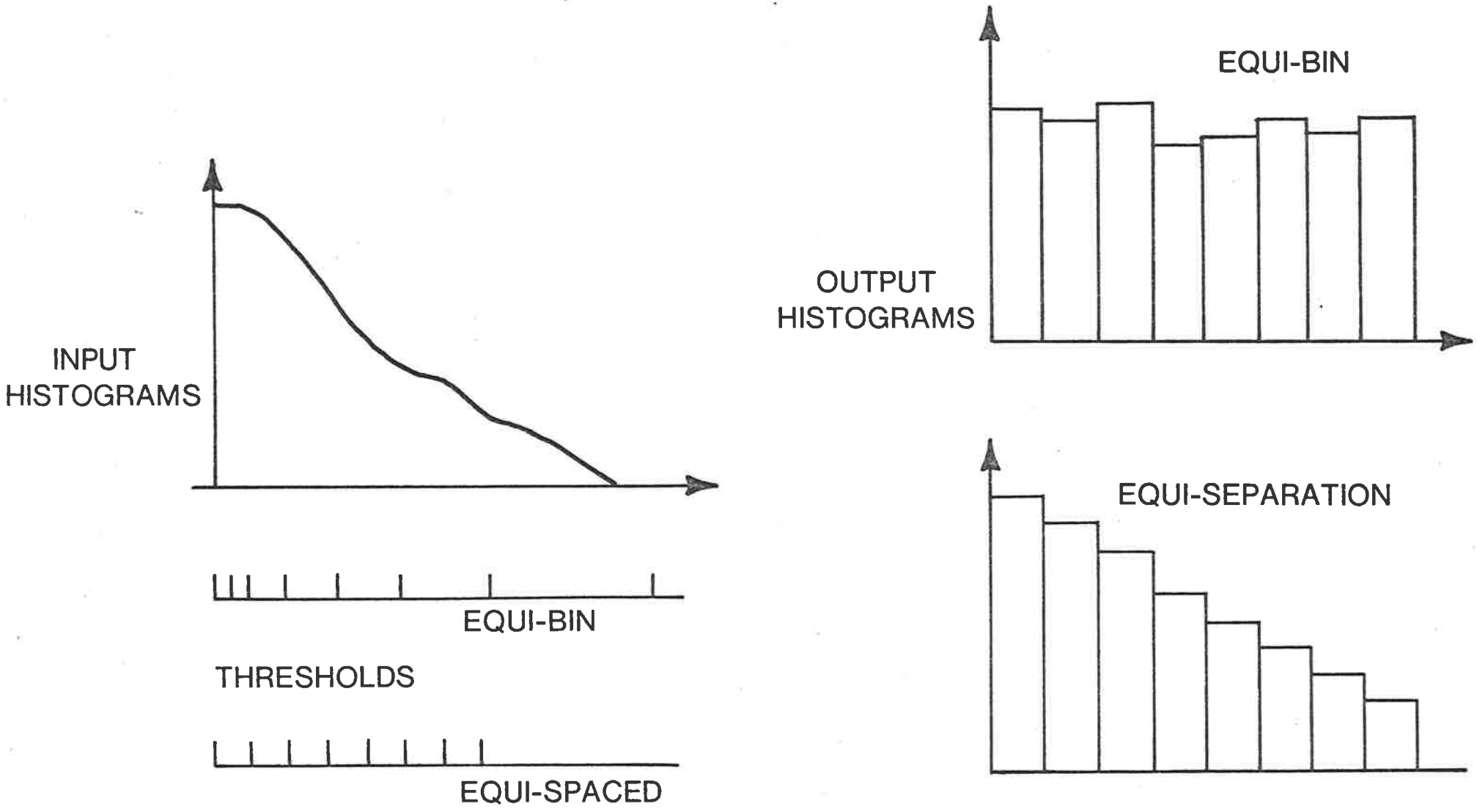


Figure 5.10 GREYLEVEL ASSIGNMENT

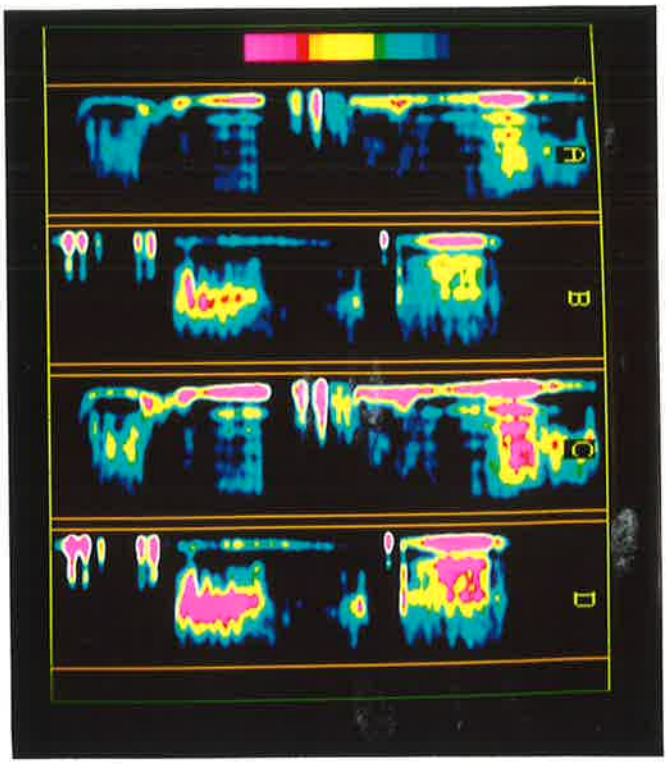


FIG.5.11 Part of 'we were away--' processed by different grey-level /colour assignment techniques. A and B are equi-level and C and D equi-bin.

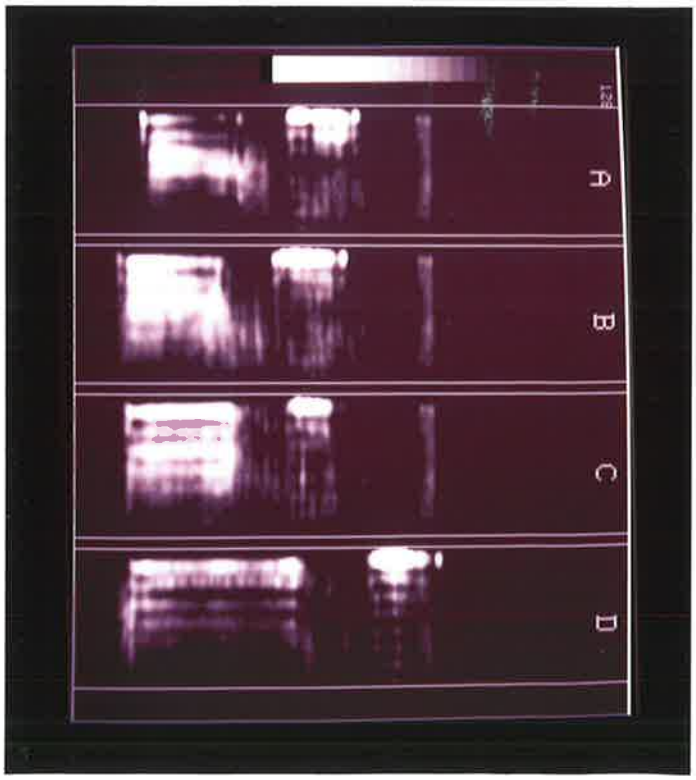


FIG.5.12 'Standard' grey level areagram with 18 levels.

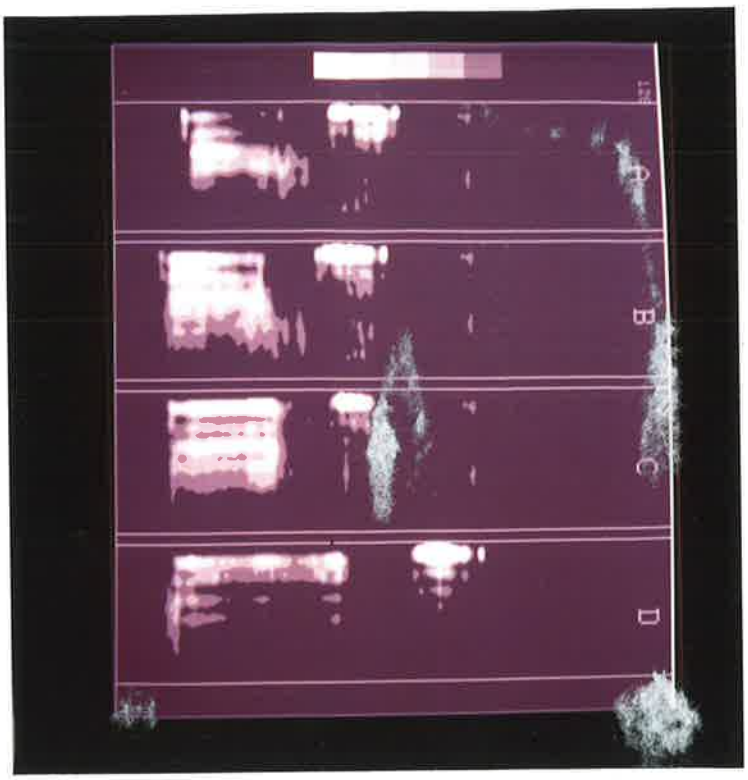


FIG.5.13 Same as figure 5.12 but with only 5 levels. Some improvement is apparent.

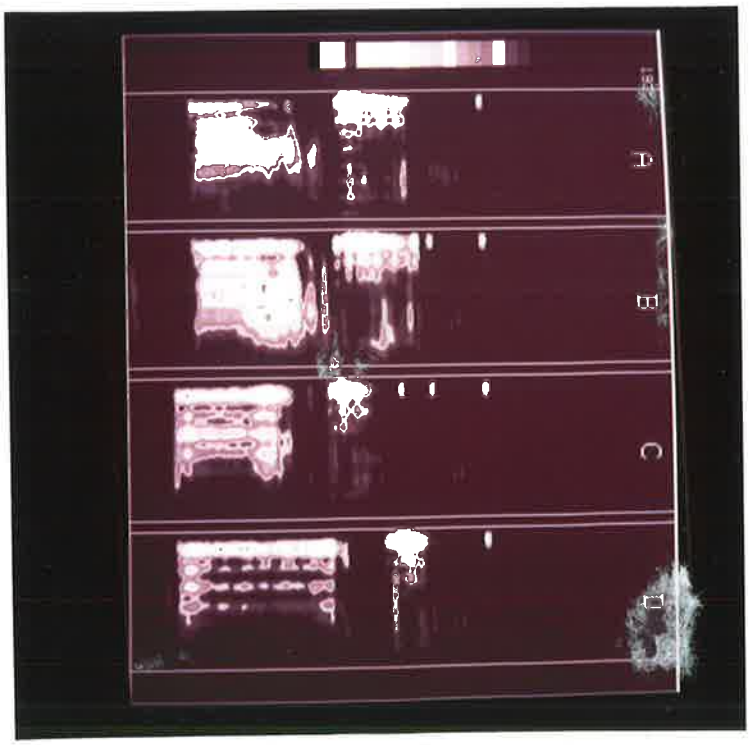


FIG.5.14 Some discontinuity in the grey scale leads to an 'outlining' of structure.

These aspects are discussed further in the next chapter.

Figure 5.14 illustrates a different approach to the grey level areagram problem. Here several discontinuities have been introduced into the grey scale and this leads to a quite effective 'outlining' of structure.

