DEVELOPMENT OF ACOUSTIC ANALYSIS TECHNIQUES FOR USE IN DIAGNOSIS OF VOCAL PATHOLOGY

by

Peter Murphy, BSc.,

Presented to the Office of Academic Affairs in partial fulfilment of the thesis requirements for the degree of Doctor of Philosophy in the School of Physical Sciences, Dublin City University, Dublin 9

September, 1997

Under the supervision of :

Professor Martin Henry,
School of Physical Sciences,
Dublin City University,
Dublin 9

Dr. Kevin McGuigan,
Department of Physics,
Royal College of Surgeons in Ireland,
123 St. Stephen's Green,
Dublin 2

(Single Volume)

Dedicated to my late father, and to my mother.

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy (Ph.D.) is entirely my own work and has not been taken from the work of others save and as to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Peter Mawby ID No.

Date: 21/08/97

Acknowledgements

I would like to thank Professor Martin Henry for accepting the responsibility of being my external supervisor. I would like to express my gratitude to Dr. Kevin M°Guigan, whose door was always open, for efficiently correcting the thesis and for encouragement offered throughout the course of the work.

I would like to express my gratitude to Professor Michael Walsh and Dr. Michael Colreavy in the ENT department and to Yvonne Fitzmaurice, Antonio Hussey and Jenny Robinson in the speech therapy department in Beaumont hospital. Thanks also to Mary Buckley and Patricia Gillivan-Murphy in the Eye and Ear Hospital.

In what, for me, has been a mini-odyssey about the college I am left with a lot of people to show my gratitude. Firstly, I can't thank enough, everybody up in pathology and microbiology (Maura, Bernie, Christina, Michael, Dorothy, Doreen, Peadar, Tina, Ann, Jemma, Brida, and others) who took me under their wing in my first year, when lumbered with me up in the pathology museum.

Thanks to Julie Duncan and Karen Pierce in physics, both of whom have been very helpful to me over the last few years and many thanks also, to Orla Cooney, particularly in respect to the collaborative time spent together on the present project before literally taking to the alcohol.

In chemistry I would like to express my gratitude to Eimear O' Brien, Fiona Bohan, Joe Jolley, Gillian McMahon, Marie Hosie, Terry Murphy and Karen Lenehan.

Thanks also to all the old school of Jocelline Levillan, Sylvie Le Roi, Elizabeth Flannagan, Sharon Brady, Andy Finnucane, Colm Campbell, Adolpho Aquilar, Suzanne Atkinson, Ali Akbar, Michael Foley, Michaela Walshe and to everyone in the department.

I would also like to thank Chris Murphy for saving me hours of turmoil with word processing. Thanks to A.T. Bates for many gifts and to Joanne Fenlon and Ferruccio Renzoni for providing distractions.

Finally, I wish to thank everyone in the college for making what has been an exceptionally friendly working environment.

PS Many thanks to everyone in media services for all your help over the past few years.

Abstract

Acoustic analysis as used in the vocal pathology literature has come to mean any spectrum or waveform measurement taken from the digitised speech signal. The purpose of the work as set out in the present thesis is to investigate the currently available acoustic measures, to test their validity and to introduce new measures. More specifically, pitch extraction techniques and perturbation measures have been tested, several harmonic to noise ratio techniques have been implemented and thoroughly investigated (three of which are new) and cepstral and other spectral measures have been examined. Also, ratios relevant to voice source characteristics and perceptual correlation have been considered in addition to the tradition harmonic to noise ratios. A study of these approaches has revealed that many measurement problems arise and that the separation of the indices into independent measures is not a simple issue. The most commonly used acoustic measures for diagnosis of vocal pathology are jitter, shimmer and the harmonic to noise ratio. However, several researchers have shown that these measures are not independent and therefore may give ambiguous information. For example, the addition of random noise causes increased jitter measurements and the introduction of jitter causes a reduced harmonic to noise ratio. Recent studies have shown that the glottal waveform and hence vibratory pattern of the vocal folds may be estimated in terms of spectral measurements. However, in order to provide spectral characterisation of the vibratory pattern in pathological voice types the effects of jitter and shimmer on the speech spectrum must firstly be removed. These issues are thoroughly addressed in this thesis. The foundation has been laid for future studies that will investigate the vibratory pattern of the vocal folds based on spectral evaluation of tape recorded data. All analysis techniques are tested by initially running them on specially designed synthesis data files and on a group of 13 patients with varying pathologies and a group of twelve normals. Finally, the possibility of using digital spectrograms for speaker identification purposes has been addressed.

TABLE OF CONTENTS

3 Investigation into Speaker Identification using Digital Speech Spectrograms

4 Time Domain Analysis

5 Harmonic Intensity Analysis

Appendix A

Source Code for Principal Matlab Program Files

Chapter 1

Background to Acoustic Analysis of Voice

1.1 Introduction

Present day basic research on voice is a multidisciplinary endeavour involving specialists from such diverse fields as physiology, anatomy, neurology, physics, electrical and electronic engineering, computer science, speech sciences, speech therapy, otolaryngology and phonetics. Even within the field of physics alone the subject encompasses wide ranging specialities including fluid dynamics, acoustic theory, network theory, viscoelasticity, vibration/damping studies, acoustic (spectral) analysis, system analysis, imaging (digital, x-ray, stroboscopy), laryngeal biomechanics, continuum mechanics and chaos theory. The possible benefits of such basic research are manifold including, for example, improved natural sounding synthesis, enhanced speech and speaker recognition strategies, more efficient coding for communication purposes and clinical diagnosis.

The serious and contentious issue of speaker identification is addressed in chapter three but here as in all other chapters we turn our attention to the equally serious issue of

diagnosis of vocal pathology. From all of the above mentioned areas of physics we limit ourselves to a discussion on the potential usefulness of acoustic analysis for vocal quality assessment. By acoustic analysis we simply mean any computer technique that is used to analyse the digitised voice signal, whether it be an accelerometer transduced signal, an inverse filtered glottal waveform or simply a standard microphone transduction of the output radiated speech waveform. The term 'acoustic analysis' should not to be confused with the related area, termed 'acoustic theory', which has been applied to give a scientific basis to the process of speech production.

Many problems arise when attempting to characterise vocal qualities based on perceptual measures and this situation is further exacerbated in the clinical setting. Labelling pathological voice types as hoarse is a wastebasket term substituted for any one or combination of the following:

> "...aspirate, breathy, coarse, dead, dull, feeble, flat, gloomy, grating, grave, growling, guttural, harsh, hoarse, hollow, husky, infantile, lifeless, loud, metallic, monotonous, muffled, neurasthenic, passive, pectoral, pinched, rasping, raucous, rough, sepulchral, shrill, sober, strained, somber, subdued, thick, thin, throaty, tired, toneless, tremulous, weak, whining and whispered."[1]

Part of the problem lies with the speech signal itself, due to it's complexity, carrying several sub-messages, indicative of emotional state, dialect etc. of the speaker, encoded into the main message of what is primarily a communicative gesture. Some of these sub-messages carry information that is indicative of the health of the vocal cords. Other problems are due to inter-rater variability and even intra-rater variability that arises when diagnosing voice type based on perceptual measures.

Acoustic analysis provides an appealing alternative, providing objective, quantifiable measurements. However, the acoustic analyses are only as good as their correlation with their perceptual counterparts. An alternative approach can be taken however in which the acoustic measures are correlated with vibratory events as viewed for example through laryngovideostroboscopy or electroglottogram recordings. Acoustic measurements taken on the output radiated speech waveform and it's spectrum have been shown to correlate with perceptual measures of 'roughness' and 'hoarseness'[2]. In

this chapter we define some commonly used vocal qualities, describe typical clinical procedures for voice assessment and provide the basic acoustic theory for both the glottal source and subsequent resonance in the vocal and nasal cavities that motivates the possibility of applying acoustic analysis studies to clinical assessment. Finally, acoustic analysis techniques currently available for use in clinical practice offer very limited information regarding differential diagnoses or perceptual correlations[3]. These limitations along with other methodological problems encountered in applying acoustic analysis to vocal pathology are investigated and procedures for improving the diagnostic value of acoustic analyses are examined.

1.2 Vocal Quality

In describing voice qualities or more specifically pathological voice types, it will be helpful to familiarise ourselves with some of the terminology. There is some need for standardisation here[4] as different terms can take on different meanings depending on the researcher's background and also the definition may be given in terms of perceptual, acoustic or physiological aspects. It is interesting to consider the perceptual labelling of voice qualities: these descriptions can only be compared to other sounds eg. vocal fry - "similarity with popping sounds that are emitted from a hot frying pan"[5] and perhaps this is the reason that bipolar labelling is used as in for example hypofunction and hyperfunction thus indicating that one sound is opposite to another. Alternatively, the sound can be described in terms of acoustic measures, physiological function or aerodynamic measurements. A description of breathy vocal quality using these measures might be described as having a relatively high fundamental frequency, less adducted vocal folds during the closed phase and high airflow (>500 $mlsec^{-1}$). Generally all measures are used interchangeably in the description on voice qualities or phonation types.

Phonation type can be considered a broad term describing any state of the glottis that provides energy to the vocal tract and 'voice' can be defined as the regular vibration of the vocal cords at any frequency within the speaker's normal range. The term 'voice'

in everyday conversation is often used to mean 'speech' as illustrated by Deller et al[6] "One often hears a singer described as having a "beautiful voice". This may indeed be the case, but the audience does not attend the concert to hear the singer's voice! " We ask therefore, what is the correct term to use to describe the singing ? The term voice quality has been used liberally in this section without definition. We can consider this term as appropriate in describing the sound produced by the singer and it therefore describes the supra-glottal as well as glottal activity. The possibility for confusion is clear and when describing voice quality with respect to glottal activity we will state so explicitly. Furthermore, the voice quality that we are interested in is voice quality 'speech' as opposed to voice qualities associated with singing, opera, belting etc.

Term	(Loose) definition
Phonation Type	Any state of the glottis that provides acoustic energy to the vocal tract
Voice	Regular vibrations of the vocal cords at any frequency within the speaker's normal range
Modal voice	Unmarked phonation type
Breathy voice	
Murmur	Vibrating, but more abducted vocal folds
Slack voice	
Vocal fry	Very low pitch vibrations involving only parts of the vocal folds
Creaky voice	
laryngealised voice	Vibrating, but more adducted vocal cords
Stiff Voice	
Pressed voice/ glottalised voice	May refer to more adducted vocal cords but may have other connotations

Table 1.1 *Some terms for phonation types (summarised from a presentation given by Prof. Peter Ladefoged[7] at the 5th Vocal Fold Physiology Conference).*

Table 1.1 gives a list of some commonly found phonation types[7]. The first four terms have been alluded to above and describe modal and breathy voice. The terms following and including 'vocal fry' are used to describe a mode of vibration that occurs when the vocal folds are more adducted than for modal voice. 'Vocal fry' describes a very low frequency form of this vibration in which amplitude or frequency modulation of every

second period results in the perception of a fundamental frequency an octave lower. These voice quality terms will be used extensively throughout the main text, along with their corresponding modes of vibration and acoustic correlates. They serve only as a very basic guide to voice quality assessment and more elaborate classification schemes exist, most notably the phonetically based Laver's Vocal Profile Analysis[8]. Another scale (GRBAS) related specifically to pathological voice types was introduced by the Japanese Society of Logopedics and Phoniatrics and voices were rated according to the degree (five point scale) or grade of, roughness, breathiness, asthenicity and strained quality[9]. Yet another scale based on years of clinical experience in Swedish speech therapy clinics is given in table 1.2. This bipolar scale shown with it's acoustical correlates as shown in table 1.2 is based on the work of Hammarberg and Gauffin[10].

1.3 Clinical Examination of Voice

When a patient presents with abnormal voice the clinician's primary concern is whether or not abnormal voice signifies illness. The cause or causes of abnormal voice must therefore be established through thorough examination[11,12,13].

Initially, this takes the form of a standard ear, nose and throat examination. Further examination, involving a full laryngologic evaluation is carried out as required. Indirect laryngoscopy is the traditional method for viewing the vocal folds. The patient is usually siting in an upright position and his/her tongue is wrapped in gauze to protect the frenum from the lower incisors. The tongue is then pulled outward from the mouth and the slightly warmed laryngeal mirror is introduced into the mouth and guided posteriorly by pushing the uvula upward and backward and positioned in the oropharynx. The effect of mirror reversal and the illumination source directed towards the laryngeal mirror on reflection from the familiar head mirror are shown in fig.1.1. The complete glottal and supra glottal areas are carefully examined during quiet breathing and sustained phonation. In recent years most voice clinics have introduced the videostroboscope which provides an excellent view of all glottal and supra glottal areas as well as the vibrating vocal folds. Nasopharyngealfiberscopy is also used.

Voice Quality Parameter	Tentative Definition
Aphonic/intermittent aphonic	Voice is constantly or intermittently lacking phonation-there are moments of whisper or loss of voice
Breathy	Audible noise created at the glottis, probably because of insufficient glottal closure
Hyperfunctional/Tense	Voice sounds strained, as if the vocal folds are compressed during phonation
Hypofunctional/Lax	Opposite to hyperfunctional, insufficient vocal fold tension, resulting in a weak and "slack" voice
Vocal Fry/Creaky	Low-frequency aperiodic/periodic vibration: vocal folds are very close together and only a section of them is free to vibrate
Rough	Low-frequency aperiodic noise, presumably related to some kind of irregular vocal fold vibrations
Gratings/"High-frequency roughness"	High-frequency aperiodic noise, presumably related to some kind of irregular vocal fold vibration
Unstable voice quality	Voice is fluctuating in pitch or invoice quality over time
Voice Breaks	Intermittent frequency breaks
Diplophonic	Two different pitches can be simulataneously perceived
Modal/Falsetto Register	Modes of phonation
Pitch	The chief auditory correlate of fundamental frequency
Loudness	The chief auditory correlate of sound pressure level of speech

Table 1.2 *Proposed perceptual scale for clinical assessment (After Hammerberg and Gauffin[10]).*