CHAPTER 6

VISUAL RECOGNITION OF THE AREAGRAM AND SPECTROGRAM

There are a number of reasons which, potentially, make the VTAF a suitable tool for the phoneme pattern recognition task; that is for the classification, by man or machine, of individual phonemes. These include the fact that the pitch component of speech has been removed, and thus so has an important part of the difference between male, female and child speech. Pitch in itself does not contribute to phoneme recognition in English [30]. What should be then left is purely articulatory information, which is more or less constant for speakers with similar accents. Thus vowels are classified by phoneticists [31,32] as being frontal or back, closed or open according to the behaviour of the vocal tract and the position of the tongue hump. The areagram displays this information in a visible form and the hope was that visual classification of the utterance may be possible. To test this idea a data-base of simple CVC (consonant-vowel-consonant) utterances was built-up and a randomised visual presentation technique used to evaluate the visual classification potential of the areagram. For comparison a database of spectrograms of the same CVC utterances was assembled and tested. The opportunity was also taken to compare the recognisability of both colour and grey-level areagrams and spectrograms. This is an important test in its own right with implications over a much wider field than the present visible speech study.

6.1 EXPERIMENTAL PROCEDURE

The time series database consisted of 4 repetitions of the CVC utterances 'beat', 'boot', 'bat' and 'bart' by 12 Australian speakers including two females. The vowels /i/, /u/, /ae/ and /a/ were chosen because they correspond to extreme positions in the vowel quadrilateral. Thus, as seen

in figure 6.1, /i/ is frontal and open, /u/ is back and open, /ae/ is frontal and closed, and /a/ is back and closed.

The utterances were digitised and edited as described in Appendix A. This edited database was used as the input to the subsequent processing. Details of the spectral analysis and VTAF computation are given in Table 6.1.

TABLE 6.1

Database For Classification Test

Number of speakers	: 12 (10 male, 2 female)
Number of CVC utterances	: 4 (repeated 4 times)
Sampling frequency	: 16384 Hz (8 kHz Nyquist)
Computation interval	: 64 samples
Window length	: 192 samples
Window shape	: Hanning
Inverse filter coefficients (VTAF)	: 16
Maximum frequency displayed (Spectra)	: 6.0 kHz

The picture processing performed on the resulting areagrams and spectrograms was minimal viz. standard interpolation and smoothing by a 7(horiz.) x 3(vertical) boxcar window. Grey level/colour assignment was by the equi-separation method and 18 grey levels/colours were used.

Two types of pictures were required for testing. These are the multi-utterance and single utterance areagrams and spectrograms. The multi-pictures have a complete set of the four CVC utterances from a single speaker and the single pictures have just one CVC utterance. This single utterance subpicture was placed in the same location on all single pictures. No utterance identification data was included in the single pictures. Figure 6.2 shows a multi-areagram and figure 6.3 a single-areagram from the same speaker but for a different repetition. For the

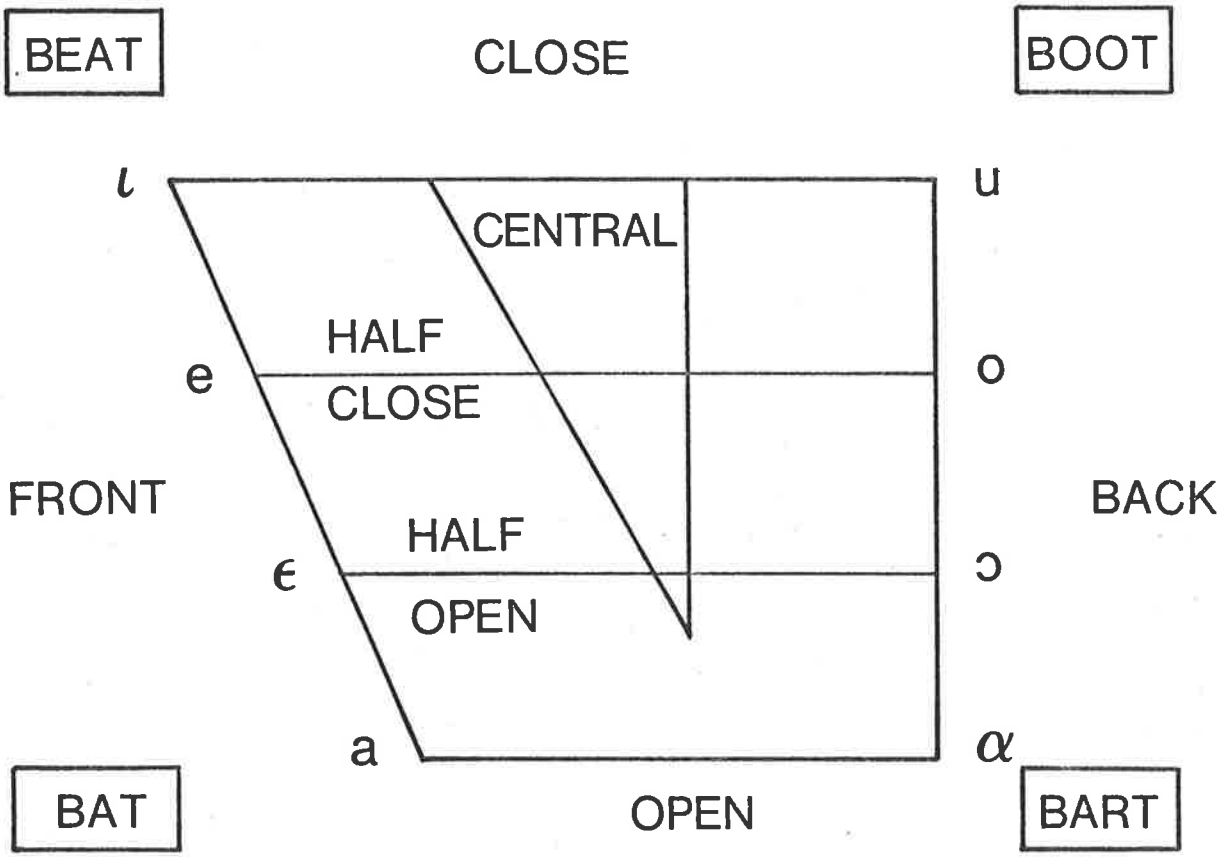


Figure 6.1 THE VOWEL QUADRILATERAL

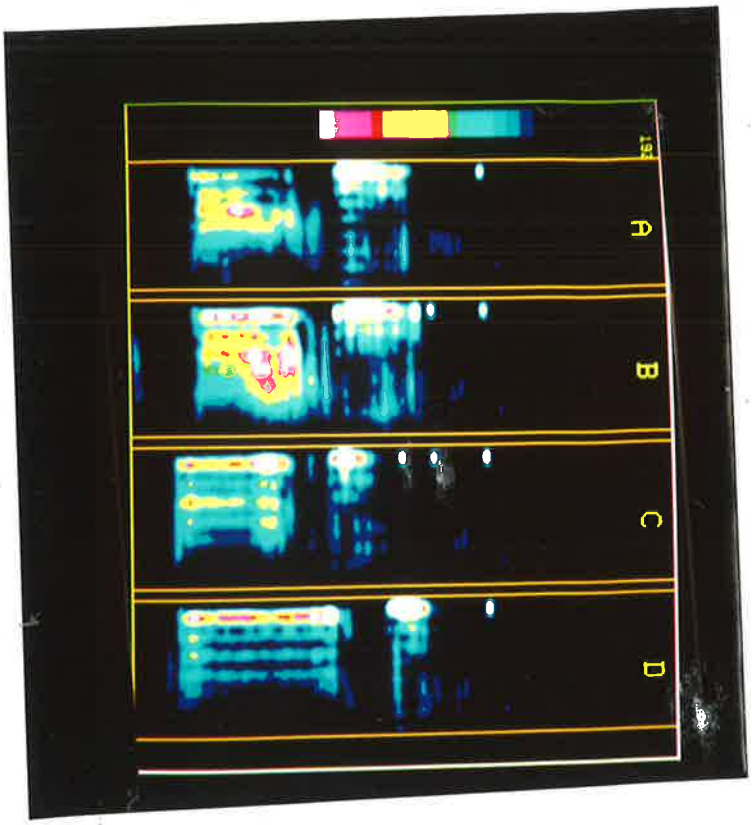


FIG.6.2 Multi-areagram of 'beat-boot-bat-bart'.

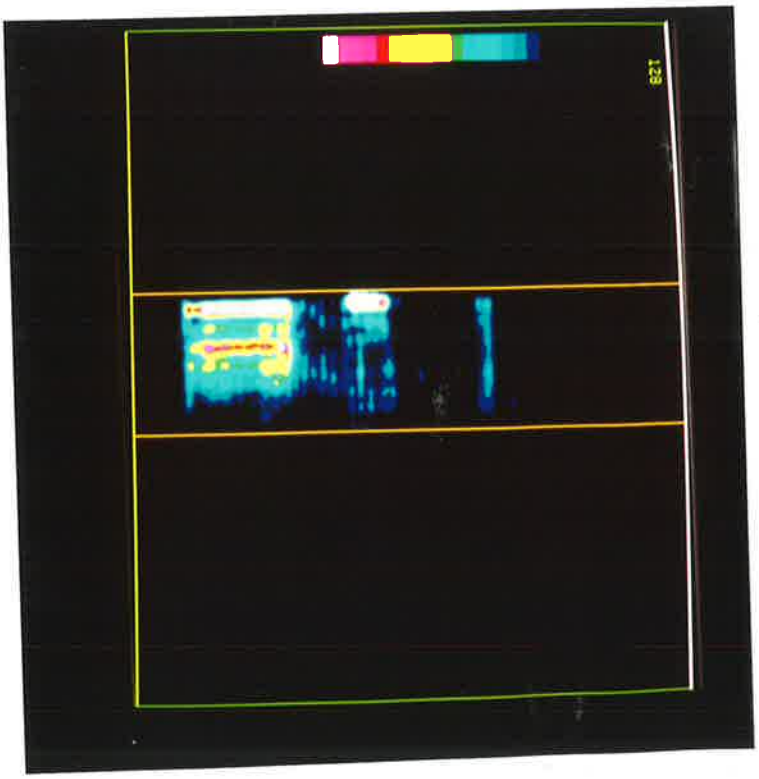


FIG.6.3 Single-areagram of 'bat' from second dataset.

comparison tests different repetitions were used for the multi and single pictures.

The test itself really consisted of four separate tests where the recognisability of grey areagrams, grey spectrograms, colour areagrams and colour spectrograms was measured. To minimise the time of each of these tests only data from 6 speakers was used, one of whom was female. For each test 6 multi-pictures was thus needed and 24 single pictures. After some experimentation prints rather than transparencies were used. This was logistically easier and reduced setting-up time. The test proceeded as follows:

(1) Each of the six subjects was given one of the multi-pictures.

(2) Each subject was handed one single picture. They then had to note on a standard form which of the four subpictures on their multi-picture this single picture most resembled. This was done by circling one of the letters A B C or D on the form. The subjects had to make one choice and could not leave blank answers.

(3) After 30 seconds a buzzer sounded and each single picture was handed on to the next subject.

(4) When each of the 24 single pictures had to be classified by all subjects they were mixed randomly and each multi-picture was passed to the next subject. The whole process then proceeded until all 24 single pictures had again been classified. This random mixing and single-picture classification was repeated a total of six times and thus each subject had the opportunity to classify each single picture against each multi-picture. Seating was arranged so that subjects could not see the results of their neighbour's classification.