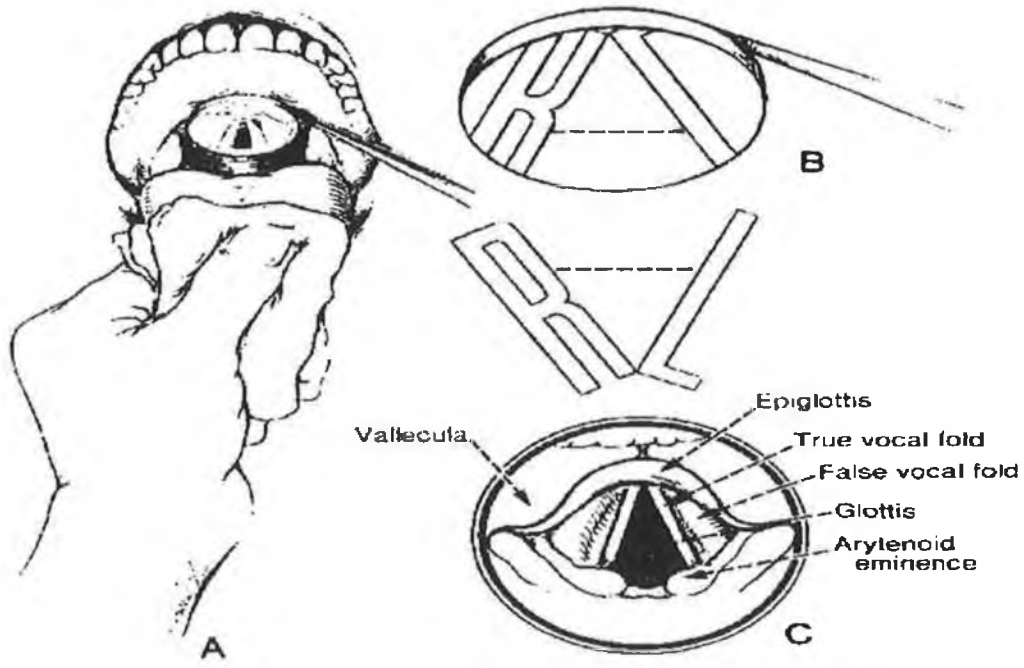


fig.1.1 *Laryngological examination illustrating the effect of mirror reversal*

This is where a fibre-optic is threaded through the nasal passages to provide the laryngeal image. Simple tests can also be performed by manual compression of the larynx to investigate the possibility of carrying out laryngeal framework surgery. Radiography and x-ray tomography techniques are also used to reveal the position, shape and size of laryngeal lesions. Additional techniques involve the use of high speed digital imaging and video fluoroscopy but these techniques are primarily used for research purposes involving laryngeal movements rather than structure.

1.4 Voice Source

In order to make meaningful inferences regarding the primary source of energy i.e. the vibration of the vocal folds based on the acoustic analysis of the output radiated speech waveform we need to have a good knowledge of source characteristics. During

phonation the respiratory muscles contract resulting in an excess pressure in the lungs which in turn causes airflow that is periodically interrupted due to the opening and closing of the vocal folds once every fundamental period. Sound is produced as a result of the interruption of the egressive airflow by the vocal folds and they do not generate any appreciable sound level due to their own mechanical vibration.

According to the myoelastic-aerodymanic theory (Van den Berg)[14] there are two primary forces acting on the vocal folds, the tension of the vocal folds themselves and the aerodynamic force exerted on them due to the exhaled air stream. The physics of the myoelastic-aerodynamic theory as given by Liberman[15] is summarised below according to Aronson[11].

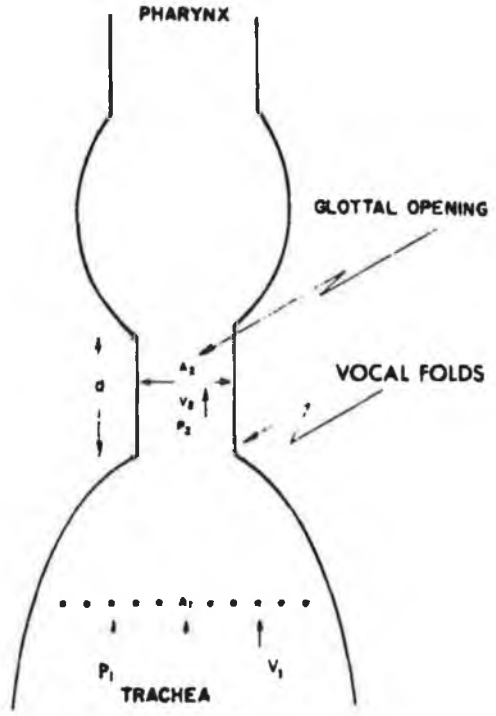


fig.1.2 Schematic diagram of forces acting on the vocal folds.
d : = Length of glottal constriction.
A_2 = Cross-sectional area of glottal constriction.
V_2 and P_2 = Particle velocity and air pressure at the glottal constriction.
A_1 = Cross-sectional area of the trachea.
V_1 and P_1 = Particle velocity and air pressure in the trachea. (*From* Lieberman, P.: Vocal cord motion in man. N.Y. Acad. Sci., 155:28–38, 1968.)

In consideration of the case when the folds are adducted and held passively in the midline fig.1.2 shows that :

1. Positive subglottic air pressure is represented by F_{AS}. When the glottis is closed this force displaces the true vocal folds outward from their adducted position.
2. The Bernoulli force, represented by F_{AB} is the negative pressure in the region of the glottis created by the high velocity airflow there.

3. Tension of the vocal ligaments that restore the vocal folds to their neutral position is represented by F_{TO} and F_{TC}.

   Interaction among the forces is as follows.

4. The aerostatic force F_{AS} resulting from the subglottic air pressure against the adducted vocal folds is maximum at the beginning of the cycle.
5. The Bernoulli effect, which is responsible for force F_{AB}, is an example of the conservation of energy; as the velocity of a gas or liquid increases as it flows from a point of lesser constriction to one of greater constriction, it's pressure decreases. Assuming that the glottal constriction contains a uniform frictionless flow of an incompressible fluid (fig.1.3):

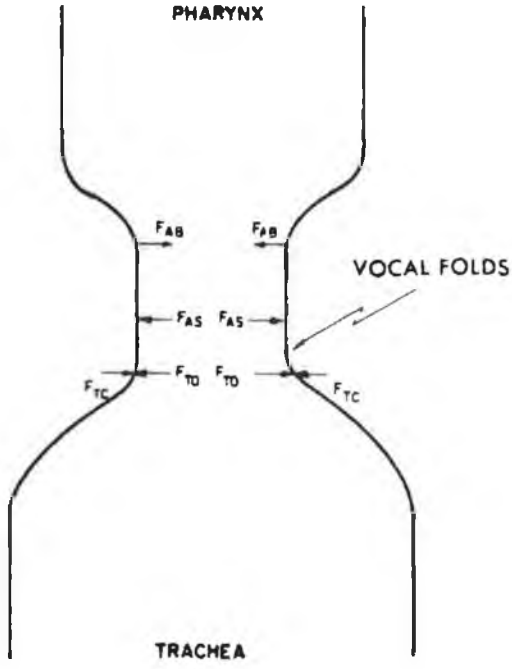


fig.1.3 Schematic diagram of forces acting on the vocal folds, in open position.
F_{AS}, Force exerted by subglottal air pressure, displacing vocal folds outward.
F_{TO} and F_{TC}, Forces acting to restore vocal folds to neutral position, owing to action of vocal ligaments.
F_{AB}, Bernoulli force generated by airflow through glottal constriction, acting to pull vocal folds inward.*From* Lieberman, P.: Vocal cord motion in man. Ann. N.Y. Acad. Sci., 155:28–38, 1968.)

a) the rate of fluid flow across A_1 is equal to $A_1V_1\rho$, where ρ is the density of the fluid, A_1 is the cross-sectional area of the trachea, and V_1 is the velocity of the fluid.

b) If the stream is steady, the same mass must travel per unit of time through the constricted portion of the pathway, so that

$$A_1V_1\rho = A_2V_2\rho \qquad \text{eqtn.1.1}$$

where A_2V_2 is the cross-sectional area times the particle velocity at the glottal constriction. Since the density p is constant, $A_1V_1=A_2V_2$. The particle velocity in the glottal constriction will thus be larger than the particle velocity in the pharynx V_1 because

$$V_2 = A_1V_1/A_2 \qquad \text{eqtn.1.2}$$

where A_2 is the cross-sectional area of the constriction. The kinetic energy of the fluid in the constriction

$$\text{K.E.} = 1/2\rho(A_1V_1/A_2)^2 \qquad \text{eqtn.1.3}$$

will, therefore, be higher in the constricted portion of the air passage. The potential energy must decrease as the kinetic energy increases, since the sum of kinetic and potential energies must remain constant. Physically, this means that the pressure of the fluid in the constriction, P_2, decreases.

c) The pressure in the constriction falls below atmospheric pressure as the cross section of the constriction decreases as the vocal folds begin to come together again and are sucked together by the pressure differential between P_2 and atmospheric.