The whole process is illustrated in figure 6.4. A single test (e.g. of grey spectrograms) took about 90 minutes. The total test program lasted

Each subject has one of six multi-image pictures. Twenty four single image pictures are rotated anti-clockwise every 30 seconds and classification (A, B, C or D) recorded.

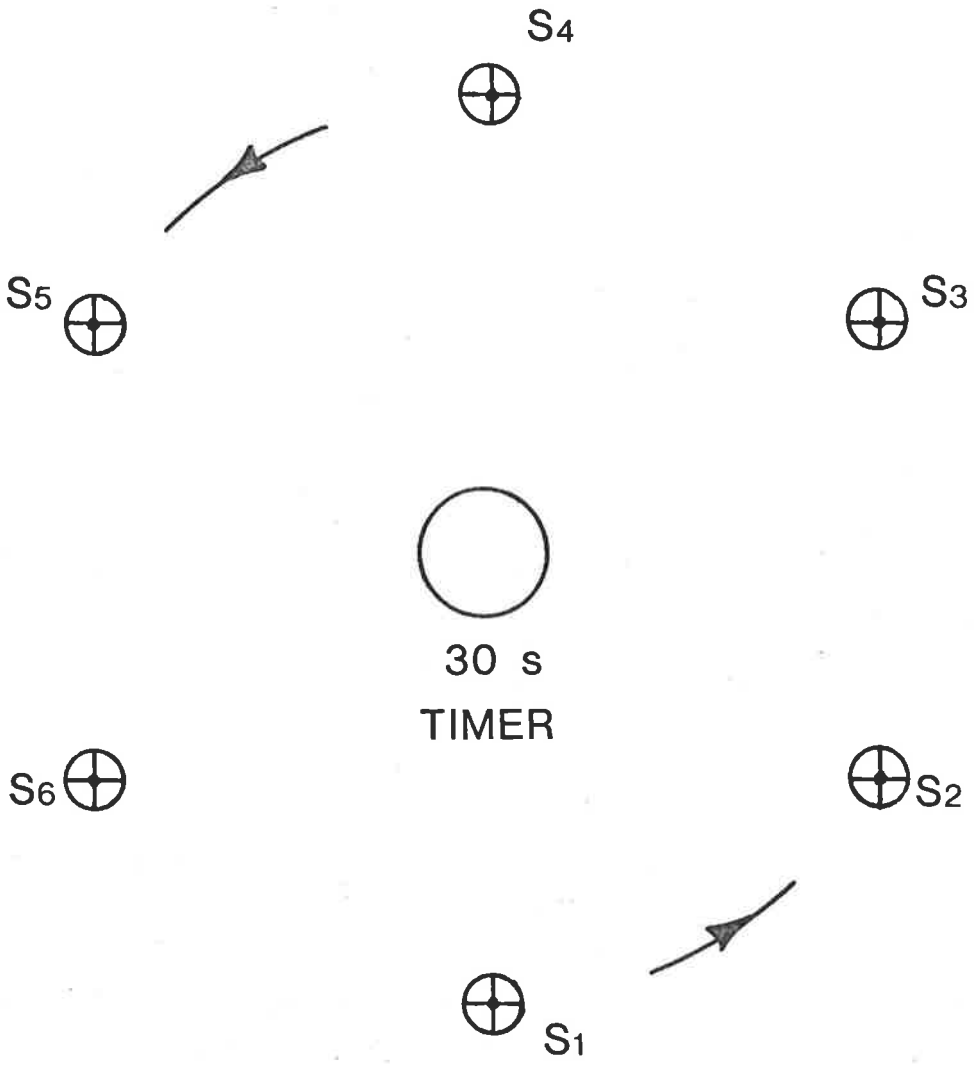


Figure 6.4 CLASSIFICATION TEST PROCEDURE

for 6 hours, and was spread over four days to minimise subject boredom. The order of carrying out these tests is shown in table 6.2.

TABLE 6.2

Day 1	Grey spectrograms
Day 2	Grey areagrams
Day 3	Colour areagrams
Day 4	Colour spectrograms

Before the start of the test program only a minimum amount of information was given to the subjects. They were simply told that they were taking part in a 'speech pattern recognition test' without any further explanation. None of the subjects had any familiarity with spectrograms or areagrams. A short 'dummy run' using another set of pictures was used before conducting the actual trial.

6.2 TEST RESULTS

During the test the actual order of the single pictures was noted by the experimenter. A cryptic identifying code on the back of the single pictures enabled him to do this. The look-up table to identify each picture was not known or available to the subjects.

All of the choices on the standard forms were entered into the computer, as were the random orderings of the single pictures.

In their most basic form the results of the tests are shown in table 6.3. This simply shows the number of correct and incorrect labels applied to the test utterances as an absolute number and also as a percentage. Results from all subjects are combined.

TABLE 6.3

LABEL ASSIGNMENTS

	Correct	Incorrect
Grey Spectrograms	646	216
Colour Spectrograms	666	218
Colour Areagrams	524	340
Grey Areagrams	395	469

Before assessing the differences between the classes the hypothesis of random result should be tested. That is did the subjects assign lables A, B, C and D randomly? Now events A, B, C and D occurred with equal frequency thus the probability of the subject guessing the label correctly is 0.25 on this hypothesis. The (null) hypothesis to be tested is then:

$$H_0: p = 1/4 \tag{6.1}$$

where p is the probability of success in each trial.

Now the type of experiment described here is of the binomial kind [33] as only two results are possible at each trial (success or failure). Also the probability of success p is constant at each trial and repeated trials are independent. The probability distribution of the binomial random variable X_1 the number of successes in n independent trials is [33]:

$$b(x;n,p) = \binom{n}{x} p^x(1-p)^{n-x} \tag{6.2}$$

An error of the first kind (type I) occurs when H_0 is accepted when it is false. The probability of this error is α and is given by:

$$\alpha = \sum_{x=X_o}^{n} b(x;n,p) \tag{6.3}$$

where X_o is the critical value. To evaluate α it is easier to use the normal distribution approximation to the binomial distribution. Thus for n trials X is approximately normally distributed with mean μ and variance σ^2 given by:

$$\begin{aligned} \mu &= n\,p \\ \sigma^2 &= n\,p(1-p) \end{aligned} \qquad (6.4)$$

Replacing X by the random variable Z, where:

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \qquad (6.5)$$

enables α to be obtained in terms of the standardized normal distribution $n(z;0,1)$. Thus:

$$\alpha \cong \int_{z=Z_o}^{\infty} n(z;0,1)\,dz \qquad (6.6)$$

where $Z_o = (X_o-\mu)/\sigma$. Choosing X_o equal to the values of successful trials in table 6.4 leads to the probabilities of type I errors for these data. As these errors are vanishingly small ($8.1\ 10^{-45}$ is the largest) they are not shown in table 6.4; rather than the α's the values of the Z_o's, in units of σ, are shown.

TABLE 6.4

H_o PROBABILITY OF OBSERVED RESULTS	Z_o
grey spectrogram	34σ
colour spectrogram	35σ
grey areagram	24σ
colour spectrogram	14σ

These values of α_{Xo} are so low that it appears H_0 can be rejected for all picture types. Thus all displayed images have highly significant (in the statistical sense) classification properties.

However the overall aim of this test is to compare the performance of the spectrogram recognition versus areagram recognition and of colour versus grey level pictures. This can be considered the problem of estimating the difference between various sets of two proportions. Thus it is intended to test:

H1: Proportion of correct grey spectrograms versus proportion of correct grey areagrams.

H2: Proportion of correct colour spectrograms versus proportion of correct areagrams.

H3: Proportion of correct colour spectrograms versus proportion of correct grey spectrograms.

H4: Proportion of correct colour areagrams versus proportion of correct grey areagrams.

Suppose that the recognition process is a binomial one with probability of being correct p_1 for one image type and p_2 for another. For samples of size n_1 and n_2 these distributions have means n_1p_1 and n_2p_2 and variances $n_1p_1(1-p_1)$ and $n_2p_2(1-p_2)$ respectively. Suppose the proportion of success in each sample is $\hat{p}_1$ and $\hat{p}_2$. It can be shown [34] that the confidence interval of $(1-\alpha)$ 100% for the difference of two binomial parameters p_1-p_2 is approximately:

$$(\hat{p}_1 - \hat{p}_2) - Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$< p_1 - p_2$$

$$< (\hat{p}_1 - \hat{p}_2) + Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \quad , \qquad (6.7)$$

where $Z_{\alpha/2}$ is the value of the standard normal curve leaving an area of $\alpha/2$ to the right.

For the present test it is arguable just how large the samples n_1 and n_2 actually are. This is because although 144 classifications are made by each observer only 24 independent subpictures are used. Randomizing the set of 24 was intended to alleviate this problem but to be conservative it will be assumed that only 24 independent labellings are done by each subject. Combining result for the six subjects gives 144 independent classifications per picture type.

This was done by dividing the results of each of the 144 classifications per subject by 6 and then adding these together. Thus the results of table 6.2 are, in effect, divided by 6 and $n_1 = n_2 = 144$. Choosing values of 95% and 99% for the confidence levels yields the intervals shown in table 6.5.

TABLE 6.5

Confidence Intervals

	$\hat{p}_1$	$\hat{p}_2$	p_1-p_2(95%)	p_1-p (99%)
H1:	0.747	0.457	0.183/0.398	0.149/0.432
H2:	0.771	0.606	0.059/0.270	0.026/0.302
H3:	0.771	0.748	-0.756/0.122	-0.106/0.152
H4:	0.606	0.457	0.035/0.263	0.000/0.299

6.3 TEST INTERPRETATIONS

Several obvious conclusions can be drawn from table 6.5. The grey spectrogram is significantly better than the grey areagram for this type of classification (H1). However this difference, whilst statistically significant, is not particularly large for areagrams and colour spectrograms (H2). There is no significant difference between colour and grey level spectrograms (H3) but a significant one between colour and grey level areagrams (H4). It is certainly clear that whatever improvement that may have arisen by removing pitch information in the areagram display is not apparent in this test. Perhaps the observed instability of the VTAF estimates (Chapter 5), compared with spectral estimates, has masked any benefit from the removal of pitch. There is some indication that this instability arises from VTAF computation and is not inherent to the linear prediction process. Thus spectra generated by linear prediction analysis are much more stable than their corresponding VTAF's (D. Fensome, private communication). Other workers have also noted instabilities in VTAF estimates [35,36].

The difference between grey level and colour VTAF's is very interesting and appear to be the first result of this nature reported. There is not a great difference between the 'channel capacity' of humans in grey levels or colours [28,37] so the result is surprising. One tentative explanation for this is that the well known intensity 'differencing' effect (known as 'lateral inhibition' in the psychology literature) is at work. This leads to a grey level wedge appearing to have bars ('Mach bands') instead of edges [38,39] and could be reducing the effectiveness of the slowly varying grey level areagram. As the spectrogram has much sharper peaks than the areagram this effect would here be much less marked, as is observed. In the areagram, by dropping out alternate grey levels, (section 5.4) an artificial partial compensation for this effect is introduced. Also the differencing is much less marked for colour wedges, which is also consistent with the observations.

CHAPTER 7

CONCLUDING REMARKS

The areagram has been proposed as a complementary display to the spectrogram to provide articulatory information to the observer in an easily absorbed format. It has been shown that the synoptic properties of this display enable processing deficiencies to be readily spotted. Using the Wakita algorithm the vocal tract areas can be quickly computed. However the effects of random noises and zeros in speech may produce gross VTAF errors. The former can be detected and suppressed but a more comprehensive speech model should be developed to accommodate the latter. Some attempts at such modelling have of course been made made [40,41] but these are very slow to compute. An efficient algorithm is needed. Reasonably fast interpolation and image processing techniques are available for the areagram. More work needs to be done on vocal tract constriction enhancement and tracking. It may be possible to extrapolate (forwards or backwards) into non-voiced speech sounds in this way. The visual classification performance of the areagram is disappointing, particularly for the grey-level display. A significant improvement may be made here by reducing instabilities in the VTAF estimates. This could be done by the use of pitch synchronous processing but at a considerable increase in processing time [35]. The significant increase in classification performance achieved by coding the areagram in colour is a most interesting result and should be investigated further.

In summary it is believed that the areagram is a useful and informative complementary display to the spectrogram and one which is worth further investigation and development.

APPENDIX A

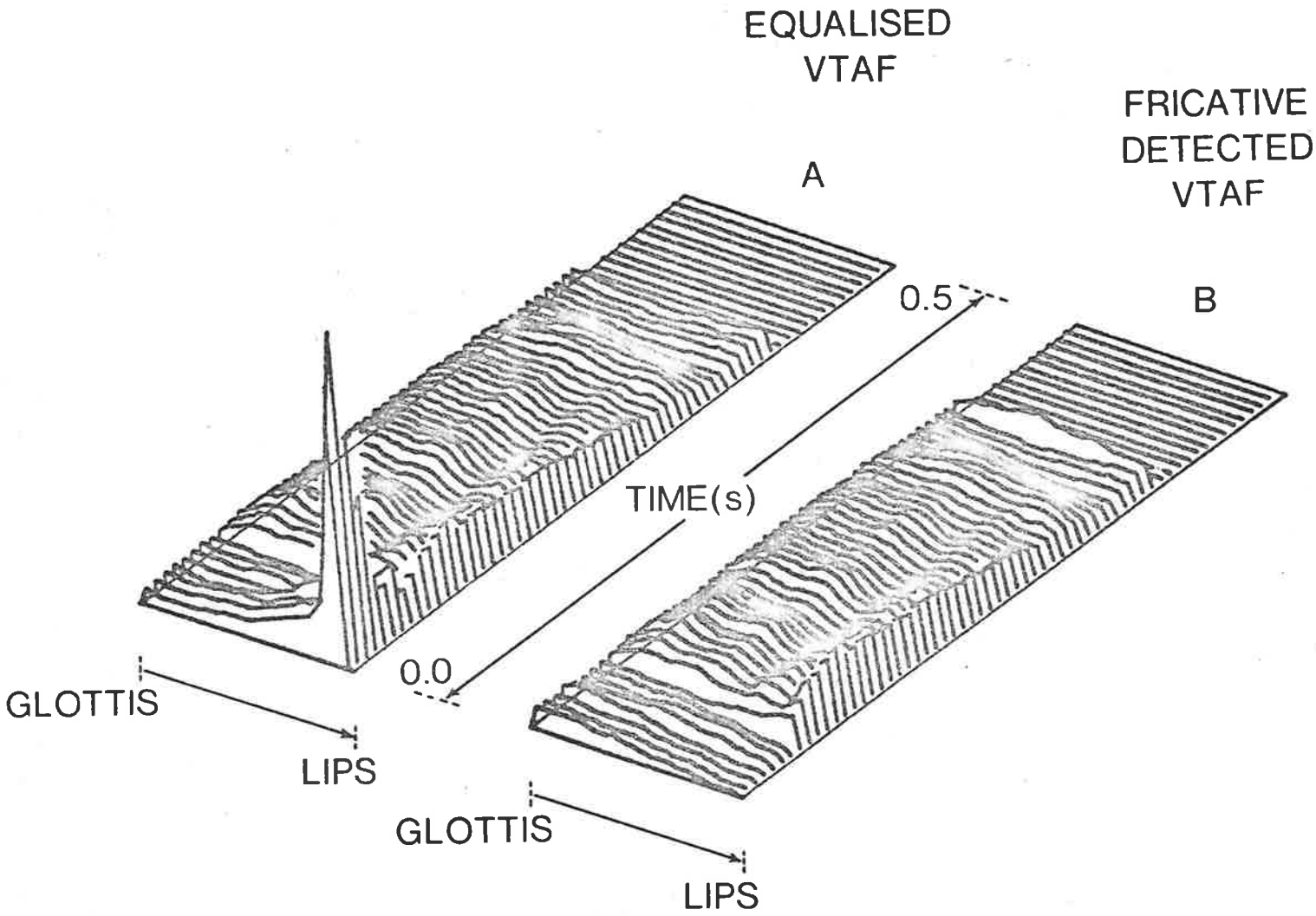
DATA ACQUISITION AND TIME SERIES EDITING

In this Appendix some aspects of the speech processing package developed for this study but not covered earlier are discussed. These are basically associated with the data acquisition and time series editing. Before discussing these a brief description of the hardware used is given.

Initially a software package, mainly is FORTRAN, was written for use on a mini-computer system based on a Data General Supernova. An extended replacement package was written in PL/I when it was decided to transfer the speech processing to a mainframe IBM 370/3033 system. In the first stage the only display available was a Tektronix 4010 which was later supplemented by a Tektronix 305 intensity modulated display. In the second stage a Tektronix 4015 with enhanced graphics mode became available. This was useful in the time series editing mode as described below. Waterfall and pseudo three-dimensional plots could be made on this. An example of the latter is shown in figure A.1. A Lexidata 3400 colour display system was later added to the mainframe and this was used to produce the areagrams and spectrograms shown earlier. The data acquisition system is now detailed.

A.1 DATA ACQUISITION

All speech used in this study was initially recorded on analog 8 mm tape. No special studio recording facilities were used; partly for convenience and partly to enable speech of a realistic quality to be obtained. A commercial portable 4 track tape recorder (Uher Reporter) and a single standard commercial microphone was always used. Recordings were made at the highest speed available (20cm/s).



SPOKEN DIGIT '2'

Figure A.1 PSUEDO — 3 D PLOT

Figure A.2 shows the method for digitizing the recorded analog speech signals. The signal was played back from the Uher and filtered using a cascaded sequence of low-pass filters set somewhat below the Nyquist frequency. Throughout this study the following parameters were used to obtain a digital speech representation.

TABLE A.1

Recording and playback speeds	: 20cm sec^{-1}
Digital sampling rate	: 1 6.384 kHz
Analog to digital convertor	: 12 bit ± 5 V
Low-pass filter cutoff	: 8.00 kHz
Filter roll-off	: 48 dB/octave
Digital Computer Compatible	: 9 tracks
Tape (CCT)	: 800/1600 BPI

As seen in figure A.2 two methods of digitizing were available. This is largely a relic of the two different systems used at different stages of this study. In the earlier stages data were transferred to computer compatible tape (CCT) at 800 BPI for the DG Supernova System or at 1600 BPI for the IBM 3033 mainframe. In its final configuration data was transferred directly to the IBM disk drives and the intermediate CCT stage avoided.

A.2 TIME SERIES EDITING

There are several features associated with the digitized raw speech signals which are undesirable. These are basically the presence of extraneous noise (e.g. microphone 'clicks', throat clearing, etc) and the varying length of silences between utterances. To remove these effects a visual display editing program was written to display the raw time series, page by