In the above description we have considered the case of a hard glottal attack where both Bernoulli and elastic forces combine to restore the perturbed folds back to the midline. There are of course many variations to this production mechanism depending on type of glottal attack, voice register and use of intrinsic and extrinsic laryngeal muscles. A few examples are considered.

In the case of a soft glottal attack, the folds are initially in the abducted position and the Bernoulli effect (i.e. sucking force) alone, explains why the folds can depart from

an initial open state without muscle action. During voiced production there exists a phase difference at closure owing to the fact that the anterior edges of the folds are the first to close. This phase difference is reduced as the pitch increases due to the greater stiffness and reduced mass of the folds. In falsetto register it is primarily the upper edges that participate in phonation. Incomplete glottal closure may occur at soft onset and decay of voicing due to incomplete inward movement of the vocal folds, or it may be due to leakage as a result of a posterior glottal chink, as occurs in breathy voices. Recent work by Hanson[16] has considered both of these cases in some detail.

Based on a simplified mechanical analysis considering only the Bernoulli effect it follows from eqtn.1.2 that the time it takes for one oscillation of the vocal folds, is inversely proportional to the square root of the subglottal pressure and proportional to the square root of the vibrating mass and to the small distance the folds have to move away before the mean pressure in the glottis switches to a negative value. An increase in subglottal pressure will therefore cause an increase in the fundamental frequency if the normal compensation of a decreased tension of the folds is not included.

Model experiments of van den Berg et al[17](1957) shows that the glottis flow resistance R_F as a function of glottis area A, and particle velocity $v = u/A$, can be decomposed into two terms $R_F = R_L + R_T$, R_L being proportional to A^{-3} and independent of the flow and R_T (due to turbulent losses) being proportional to A^{-1} and v. The former is the resistance of a very narrow slit assuming laminar streaming.

$$R_L = \frac{12\,\mu\,l\,b^2}{A^3} \qquad \text{eqtn.1.4}$$

where $\mu = 1.84\times10^{-4}$ is the coefficient of viscosity. The glottis cross section is assumed to be rectangular and of the width $a = A/b$ across the slit and of the length $b = A/a$ in the direction of the slit. The depth of the slit is l.

When the glottis area has reached about 1/6 of it's maximum value, the second term R_T obtains equal magnitude and dominates at higher area values. This resistance is due to turbulent losses and was found to 7/8 of the resistance R_B associated with the kinetic pressure of the Bernoulli equation

$$p = \rho v^2/2 \qquad \text{eqtn.1.5}$$

where p is the pressure fall at the constriction. The resistance is

$$R_B = p/u = \rho v/2A = \rho u/2A^2 \qquad \text{eqtn.1.6}$$

There is also a resistive term of turbulent origin. Stevens et al[18] have investigated the nature of turbulent noise at the glottis which shows a somewhat high pass response up until about 1kHz and thereafter shows a flat spectral characteristic (fig.1.4).

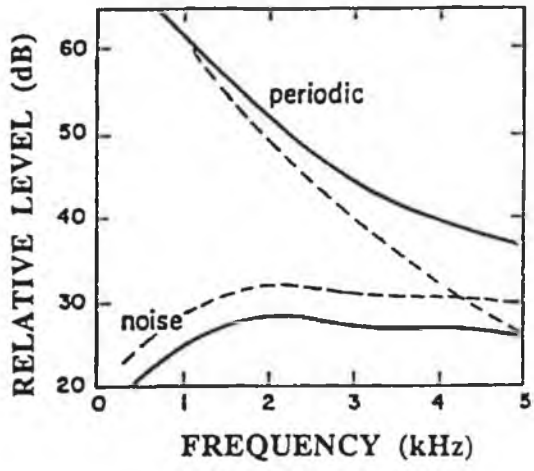


fig.1.4 *Spectra of volume velocity and turbulent noise source for two different glottal configurations (The minimum glottal opening has increased -dashed line).*

More basic experimentation is required in order to find out more about noise generation when the folds contain for example mass lesions. Turbulent flow arise from two possibilities, both of which are satisfied by the presence of mass lesions at the glottis. Turbulence arises due to a constriction of the flow causing the air particles to accelerate, forming a jet of air shot at high speed through the passage. The jet is associated with circulation effects and eddies, partially of a random nature. Alternatively, a particle hit by a jet of air gives rise to a turbulent source that can be of greater intensity than the noise produced in the passage. The Reynolds number is of basic interest in determining the onset of turbulence.

$$Re = vh/\upsilon \qquad \text{eqtn. 1.7}$$

h = width of passage

v = particle velocity

υ = kinematic coefficient of viscosity

1.5 Vocal Tract

In order to provide a completely detailed acoustic theory of sound propagation in the vocal tract all of the following must be considered:

1. Time variation of the vocal tract shape.
2. Losses due to heat conduction and viscous friction at the vocal tract walls.
3. Softness of the vocal tract walls.
4. Radiation of sound at the lips.
5. Nasal coupling.
6. Excitation of sound in the vocal tract.

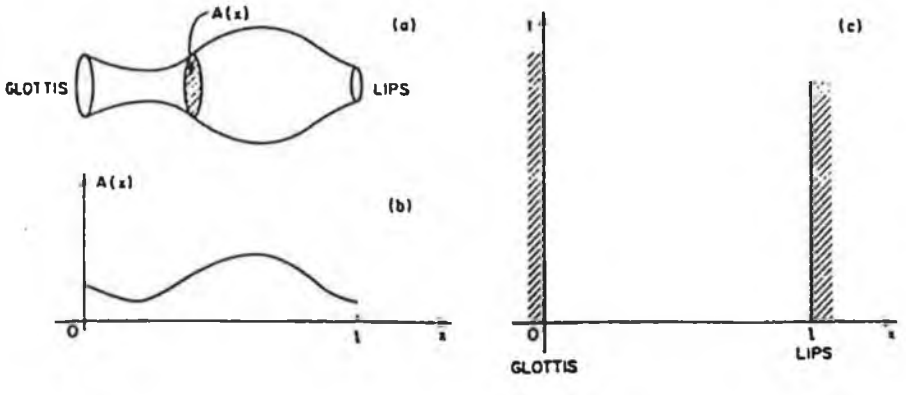


fig. 1.5 (a) *Schematic diagram of the vocal tract,* (b) *corresponding area function and* (c) *x-t plane for solution of wave equation.*

However, many simplifications are required in order to provide a useful numerical model of speech production. The schematic diagram in fig.1.5 shows the simplest physical configuration of practical interest. The vocal tract is modelled as a tube of non-uniform, time varying cross-section. Plane wave propagation is assumed for all frequencies below 4 kHz i.e. wavelengths that are long compared to the dimensions of the vocal tract. Furthermore, no energy losses due to either thermal conduction or viscosity are assumed to occur. With these assumptions, applying the laws of conservation of mass, momentum and energy to sound waves in the tube of fig.1.5, Portnoff[19] has shown that the following pair of partial differential equations are satisfied:

$$\frac{-\partial p}{\partial x} = \rho \frac{\partial \left(u / A \right)}{\partial t} \qquad \text{eqtn.1.7}$$

$$\frac{-\partial u}{\partial x} = \frac{1}{\rho c^2} \frac{\partial (p A)}{\partial t} + \frac{\partial A}{\partial t} \qquad \text{eqtn.1.8}$$

where

$p = p(x,t)$ is the variation of sound pressure in the tube at position x and time t.

$u = u(x,t)$ is the variation in volume velocity flow at position x and time t.

ρ is the density of air in the tube

c is the velocity of sound

$A = A(x,t)$ is the "area function" of the tube; i.e. the value of cross-sectional area normal to the axis of the tube as a function of a distance along the tube and as a function of time.