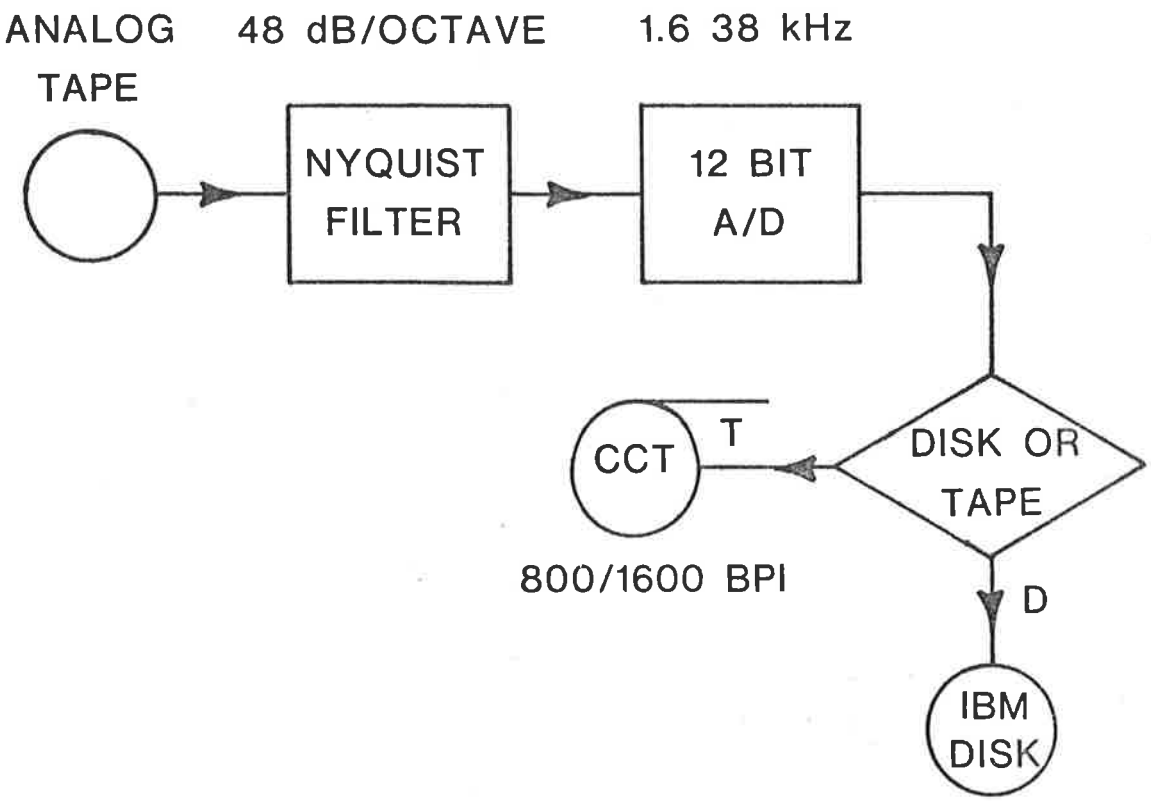


Figure A.2 DIGITISING SYSTEM — MODULE 1

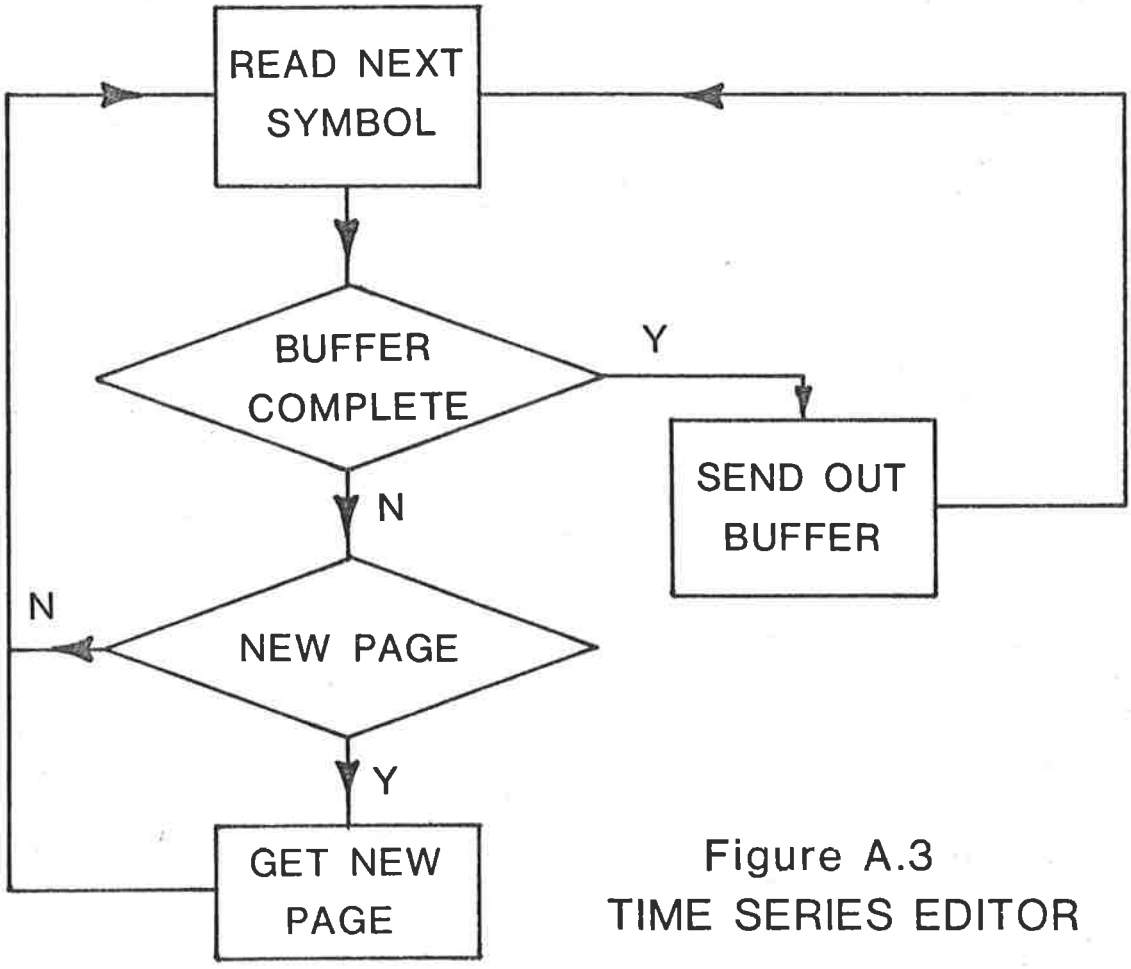


Figure A.3
TIME SERIES EDITOR

page, and to allow the operator to remove the noisy segments and standardise the silence duration between utterances. When editing is complete the operator can rewrite the edited speech to a new disk dataset which is then available to the next stage of the processing.

Figure A.3 shows the steps involved in the editing process. After selecting the required raw dataset the operator inputs the starting point (in record lengths of 1024 words) and the first page of time series data is displayed. A 'page' corresponded to 8 seconds of speech data and was displayed, as shown in the example of figure A.4, on a single line of seconds long. As only 4096 screen points can be addressed by the 4015 in the x direction only every 32nd of the 131072 data points per line could be displayed. The first stage of editing is setting the vertical gain for display. This is done recursively by accepting values from the keyboard and displaying the first page of data until the operator is satisifed. This does not take long as the maximum values of the signal are known a priori and it is simple to set a default value for y gain based on this.

Having set this value editing can begin. This is done by enabling the cross-hair cursor of the 4015 and using this to select the beginning and end-points of the required speech segments. At each stage of the process the operator can input various alphanumeric characters to achieve the following:

(1) Select start of required segment.

(2) Select end of segment and transfer segment to output buffer.

(3) Display new page.

(4) End editing and transfer output buffer to disk.

By typing two start keys in succession a selection can be aborted and to prevent confusion when the page was 'turned' a warning symbol is displayed

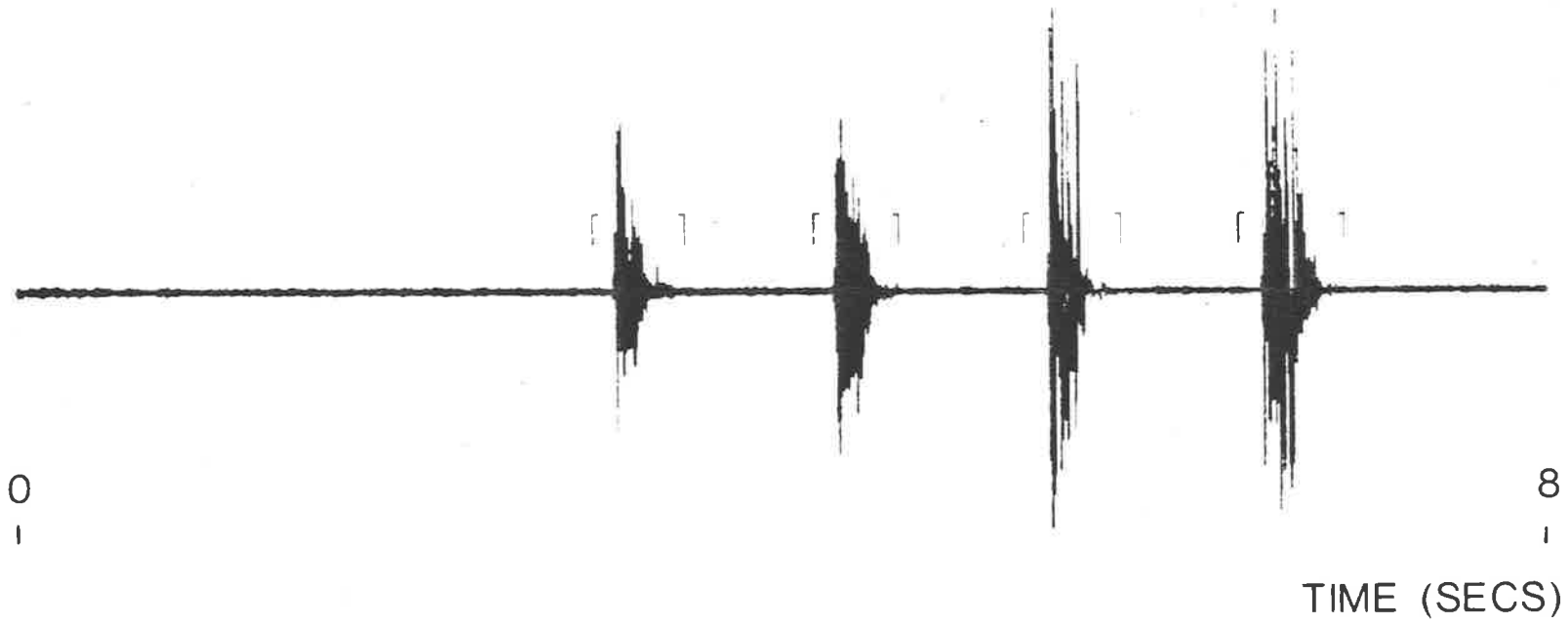


Figure A.4 TIME SERIES EDITOR

if a segment selection had already begun on the previous page. In general this program works very satisfactorily and large amounts of speech data can be rapidly edited.

REFERENCES

No.	Author	Title
1	R.K. Potter, G.A. Kopp and H.C. Green	"Visible Speech", Van Nostrand, New York, 1947
2	B.S. Atal and S.L. Hanauer,	"Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Amer., vol.50 (pt.2), page 637, 1971
3	H. Wakita	"Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms", IEEE Trans. on Audio and Electroacoustics, vol.AU-21, page 417, 1973
4	J.D. Markel, A.H. Gray, H. Wakita	"Linear Prediction of Speech-Theory and Practice". SCRL Monograph No.10, Speech Communications Research Laboratory, Santa Barbara, 1973
5	J. Makhoul	"Linear Prediction: A Tutorial Review", Proc. IEEE, vol.63, page 561, 1975

No.	Author	Title
6	H. Wakita	"Estimation of Vocal-Tract Shapes from Acoustical Analysis of the Speech Wave: The State of the Art", IEEE Trans. Acoust., Speech and Signal Processing, vol.ASSP-27, page 281, 1979
7	N. Levinson,	"The Wiener RMS Error Criterion in Filter Design and Prediction". J. Math. Phys. vol.25, page 261, 1947
8	A.V. Oppenheim and R.W. Schafer	"Digital Signal Processing", page 512, Prentice-Hall, Englewood Cliffs, 1975
9	J.D. Markel and A.H. Gray	"Linear Prediction of Speech". Section 4.3, Springer-Verlag, Berlin, 1976
10	R.W. Schafer and L.R. Rabiner	"Digital Representation of Speech Signals", Proc. IEEE vol.63, page 662, 1975
11	D.G. Nichol and R.E. Bogner	"Quasi-Periodic Instability in a Linear Prediction Analysis of Voiced Speech", IEEE Trans. Acoust. Speech and Signal Processing, vol.ASSP-26, page 210-216,1978
12	J.D. Markel and A.H. Gray	ibid, page 66

No.	Author	Title
13	R.E. Bogner and J.A.V. Rogers	"Determination of Vocal Tract Area Functions from a Pole Description of Speech Spectra", In Proc. Int. Conf. on Speech Commun. and Processing, page 368, 1972
14	G. Fant	"Acoustic Theory of Speech Production", Monton, The Hague, 1960
15	J.L. Flanagan	"Speech Analysis, Synthesis and Perception', Section 3.75, 2nd Edition Springer-Verlag, 1972
16	A.V. Oppenheim	"Speech Spectrograms using the Fast Fourier Transform", pages 57, IEEE Spectrum, 1970
17	P.M. Morse and H. Feshbach	"Methods of Theoretical Physics", page 1354, McGraw-Hill, New York, 1953
18	Markel and A.H. Grey	ibid Section 4.3

No.	Author	Title
19	A.H. Gray and J.D. Markel	"A Spectral Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis", IEEE Trans.ASSP-22, page 207, 1974
20	R.K. Potter, G.A. Kopp and H.C. Green	ibid
21	E.T. Whittaker	"Expansions of the Interpolation Theory", Proc. Roy. Soc. Edinburgh, vol.35, page 181, 1915
22	S.D. Stearns	"Digital Signal Analysis", page 62, Hayden, Rochelle Park, 1975
23	S.D.Stearns	ibid page 63
24	S.S. McCandless	"An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", IEEE Trans. Acoust. Speech and Signal Processing, vol.ASSP-22, page 135, 1974
25	J.D. O'Connor	"Phonetics", page 34, Pelican, Harmondsworth, 1973

No.	Author	Title
26	D.G. Nichol	"The Processing of Bathythermograph Data: A Picture Analysis Approach", Pattern Recognition, vol.8, page 209, 1976
27	A. Rosenfeld	"A Nonlinear Edge Detection Technique", Proc. IEEE, vol.58, page 814, 1970
28	C.W. Eriksen and H.W. Hake	"Absolute Judgements as a Function of the Stimulus Range and the Number of Stimulus and Response Categories", Journal of Experimental Psychology, vol.49, page 323, 1955
29	D.G. Nichol	"Some Automatic Techniques for Enhancing and Extracting Oceanic Features from Satellite Infrared Pictures", IEEE-MTS Oceans 78 Digest, page 433, 1978
30	J.D. O'Connor	ibid, chapter 7
31	D. Abercombie	"Elements of General Phonetics", Edinburgh University Press, Chicago, 1967
32	N.S. Trubetzkoy	"Principals of Phonology", University of California Press, 1969
33	R.E. Walpole and R.H. Myers	"Probability and Statistics for Engineers and Scientists", 2nd Edition, Section 3.3, Collier Macmillan, New York,

No.	Author	Title
1978		
34	R.E. Walpole and R.H. Myers	ibid, Section 6.6
35	J.A.V. Rogers	"Determination of Articulatory Parameters from Speech Waveforms", Ph.D. Thesis, Imperial College, University of London, London, 1974
36	S. Chandra and W.C. Lin	"Experimental Comparison Between Stationary and Non-Stationary Formulations of Linear Prediction Applied to Voiced Speech Analysis", IEEE Trans. Acoust. Speech and Signal Processing, vol.ASSP-22, page 403, 1974
37	R.M. Halsey and A. Chapanis	"Chromaticity-Confusion Contours in a Complex Viewing Situation", Journal of the Optical Society of America, vol.46, page 442, 1954
38	F. Ratliff	"Machbands: Quantative Studies on Neural Networks in the Retina", Holden-Day, San Francisco, 1965
39	T.N. Cornsweet	"Visual Perception", Academic Press, New York, 1970
40	K. Steiglitz	"On the Simultaneous Estimation of Poles and Zeros", IEEE Trans. Acoust. Speech, Signal Processing, vol.ASSP-25, page 299, 1977

No.	Author	Title
41	H. Morikawa and H. Fujisaki	"Adaptive Analysis of Speech Based on a Pole-Zero Representation", IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-30, page 77, 1982

1977 IEEE Int. Conf. on Acoust., Speech, and Signal Processing, May 9–11, 1977, pp. 212–215.
[7] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Processing*, to be published.
[8] J. S. Lim, "Evaluation of autocorrelation subtraction method for enhancing speech degraded by additive white noise," submitted to *IEEE Trans. Acoust., Speech, Signal Processing*.
[9] P. Eykhoff, *System Identification: Parameter and State Estimation*. New York: Wiley, 1974.
[10] F. C. Schweppe, *Uncertain Dynamic Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1973.
[11] H. L. Van Trees, *Detection, Estimation and Modulation Theory*. New York: Wiley, 1968.
[12] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 55, pp. 637–655, 1974.
[13] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
[14] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Heidelberg, New York: Springer-Verlag, 1976.
[15] F. Itakura and B. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," presented at the 6th Int. Cong. Acoust., Y. Kohasi, Ed., Tokyo, Japan, August 21–28, 1968, paper C-5-5.
[16] J. D. Markel and A. H. Gray, Jr., "A linear prediction vocoder simulation based upon the autocorrelation method," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 124–134, Apr. 1974.
[17] N. Levinson, "The wiener RMS (root mean square) error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261–278, 1947.
[18] J. Durbin, "The fitting of time-series models," *Rev. Inst. Int. Statist.*, vol. 28, pp. 233–243, 1960.
[19] A. Gelb *et al.*, *Applied Optimal Estimation*, A. Gelb, Ed. Cambridge, MA: M.I.T. Press, 1974.
[20] J. D. Gibson, J. L. Melsa, and S. K. Jones, "Digital speech analysis using sequential estimation techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 362–369, Aug. 1975.
[21] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1965.
[22] K. Steiglitz and L. E. McBride, "A technique for the identification of linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-10, pp. 461–464, 1965.
[23] J. K. Aström and P. Eykhoff, "System identification—A survey," *Automatica*, vol. 7, pp. 123–162, 1971.

Quasi-Periodic Instability in a Linear Prediction Analysis of Voiced Speech

D. G. NICHOL, MEMBER, IEEE, AND R. E. BOGNER, MEMBER, IEEE

Abstract—A significant semiperiodic fluctuation of the vocal tract area functions derived by linear prediction of the speech waveform has been noted during apparently stationary voiced segments of speech. In one example some values of the area function varied over a range of 9 : 1 over a few pitch periods. The phenomenon is attributed to "beating" of the pitch period and the time interval between successive computations which causes variations of the time relationship between glottal pulse and analysis window. This is supported by the fact that no fluctuations occur in the area function derived from natural or synthetic speech when the computation interval is equal to the pitch period. Any slight difference between the two leads to significant pulsations, however. A simple theoretical model is used to show how the positioning of the analysis window can influence area function estimates.

The problem can be largely overcome by using longer time windows (greater than 2.5 pitch periods), or alternatively, by averaging the area functions over several adjacent intervals.

Manuscript received August 1, 1977; revised December 16, 1977, and January 18, 1978. This work was supported in part by the Department of Defence and the Australian Research Grants Committee.

D. G. Nichol is with the Weapons Research Establishment, Adelaide, South Australia.

R. E. Bogner is with the Department of Electrical Engineering, University of Adelaide, Adelaide, South Australia.

I. INTRODUCTION

SEVERAL attempts have been made recently to use linear prediction analysis of speech for isolated-word [1], [2] and spoken-digit recognition [3], [4]. The feature chosen for the recognition algorithm in these studies was the set of linear prediction coefficients. It is well known that an estimate of the vocal tract area function can be derived from these coefficients [5]–[7] and the present paper arose from a study of the usefulness of this function for both speech and voice recognition. Because of the extensive information available from phonetic and articulatory studies of speech production, it was believed that the vocal tract area function (VTAF)

0096-3518/78/0600-0210$00.75

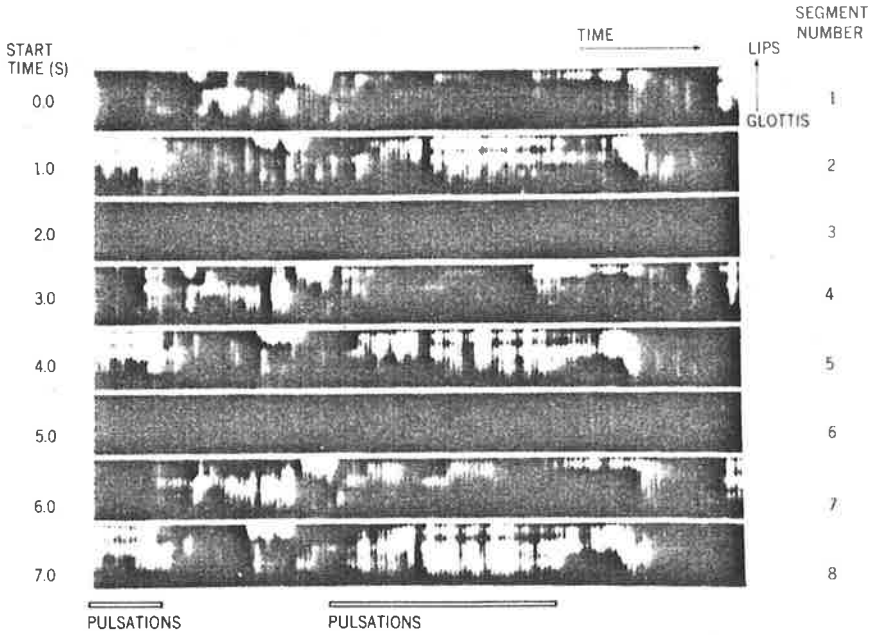


Fig. 1. Vocal tract area function for sentence "Speak to me now, bad kangaroo!"

would be an advantageous feature for the pattern recognition process. To test this idea, and also to compare the various formulations of the linear prediction models, it was decided to display the VTAF as an intensity-modulated picture of vocal tract position versus time, with the area plotted as a grey level. This is, of course, a similar display to the well-known spectrogram. We shall refer to these displays as VTAF pictures.

Fig. 1 shows a typical VTAF picture produced from real speech. The most obvious feature of this picture is the strong pulsations seen, for example, in the segments labeled 2, 5, and 8. It is believed that these pulsations are artifacts of the analysis as no evidence of them is apparent in the time series or spectrogram.

Before discussing this phenomenon in more detail, we describe briefly the production of the pictures.

II. Experimental Results

A. The VTAF Picture

The first linear prediction model used in this study was that due to Wakita [7]. This is a so-called autocorrelation technique and was chosen because, for non-pitch-synchronous analysis, these formulations are generally more stable and robust than the "covariance" methods, although for pitch-synchronous analysis the latter are capable of giving better estimates of the actual vocal tract [8], [9].

Suppose that the antialiasing filtered speech signal is sampled at frequency $f_s = 1/T$, and that n_W samples are included in each autocorrelation window and that a new computation of the VTAF is made every n_c samples. If m_1 linear prediction coefficients are used, then $m_V = m_1 + 1$ vocal tract areas are produced at time intervals of t_c where

$$t_c = n_c T. \tag{1}$$

Denoting the array of vocal tract areas $a_i(t)$ obtained at time t as a vector $\alpha(t)$ we have

$$\alpha(t) = a_1(t), a_2(t) \quad a_3(t) \cdots a_{m_V}(t). \tag{2}$$

If n successive estimates of $\underline{\alpha}(t)$ are evaluated, then the resulting sets of these $\underline{\alpha}(t)$ may be regarded as an $(m \times n)$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{1n} \\ a_{21} & a_{22} & a_{2n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{m1} & & a_{mn} \end{bmatrix} \tag{3}$$

where we have written a_{mn} for $a_m(nt_c)$.

This matrix can be plotted as an $(m \times n)$ digital picture where the grey levels are assigned by some mapping from the values of the elements of A to the set of grey levels.

To produce Fig. 1 the values $f_s = 8192$ Hz, $m_V = 9$, $n_W = 64$, $n_c = 64$, and $n = 1024$ were used. Now a 9×1024 picture is a very cumbersome shape, and so this was split into eight 9×128 subpictures, which for display purposes were interpolated (by a two-dimensional fast Fourier transform) into eight 36×512 subpictures. These eight subpictures were plotted, one below the other, as in Fig. 1 on an intensity-modulated CRT. The bottom of each subpicture represents the glottis and the top the lips. The grey levels have been assigned such that the larger the area the greater the whiteness. Thus the point of maximum constriction is the darkest region in each column. It should be noted that the Wakita model assumes a constant glottis area and thus the lower edge of each picture is a constant grey level. Some regions of the picture are blank. This is due to use of an energy-detecting algorithm which assigns arbitrary zero levels to the VTAF's when the total signal occurring in the time series window is below a given threshold (as during silences between utterances). Each subpicture represents 1 s of real time, and thus 8 s is shown overall.

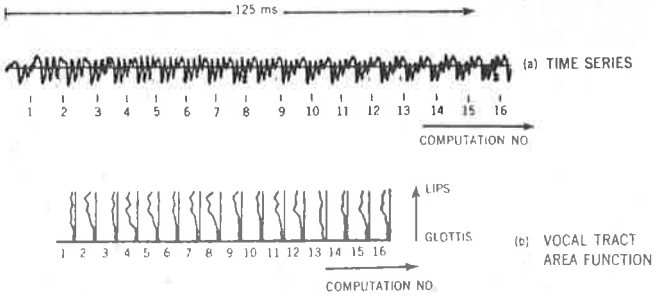


Fig. 2. (a) Speech waveform for /ae/. (b) Vocal tract area function for /ae/.