Using a variety of simplifications and approximations some straight forward solutions are possible. Considering a constant area function for the vocal tract which is

approximately correct for the neutral vowel /UH/ reduces the partial differential equation to the following form

$$\frac{-\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t} \; ; \; \frac{-\partial u}{\partial x} = \frac{\rho}{A}\frac{\partial p}{\partial t} \qquad \text{eqtn.1.9}$$

which have the familiar travelling wave solutions

$$u(x,t) = [u^{+}(t - x/c) - u^{-}(t + x/c)]$$

$$p(x,t) = \frac{\rho c}{A}[u^{+}(t - x/c) - u^{-}(t + x/c)] \qquad \text{eqtn.1.10}$$

The frequency domain representation of this model is obtained by assuming a boundary condition at x = 0 of

$$u(0,t) = u_G(\Omega) = u_G e^{j\Omega t} \qquad \text{eqtn.1.11}$$

that is, the tube is excited by a complex exponential variation of volume velocity of radian frequency ω and complex amplitude, $U_G(\omega)$. Since equation 1.9 is linear , the solution $u^{+}(t-x/c)$ and $u^{-}(t+x/c)$ must be of the form

$$u^{+}(t - x/c) = K^{+} e^{j\Omega(t - x/c)}$$

$$u^{-}(t + x/c) = K^{-} e^{j\Omega(t + x/c)} \qquad \text{eqtn.1.12}$$

Substituting these equations into eqtn.1.10 and applying the boundary condition p(l,t) = 0 at the lip end of the tube and eqtn.1.11 at the glottis end we can solve for the unknown constants K^{+} and K^{-}. The resulting sinusoidal steady state solutions are

$$p(x,t) = jZ_0 \frac{\sin[\Omega(l-x)/c]}{\cos(\Omega l/c)} U_G(\Omega) e^{j\Omega t}$$

eqtn.1.13

$$u(x,t) = \frac{\cos[\Omega(l-x)/c]}{\cos(\Omega l/c)} U_G(\Omega) e^{j\Omega t}$$

where

$$Z_0 = \rho c/A \qquad \text{eqtn.1.14}$$

is by analogy to transmission line theory called the characteristic acoustic impedance of the tube. The frequency response allows us to determine the response of the system to arbitrary inputs, not only sinusoids, through the use of Fourier analysis. For more realistic models, including the effects of vocal tract losses and radiation at the lips, the reader is referred to Rabiner and Schafer[20].

1.6 Acoustic Analysis of Pathological Voice

Acoustic analysis as used in the vocal pathology literature and as mentioned above has come to mean any spectrum or waveform measurement taken from the digitised speech signal. The purpose of the present thesis is to investigate the currently available acoustic measures[2], to test their validity and to introduce new measures. A study of the presently available approaches has revealed that (1) they offer limited information for use in clinical investigations and (2) many measurement problems arise and that the separation of the acoustic indices into independent measures is not a simple issue[21]. More specifically, the most commonly used acoustic measures for diagnosis of vocal pathology are jitter, shimmer and the harmonic to noise ratio. However, several researchers have shown that these measures are not independent and therefore may give ambiguous information. For example, the addition of random noise causes increased jitter measurements and the introduction of jitter causes a reduced harmonic to noise ratio. The previous section was included in order to show the effect of the

vocal tract on the output radiated speech waveform. The effect of these tract resonances have been cancelled by various strategies using inverse filtering of a high fidelity true phase recording of the output airflow from the lips. Recent studies have shown that the glottal waveform may be estimated from tape recorded speech samples using a frequency domain parameter set[22]. Therefore more is being learnt about the glottal flow and hence vibratory pattern of the vocal folds in terms of spectral measurements. Hanson[16], Holmberg[23], Karlsson[24] and others have shown that many useful acoustic parameters can be obtained from the acoustic speech waveform. However, in order to provide spectral characterisation of the vibratory pattern in pathological voice types the effects of jitter and shimmer on the speech spectrum must firstly be removed.

These issues have been thoroughly addressed in this thesis and the foundation has been laid for future studies that will investigate the vibratory pattern of the vocal folds based on spectral evaluation of tape recorded data. Firstly, an attempt has been made to spectrally characterise the four perturbation measures of additive noise, random jitter, cyclic jitter and shimmer, therefore providing a means of taking quantitative perturbation specific measurements from the speech spectra. Secondly, novel analysis programs have been written in order to overcome the contaminating effects of the perturbation measures and therefore provide a means of assessing the vibratory characteristics of the vocal folds. Time domain measures have also been investigated and the indications are that this requires further study. It is hoped that these research efforts will complement the work of Hanson[16], Holmberg[23], Karlsson[24] and others in providing more reliable acoustic indices with which to investigate both the vocal mechanism and voice quality.

Another important issue is whether future improvement in modelling voice production (Flannagan[25], Fant[22], Hirano[26], Fujimura[27], Titze[28], Farley[29]) and enhanced acoustic analysis will be able to provide differential diagnoses with respect to organic and psychogenic disorders. This is not a simple question to answer directly, but what is definitely true is that improved modelling and analysis of pathological voice types will definitely occur. One can envisage the culmination of several research efforts dedicated to voice, providing, not in the too distant future, a 3-D computer model of the larynx where many physiological parameters relevant to voice are included and

manipulated and the user is provided with synthesis feedback and spectral information regarding the voicing possibilities associated with a given configuration. Images taken from patient larynges via cinematography or ultrasound could then be matched to the model and if after matching the synthesis sounds the same as the patient it could be assumed that the correct model has been obtained. If the synthesis sounded different further model alterations could be made until the synthesis matched. Having obtained the correct match, correct alterations could be made until 'normal' voice was obtained. However this is of course beyond the scope of this thesis and here we concern ourselves with developing new analysis techniques that differentiate between normal and pathological voice types.

1.7 Bibliography

1. Robbins, SD. A dictionary of speech pathology and therapy. Cambridge, Mass.: Sci-Art publishers, 1963
2. Emanuel, FW. And Sansone FE. Some spectral features of 'normal' and simulated 'rough' vowels. Folia Phoniat. 1969; **21**:401:415
3. Bielamovicz, S. et al Comparison of voice analysis systems for perturbation measurement. J. Speech Hear. Res. 1996; **39**: 126-134
4. Titze, IR. Towards standards in acoustic analysis of voice. NY: Raven Press J. Voice 1994; **8**: 1-7
5. Titze, IR. Definitions and nomenclature related to voice quality, In O. Fujimura and M. Hirano (Eds.), Vocal Fold Physiology : Voice quality control, Singular publishing group, San Diego, 1995; pp. 335-342
6. Deller JR. et al Discrete time processing of speech signals, New York:Macmillan, 1993, pp.110
7. Ladefoged, P. Discussion of phonetics : a note on some terms for phonation types, In O. Fujimura (Ed.), Vocal fold physiology : Voice production, mechanisms and functions, New York: Raven Press, 1988; pp.373-376
8. Laver, J. The phonetic description of voice quality. New York:Cambridge University Press, 1980.
9. Imaizumi, S. Annual Bulletin RILP, University of Tokyo, Tokyo, **19**: 179-190, 1985
10. Hammarberg, B and Gauffin, J. Perceptual and acoustic characteristics of quality differences in pathologic voices as related to physiological aspects, In O. Fujimura and M. Hirano (Eds.), Vocal Fold Physiology : Voice quality control, Singular publishing group, San Diego, 1995; pp.283-303
11. Aronson, AE. Clinical voice disorders. New York: Thieme, 1990
12. Hirano, M. Clinical examination of voice. New York: Springer-Verlag , 1981
13. Baken, RJ. Clinical measurement of speech and voice. London: Taylor and Francis Ltd., 1987

14. van den Berg, J. Myoelastic-aerodynamic theory of voice production. J. Speech and Hearing Res. 1958; **1**:3, 227-245.