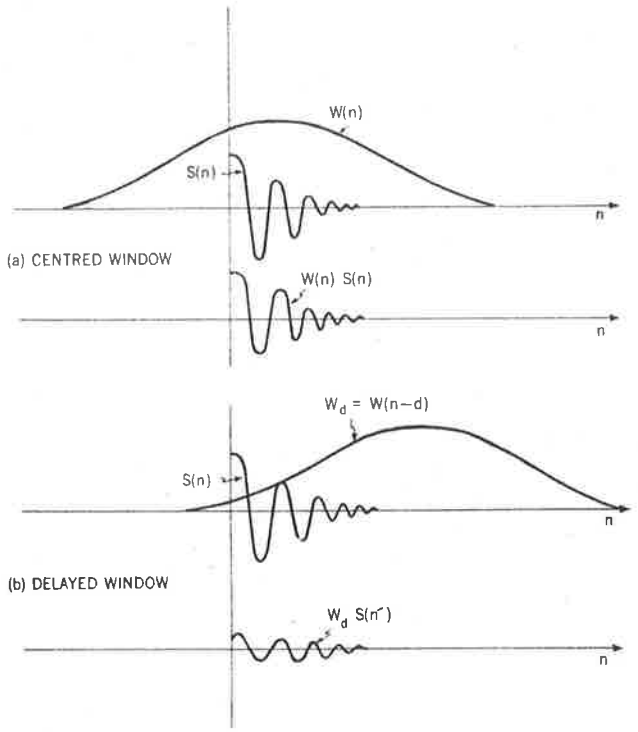


Fig. 3. Effect on windowed signal $s_w(n)$ of relative shift between signal and window. (a) Window centered on signal. (b) Window delayed with respect to signal.

B. Observations

In Fig. 1 the occurrence of periods of pulsations is easily observed; the most obvious of these are indicated in the picture. The utterance shown in this picture is the phrase "Speak to me now, bad kangaroo!" repeated three times by an Australian female speaker. The observed pulsations occur during constant vowel segments where little or no real change in the vocal tract is occurring.

This is supported by an examination of the speech time series corresponding to the utterance. Fig. 2(a) shows the speech waveform corresponding to the /æ/ in "bad" which for this speaker is remarkably stationary. The corresponding VTAF's are plotted in Fig. 2(b) and are clearly fluctuating, an effect which does not auger well for using the VTAF's in any automatic speech recognition process. The plots in Fig. 2(b) are in fact the square root of the VTAF, and thus the estimates of area actually vary by the order of 9:1 during this apparently stationary segment.

We had not observed this phenomenon previously though several VTAF pictures of Australian male speak peating the same phrase had been made. This suggeste the phenomenon may be sensitive to pitch period.

Now for linear prediction models of the Wakita typ analysis begins by windowing the time series by a H weighting function. The only parameter which is ch during a stationary segment is thus the position of the pulse within the Hanning window (unless the analysis i synchronous). We shall now examine this effect and s can cause the observed phenomenon.

III. Analysis of Window Position Effects

We examine the effect of the time relationship betw autocorrelation window and the speech waveform (I To facilitate analysis we use a model comprising a s order (two-junction, three-section) vocal tract yielding pulse response of the form

$$h(n) = r^n \cos n\omega T, \qquad n = 0, 1, 2 \cdots.$$

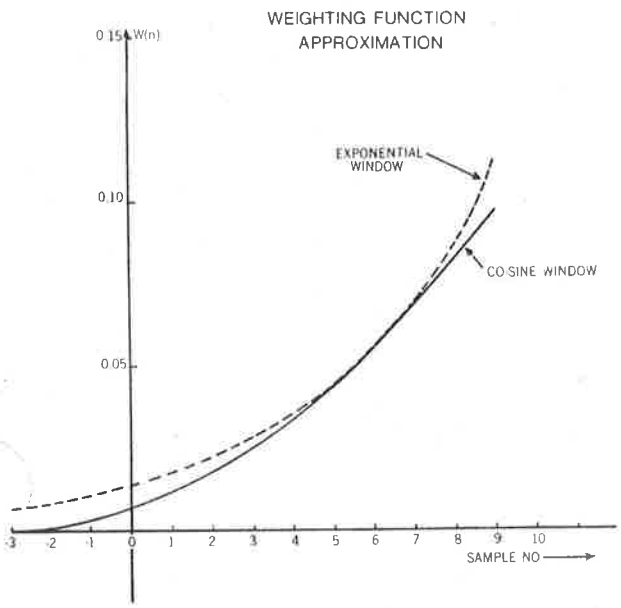


Fig. 4. Approximation to a delayed window $w_d(n) = 0.58\ [1 + \cos (2\pi/128)\ (n - 61)]$ by a rising exponential 0.014×1.26^n.

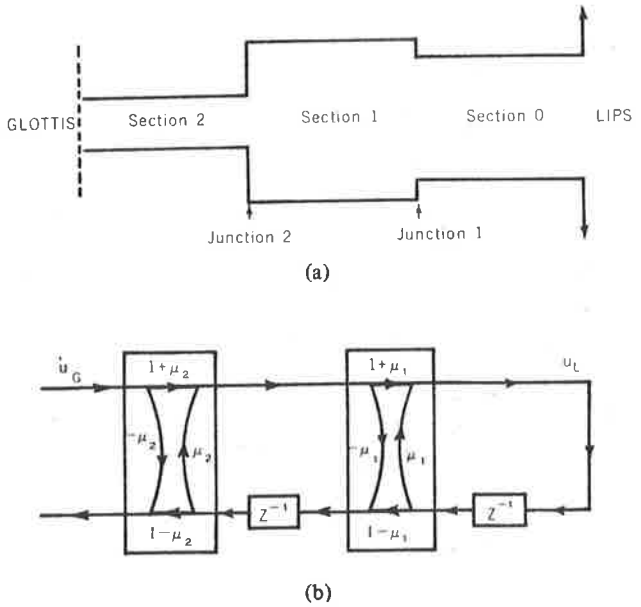


Fig. 5. (a) Physical tube model. (b) Signal flow model.

We use a window of the form

$$w(n) = \frac{1}{2} + \frac{1}{2}\cos\frac{2\pi(n-d)}{N}, \qquad n = -\frac{N}{2} \text{ to } +\frac{N}{2} \tag{5}$$

where d is the delay whose effect is of interest, and N the duration of the window is large compared with $1/(r-1)$, i.e., the interval over which the impulse response has significant magnitude. The excitation is taken to be a unit pulse, and thus the model speech signal $s(n)$ is the same as $h(n)$.

This model is not realistic, but it does aid our appreciation of effects which can arise, and yields a sufficient explanation of our observations.

Fig. 3 shows how variation of the delay affects the windowed signal $s_w(n)$. Two cases are shown, viz., 1) centered window, in which $w(n) \doteqdot 1$ over the effective duration of $s(n)$, i.e., we have $s_w(n) \doteqdot s(n)$, and 2) delayed window, in which the curved rise of $w(n)$ progressively magnifies the signal, producing a compensation of the damping of $s(n)$.

Fig. 4 shows, for example, the shape of the window

$$w_d(n) = 0.5\left[1 + \cos\frac{2\pi}{128}(n - 61)\right] \tag{6}$$

compared with the exponential Rr_1^n with $R = 0.014$ and $r_1 = 1.26$, which were chosen in an ad hoc manner simply for demonstration purposes. The exponential appears to be a reasonable approximation to $w_d(n)$.

We see that for $s(n)$ of the form

$$s(n) = r_s^n \cos n\omega T \tag{7}$$

with $r < 1$ the delayed window would cause $s_w(n)$ to be approximately

$$s_w(n) = s(n)\, w_d(n) \doteqdot R(r_s r_1)^n \cos n\omega T. \tag{8}$$

Now, for speech sampled at 10 000 s^{-1}, the value of r_s is likely to be in the range 0.985–0.9 (i.e., approximately 50–300 Hz formant bandwidth, respectively). Thus, the apparent value of the ratio corresponding to r_s in $s(n)$ as given by (7) becomes the value $r_s r_1$ in (8) and the latter may be grossly in error, and even exceed unity, as with the example values of $r_1 = 1.26$ and $r_s = 0.9$.

Next we study the effect of a discrepancy in the value of r_s on the area function of a model vocal tract.

Fig. 5 shows an acoustic tube (or transmission line model) of a vocal tract in which there are three sections of area a_m, $m = 0$ at the lips end, 1 for the middle section, and 2 at the glottis. There are thus two junctions whose volume velocity reflection coefficients μ_m, $m = 1$ and 2 are given by [10, p. 66].

$$\mu_m = \frac{a_{m-1} - a_m}{a_{m-1} + a_m}. \tag{9}$$

The termination at the lips is assumed to be equivalent to a tube section of infinite area, resulting in a volume velocity reflection coefficient of -1. For convenience in analysis, we associate all the delay (i.e., sum of delays for forward and backward traveling waves) with the backward traveling wave in each section. The physical model of Fig. 5(a) may then be represented by the signal flow model of Fig. 5(b).

Analysis of this model shows that the transfer function $H(z) = U_L(z)/U_G(z)$ is given by

$$H(z) = \frac{(1+\mu_1)(1+\mu_2)}{1 + z^{-1}\mu_1(1+\mu_2) - z^{-2}\mu_2} \tag{10}$$

$$= \frac{4\,a_0 a_1}{(a_0 + a_1)(a_1 + a_2)} \cdot \frac{1}{1 + z^{-1}\left(\frac{2a_1}{a_1 + a_2}\right)\left(\frac{a_0 - a_1}{a_0 + a_1}\right) + z^{-2}\left(\frac{a_2 - a_1}{a_2 + a_1}\right)}. \tag{11}$$

The impulse response of this system is of the form

$$h(n) = h(0)\, r^n \cos n\omega T \tag{12}$$

where

$$h(0) = (1 + \mu_1)(1 + \mu_2) = \frac{4a_0a_1}{(a_0 + a_1)(a_1 + a_2)} \tag{13}$$

$$r^2 = -\mu_2 = \frac{a_2 - a_1}{a_2 + a_1} \tag{14}$$

and

$$\cos \omega T = -\frac{1}{2r}\mu_1 (1 + \mu_2) = \frac{1}{2r}\left(\frac{2a_2}{a_1 + a_2}\right)\left(\frac{a_0 - a_1}{a_0 + a_1}\right). \tag{15}$$

From (14) we see that, for this two-junction model, the damping ratio r depends only on the reflection coefficient of the junction closest to the glottis, i.e., on the area ratio at this junction. We might query the physical meaning of the possibility that $a_2 - a_1 < 0$, i.e., $r^2 < 0$ in (14). Detailed analysis shows that the impulse response is then not oscillatory, corresponds to real poles, and is not of interest in the present study.

To apply this two-junction model to realistic speech parameters, we set the length of each section equal to half the length of the vocal tract, i.e., about 9 cm. To tie in with the previous discussion of the 10 000 Hz sampling rate, it is convenient to let each of the sections be equivalent to an each-way delay of 3×10^{-4} s. The delay T in the second-order model described by (8) is thus 6×10^{-4} s, and the relevant values of r for use in these equations are $(r_s r_1)^6$ or $(r_s)^6$. Of course, this change in fact replaces the second-order system by a 12th-order system if the original sampling rate is maintained, since denominator factors of $H(z)$ in (12) of the form

$$(1 - z^{-1}\, r\, e^{j\omega T})$$

are replaced by factors of the form

$$(1 - z^{-6}\, r^6\, e^{j6\omega T}).$$

Each of these factors results in six poles, but the base pole of each is the same as previously, i.e., at $z = re^{j\omega T}$.

From (14) we find

$$a_2 = \frac{1 + r^2}{1 - r^2}\, a_1 \tag{16}$$

and

$$\left.\frac{\frac{da_2}{a_2}}{dr}\right|_{a_1} = \frac{4r}{1 - r^4} \tag{17}$$

and

$$\left.\frac{\frac{da_1}{a_1}}{dr}\right|_{a_2} = \frac{-4r}{1 - r^4} \tag{18}$$

From (17) and also (18) we see that the proportional variation of either area a_1 or a_2 with r, while the other is fixed becomes very great as $r \to 1$. We found earlier that the effect of the delayed window on the apparent damping was sufficient to make r pass through unity, and thus the system may incur such great sensitivities. For example, varying r from 0.99 causes a_2/a_1 to change from 9.53 to 99.5. Clear effect is sufficient to account for variations as large a observed in Section II-B.

TABLE I
POLE POSITIONS IN TERMS OF FORMANT FREQUENCIES AND BAND (AFTER FANT [13])

VOWEL	FIRST FORMANT FREQ.	FIRST FORMANT B-WIDTH	SECOND FORMANT FREQ.	SECOND FORMANT B-WIDTH	THIRD FORMANT FREQ.	THIRD FORMANT B-WIDTH	FOURTH FREQ.
/æ/	616	57	1072	72	2430	130	3410
/e/	432	39	1959	95	2722	170	3500
/I/	222	60	2244	75	3140	240	3700
/ɒ/	510	54	900	65	2400	100	3220
/ʌ/	231	69	615	50	2375	110	3320

For completeness, we study the effect at the first ju We note that ωT is not affected by the window dela nomenon, and thus we set $d(\cos \omega T) = 0$ when di tiating (15).

We find

$$a_0 = a_1 \frac{r}{1 + r^2} \cos \omega T$$

and

$$\left.\frac{\frac{da_0}{a_0}}{dr}\right|_{a_1} = \frac{1 - r^2}{r(1 + r^2)}$$

which shows that the area ratios at the first junction a strongly influenced by r.

Note also that there is a gross effect on the initial valu of the windowed response. Via (13) we see that this can the product of the junction transmission coefficients $(1 + \mu_1)(1 + \mu_2)$. The effect on a particular feature, hov is not explicit.

For more complex vocal tract models, the effects are complex, but we have demonstrated a sufficient mech to account for the observations. One previous study showed that under moderate variation of the pole dampi a five-pole signal, the resultant VTAF retained its gros tures, but underwent a gradual smooth change.

IV. EXPERIMENTAL DIAGNOSIS

To test these ideas, synthetic vowels were generated (i computer) using an all-pole filter and known excitation tion. Details of the synthesis algorithm used are giv Rogers [12]. The four poles used were derived fron values of formant positions and bandwidths given by [13] (Table I).

Fig. 6 shows plots of the VTAF's obtained for the synt vowel /ae/ when the pitch period n_p has been made equ the computation interval n_c. In each of the four colu however, the "phase" of the excitation function relati the computation window is different (as indicated ir figure). Clearly, the areas calculated vary with this p This means that when $n_p \neq n_c$, the area calculated from a stant waveform will fluctuate as the position of the excit

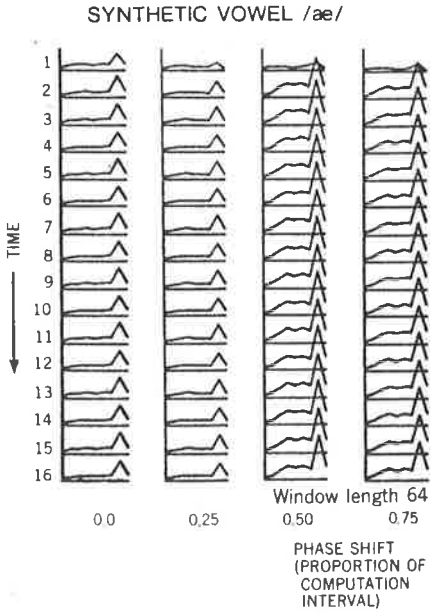


Fig. 6. Pitch period and computation interval are equal but with different relative positions in each column.

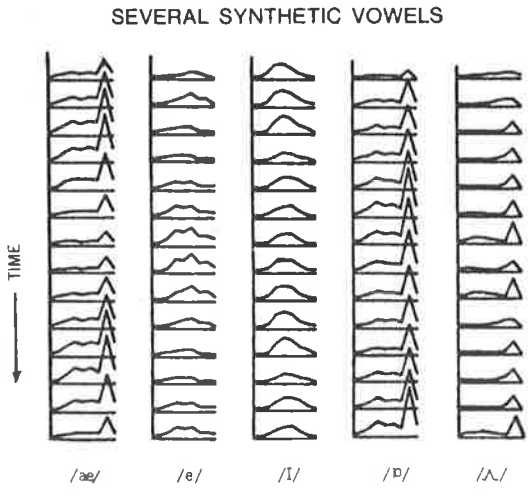


Fig. 7. Changing position of pitch pulse within window leads to pulsations when $n_c = n_w$.

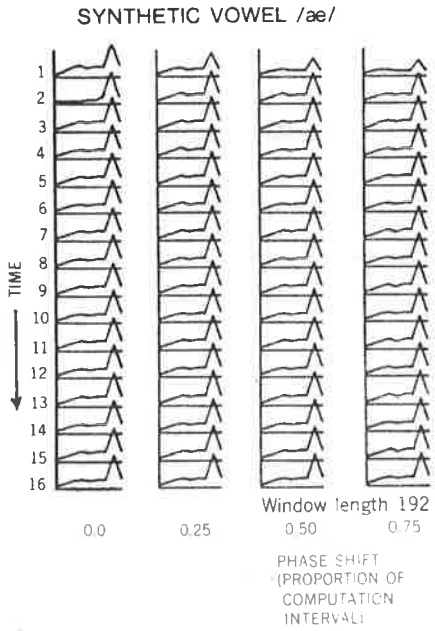


Fig. 8. Increasing n_w to 3.5 n_p suppresses the fluctuations for synthetic vowel /æ/.

impulse changes within the window. Fig. 7 shows this happening when $n_p = 0.8\ n_c$ for five different synthetic vowels.

The reason that this effect had not been observed in previous VTAF pictures of male speakers is believed to be that the computation interval used (64 samples, equivalent to 7.81 ms at 8192 Hz sampling rate) is quite close to the pitch period of the speakers analyzed. Thus, fluctuations are not observed as the excitation function remains in a nearly constant position in the Hanning window. It should be remembered, however, that the errors may still be present in the analysis but not show up as fluctuations. For the female speaker $n_p \approx 0.9\ n_c$ and the fluctuations are obvious (Fig. 1). This interpretation is supported by the fact that fluctuations did appear in male VTAF's that have been reprocessed with larger values of n_c.

V. Suppressing the Fluctuations

It appears from the above discussion that a partial cure for the problem of fluctuations would be to increase the size of the Hanning window used to estimate the autocorrelation function. This should improve the estimate of r_s. Fig. 8 shows the synthetic vowel /æ/ as shown in Fig. 6 but with $n_w = 3.5 n_p$. We see that the variations are suppressed. To test this on real speech the VTAF picture (Fig. 1) was reprocessed with $n_w = 192$ (approximately 3.5 n_p) and the result is shown in Fig. 9. The fluctuations have indeed been largely suppressed.

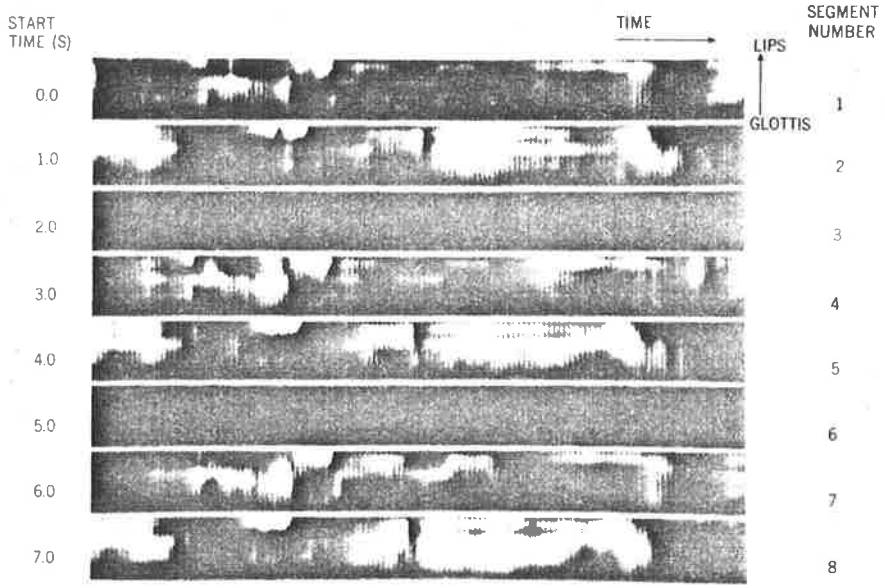


Fig. 9. Vocal tract area function for "Speak to me now, bad kangaroo!" for window $n_w = 3.5$ times pitch period n_p.

VI. Concluding Remarks

The autocorrelation methods of linear prediction have a certain attraction in terms of robustness and economy of computing effort. We have shown that care must be taken in choosing window lengths for the analysis, but provided this is done, then consistent estimates of the vocal tract area are obtained. If Hanning windows of length $>2.5\ n_p$ are used, the resultant area functions appear to have the robustness desirable for automatic speech recognition, or for use in visual displays for speech training and phonetic studies. We recognize that while the acoustic tube models so obtained may be consistent, they may not represent the shape of the vocal tract accurately.

References

[1] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.

[2] H. Fujisaki and Y. Sato, "Evaluation and comparison of features in speech recognition," Faculty Eng., Univ. Tokyo, Tokyo, Japan, *Annual Rep. Eng. Res. Inst.*, vol. 73, pp. 213–218, 1973.

[3] M. R. Sambur and L. R. Rabiner, "A speaker-independent digital recognition system," *Bell Syst. Tech. J.*, vol. 54, pp. 81–102, 1975.

[4] L. R. Rabiner and M. R. Sambur, "Some preliminary experiments in the recognition of connected digits," *IEEE Trans. Speech, Signal Processing*, vol. ASSP-24, pp. 170–182, A

[5] B. S. Atal and S. L. Hanauer, "Speech analysis and syn linear prediction of the speech wave," *J. Acoust. Soc. A* 50, pp. 637–655, 1971.

[6] J. D. Markel, "Formant trajectory estimation from a lin squares inverse filter formulation," Speech Commun. F Santa Barbara, CA, SCRL Monograph 7, 1971.

[7] H. Wakita, "Estimation of the vocal tract shape by op verse filtering and acoustic/articulatory conversion m Speech Commun. Res. Lab., Santa Barbara, CA, SCRL M no. 9, 1971.

[8] S. Chandra and W. C. Lin, "Experimental comparison stationary and nonstationary formulations of linear p applied to voiced speech analysis," *IEEE Trans. Acoust. Signal Processing*, vol. ASSP-22, pp. 403–415, Dec. 1974

[9] J. Makhoul, "Linear Prediction: A tutorial review," *Pro* vol. 63, pp. 561–580, 1975.

[10] J. D. Markel and A. H. Gray, *Linear Prediction of Spee* lin: Springer-Verlag, 1976.

[11] R. E. Bogner and J. A. V. Rogers, "Determination of vo area functions from a pole description of speech spe *Proc. Int. Conf. on Speech Commun. and Processin* pp. 368–371.

[12] J. A. V. Rogers, "Determination of articulatory pa from speech waveforms," Ph.D. thesis, Univ. London, England, 1974.

[13] G. Fant, *Acoustic Theory of Speech Production*. Th The Netherlands: Mouton, 1960.