15. Lieberman, P. Vocal cord motion in man. Ann. N.Y. Acad. Sci., **155**:28-38, 1968

16. Hanson, HM. Glottal characteristics of female speakers; Acoustic correlates. J. Acoust. Soc. Am. 1997; **101:**466:481

17. van den Berg, J. On the air resistance and Bernoulli effect of the human larynx. J. Acoust. Soc. Amer. 1957, **29**:626-631

18. Stevens, KN. Airflow and turbulent noise for fricative and stop consonants , J. Acoust. Soc. Am. 1971, **50**:1180:1192

19. Portnoff, MR. and Schafer RW. Mathematical considerations in digital simulations of the vocal tract, J. Acoust. Soc. Am. 1973, **53**:294

20. Rabiner L. and Schafer R. Digital processing of speech signals. Englewood Cliffs, N.J.: Prentice Hall, 1978

21. Hillenbrand, J. A methodical study of perturbation and additive noise in synthetically generated voice signals. J. Speech Hearing Res. 1987, **30**:448-461

22. Fant, G. and Lin, Q. Frequency domain interpretation and derivation of glottal flow parameters. STL-QPSR 1988, **2-3**:1-23

23. Holmberg, EB. Comparisons among aerodynamic, electroglottographic and acoustic spectral measures of female voice. J. Speech Hearing Res. 1995, **38**:1212:1223

24. Karlsson, I. Glottal waveforms for normal female speakers. J. Phon. 1986, **14**:415:419

25. Flanagan, JL.et al. Synthesis of speech from a dynamic model of the vocal cords and vocal tract. Bell System Tech. J. 1975, 54:485-506

26. Hirano, M. Morphological structure of the vocal cord as a vibrator and it's variations. Folia Phoniat. 1974, **26**:89-94

27. Fujimura, O. Body-cover theory of the vocal fold and it's phonetic implications. In K. Stephens and M. Hirano (Eds.) Vocal fold physiology. Tokyo:University of Tokyo Press 1981, pp.271-288.

28. Titze, I. Preliminaries to the body-cover theory of pitch control. J. Voice, 1988, **1**:314-319

29. Farley, GR. A biomechanical laryngeal model of voice f0 control and glottal width control. J. Acoust. Soc. Amer. 1996, **100**:3794-3812

Chapter 2

Experimental Apparatus, Technique and Data

2.1 Speech Analysis Environment

The equipment necessary to carry out speech analysis research is well within the budget resources of any university or speech therapy department[1]. The basic requirements are a standard personal computer (PC) with an additional plug-in I/O module and some means of recording the acoustic speech waveform. This comprises a surprisingly powerful analysis environment with dedicated digital signal processing (DSP) chips providing real-time processing and feedback if required at moderate extra cost. The system implemented in the present study is shown schematically in fig. 2.1.

2.1.1 Data Acquisition

Speech samples were recorded using a standard linear dynamic microphone (SONY F-VS3N, Tokyo, Japan) connected to a CT-W851R PIONEER double cassette deck tape recorder (Pioneer T-W851R, Tokyo, Japan). TDK chrome tape cassettes of 57dB

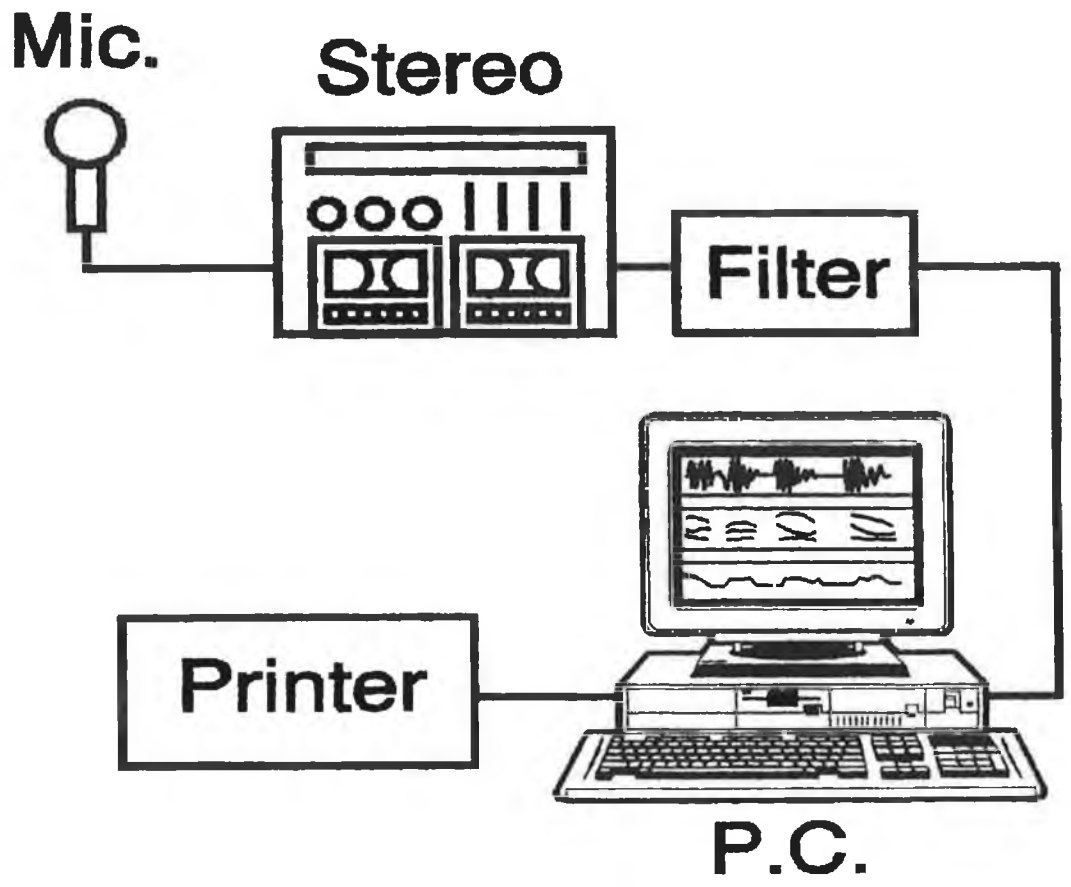


fig.2.1 *Schematic Diagram of Speech Analysis System*

signal to noise ratio were used with subsequent playback through a stereo amplifier unit (Sony F170, Tokyo, Japan). Alternatively, direct digitisation was also possible, with the tape deck set in record mode. The resulting continuous time signal had then to be band limited prior to sampling in order to avoid aliasing, an unfortunate consequence of the well known sampling theorem[2]. An eight order Chebychev low pass filter[3,4] with -48 dB/octave roll off at 3.8 kHz and 2 dB ripple across the pass band was constructed for this purpose. The filter response was examined by applying signals in the frequency range from D.C. to 10 kHz from a Thurlby/Thandar TG220 2Mhz Sweep/Function Generator (Huntington, Camb., England) (fig. 2.2). This bandwidth limited analog signal could now be digitised.[5] A sampling rate of 10 kHz using pacer trigger mode conversion was chosen from the C software driver for a 14-bit resolution,

variable sampling frequency, data acquisition expansion card (Integrated Measurement Systems PCL-814, Southampton, UK) installed in an 80486DX LEO PC.

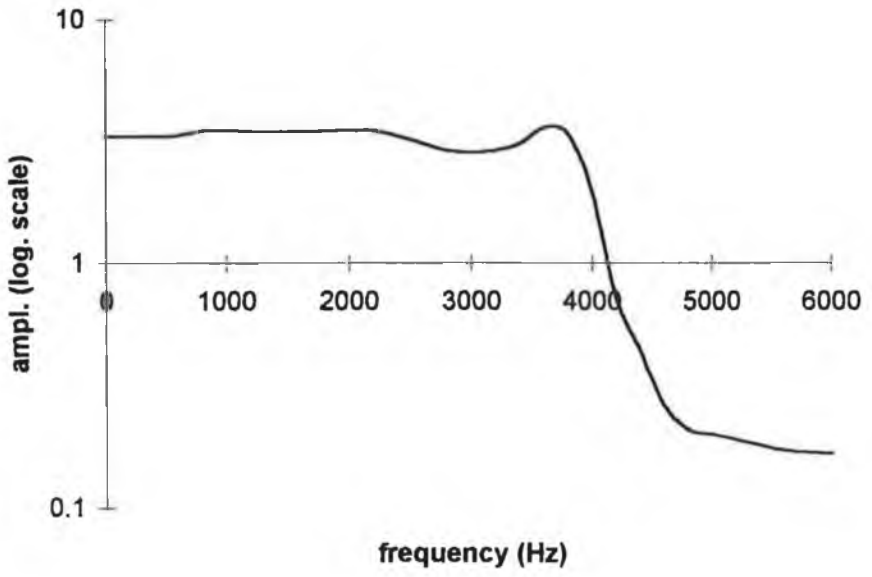


fig.2.2 *Frequency response of Chebychev low pass filter*

The resulting digitised samples were stored in 2's complement (integer type) binary form in two separate data buffers giving a total sample length of approximately 6.5 seconds. The data was then routinely saved to disk in binary file format for subsequent analysis.

2.1.2 Software Programming

Both Borland's Turbo C++ (Scott's Valley CA, USA) and Matlab (The Math Works Inc., Natick, Mass., USA) programming environments were used for analysis. In the case of Turbo C++ the compiler was a DOS application operating in an Integrated Development Environment. The project file option available with this compiler made for efficient programming with separately compiled files being linked together at run time. For the present application the main modules of a project file generally consisted of a) the software driver for the A/D card, b) the FFT radix-4 algorithm from Numerical Recipes in C (Cambridge University Press, Portchester, CA, USA) [6,7] and c)

the user written specific application program. The main user written C++ analysis program files were used for spectrogram production purposes.

The Window's based Matlab technical computing environment was introduced at a later stage and greatly decreased the time necessary for coding the required analysis and display algorithms owing to it's high level language interface. The accompanying Digital Signal Processing Toolbox (The Math Works Inc., Natick, Mass., USA) with it's specialised DSP functions was also obtained to provide the complete analysis system. The final versions of the principal analysis files written in Matlab are given in appendix A.

2.2 Recordings

a) Speaker Identification Experiment

All recording were taken in a quiet room in the college using the analysis equipment as outlined in paragraph 2.1.1. Experimental details are given in the next chapter as appropriate, alongside the description of the speaker identification experiment.

b) Diagnostic Investigations

The above recording procedure could not be followed in the clinical setting. Here, recordings were made of the participants phonating the sustained vowel a/ and uttering the phonetically balanced sentence "Joe took father's shoe bench out" at their comfortable pitch and loudness level. All recordings were taken by a member of the research group using a Tandberg audio recorder (AT 771, Audio Tutor Educational, Japan) prior to the participants (thirteen in all) undergoing laryngovideostroboscopic (LVS, Endo-Stroboskop, Atmos, Germany) evaluation[8] at the outpatient's ENT clinic in Beaumont Hospital, Dublin. The videostroboscopic evaluation was carried out by the otolaryngologist in collaboration with the speech therapist. The LVS system (fig.2.3) consists of a rigid endoscope which is guided posteriorly through the patient's

oropharynx until a clear image of the vocal folds is obtained. The patient is asked to phonate the vowel a/ or i/ while under examination. Illumination is provided via strobe pulses reflected from a laryngeal mirror attached to the end of the endoscope.

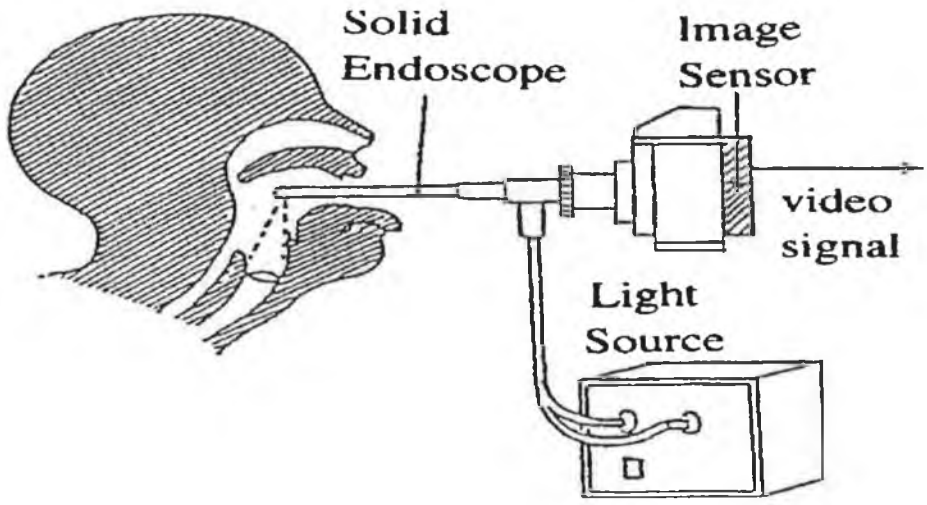


fig.2.3 *Schematic Diagram of Laryngeal examination using videostroboscopy*

When the pulse rate 'matches' the pitch frequency a clear video image is obtained of the vocal fold vibratory pattern. Supra-laryngeal structures may also be viewed. Hence, it provides a site specific, quantifiable assessment of the larynx. A word of caution is needed however, in respect to interpreting these images, especially in cases involving vocal pathology. As the images were obtained under strobe lighting, the apparent glottal cycle that the observer views are taken over several cycles (typically 24) of actual vocal fold movements. Therefore, there is an inherent assumption that the signal source is periodic which is clearly not the case with many vocal fold pathologies. So what results in one apparent cycle may have come from several cycles which vary widely and in many cases obtaining an image is not possible as was the case here. Along with the results of the stroboscopic examination, full medical details regarding the vocal pathology were taken for each patient as well as any further diagnostic comments at the time of assessment. Patient details are outlined in table 2.1.

PATIENT NO.	AGE	SEX	PATHOLOGY
1	39	f	Vocal cord nodules (bilateral)
2	70	m	hoarseness
3	43	f	vocal cord oedema /nodules
4	33	m	vocal nodule
5	22	f	left vocal cord nodules (bilateral)
6	22	f	hoarseness (on/off)
7	43	m	verucous carcinoma of both folds
8	65	m	Hyperkeratosis and parakeratosis
9	57	f	mild swollen vocal cords
10	74	m	carcinoma post-radiation right vocal cord immobile
11	23	f	left vocal cord palsy Immobile- well compensated right cord
12	54	m	laryngeal papilloma ptosis
13	57	f	abductor palsy

Table 2.1 *Patient listing and details. Mean age 46.3 , std. dev.20.6*

The audio (and video) data from the stroboscopic evaluation was recorded using SONY SVHS (E-180, France) cassettes. Twelve normals were subsequently recorded under the same conditions.

2.3 Aim of Acoustic Evaluation

From the data recorded in 2.2 a number of investigations are possible

1) To separate the patients and normals based solely on acoustic analysis of the audio recording[9,10].
2) To correlate the acoustic findings with assessments based on the stroboscopic assessment[11,12] , the overall medical evaluation or a perceptual evaluation[13,14].
3) To assess the effects of the endoscope on normal phonation.

This thesis reports the results of investigation number one above. Number two could not be attempted, unfortunately, due to lack of viewing facilities in the case of the LVS recordings and no GRBAS scale rating[15] or equivalent in the case of the perceptual evaluation. However, a simple perceptual rating scale, based on a system proposed by Hammarberg et al was used for both the patient and normal data (Table 2.2 and 2.3) in order to provide a more complete assessment with respect to number one above. Part three forms the basis of ongoing research, the results of which will be presented elsewhere.

NORMAL NO./ VOICE QUALITY	1	2	3	4	5	6	7	8	9	10	11	12
Normal (Quality)	✓	✓	✓	✓		✓		✓	✓	✓	✓	✓
Breathy					✓		✓					
Hyperfunctional							✓					
Roughness					✓							
Unstable Pitch/					✓							

Table 2.2 *Perceptual Evaluation for 'Normals'. Mean age 26.5, std. dev.3.5*

PATIENT NO./ VOICE QUALITY	1	2	3	4	5	6	7	8	9	10	11	12	13
Aphonic													
Breathy				✓			✓	✓	✓		✓		
Hyperfunctional	✓			✓		✓	✓	✓	✓	✓		✓	
Hypofunctional													
Fry/Creaky		✓	✓										
Roughness			✓				✓			✓		✓	
Gratings							✓			✓		✓	✓
Unstable pitch		✓	✓				✓			✓			
Voice breaks			✓			✓				✓	✓		✓
diplophonia											✓		✓

(a)

PATIENT NO./ VOICE QUALITY	1	2	3	4	5	6	7	8	9	10	11	12	13
Aphonic													
Breathy	✓			✓			✓	✓					
Hyperfunctional			✓		✓	✓		✓	✓	✓	✓	✓	
Hypofunctional				✓					✓				
Fry/Creaky		✓	✓										
Roughness							✓			✓		✓	
Gratings													
Unstable pitch		✓	✓			✓	✓			✓			
Voice breaks			✓				✓			✓	✓		
diplophonia											✓		✓

(b)

Table 2.3 *Perceptual Evaluation for patients* a) *Therapist 1* b) *Therapist 2*

In the case of the normal data, two of the 'normals' showed deviant voice qualities, the equivalent of a rating of one on a five point scale where zero represents normal and four indicates severe dysphonia. All patient data show deviant qualities (as rated by two speech therapists) but unfortunately the degree is not given and therefore the perceptual ratings were used simply as a accompaniment to the acoustic findings, rather than as the basis for correlation.

2.4 Vowel Synthesis

In order to test the analysis programs for evaluation of vocal pathology in a systematic way vowel synthesis data files were produced. These were designed to simulate various commonly found acoustic characterisations of vocal pathology such as jitter and shimmer. The discrete time system model for speech production[16,17,18] shown in fig. 2.4(a) forms the basis for this approach.

Adequate synthesis can be performed using this model to produce continuous speech where the vocal tract parameters vary with time as appropriate. In order to introduce the various perturbation measures certain adjustments to the model are required as shown in fig. 2.4(b). Instead of simply replacing the traditional exclusive OR gate switch for voiced/unvoiced excitation with an OR gate in order to simulate conditions of turbulent glottal flow concurrent with normal voicing, a signal dependent random noise component was introduced at the glottal source as shown in part (b) of the figure. The noise component was introduced in this manner in acknowledgment of the fact that for voiced fricatives, frication is correlated with the peaks of the glottal flow. What is the most pertinent way to represent the noise component for conditions involving vocal pathology is uncertain and is most likely somewhat variable depending on the specific pathology under investigation and certainly merits further study. The vocal tract parameters are kept constant in order to produce a sustained vowel. A randomised gain factor may also be added to the impulse train generator in order to produce amplitude perturbation of the glottal source. Each stage of the model is

examined below but we can get some idea of the complexity of the problem posed by acoustic analysis of vocal pathology through examining fig. 2.4 (b) and considering that we are trying to reveal or separate (if possible) the source of the abnormality introduced at A), B), C) or D) by analysing a signal that has been convolved with the vocal tract response and radiated at the lips. Furthermore, the model assumptions of source/tract separability and non time-varying vocal tract parameters are only approximately correct, even in the case of a sustained vowel phonation.

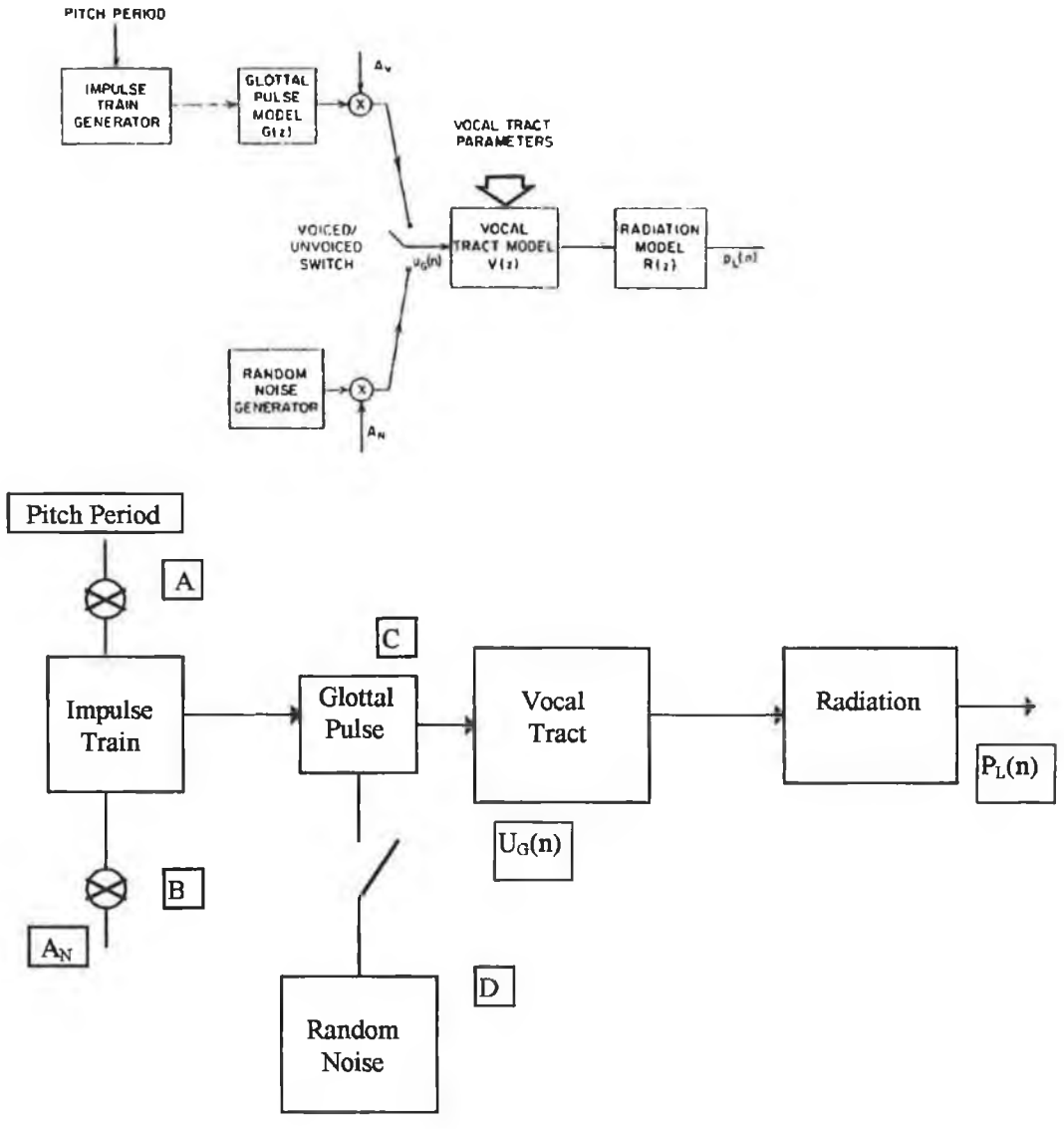


fig.2.4 (a) *Discrete time system model for speech production and* (b) *modification of the model for use in investigation of vocal pathology*

2.4.1 Excitation

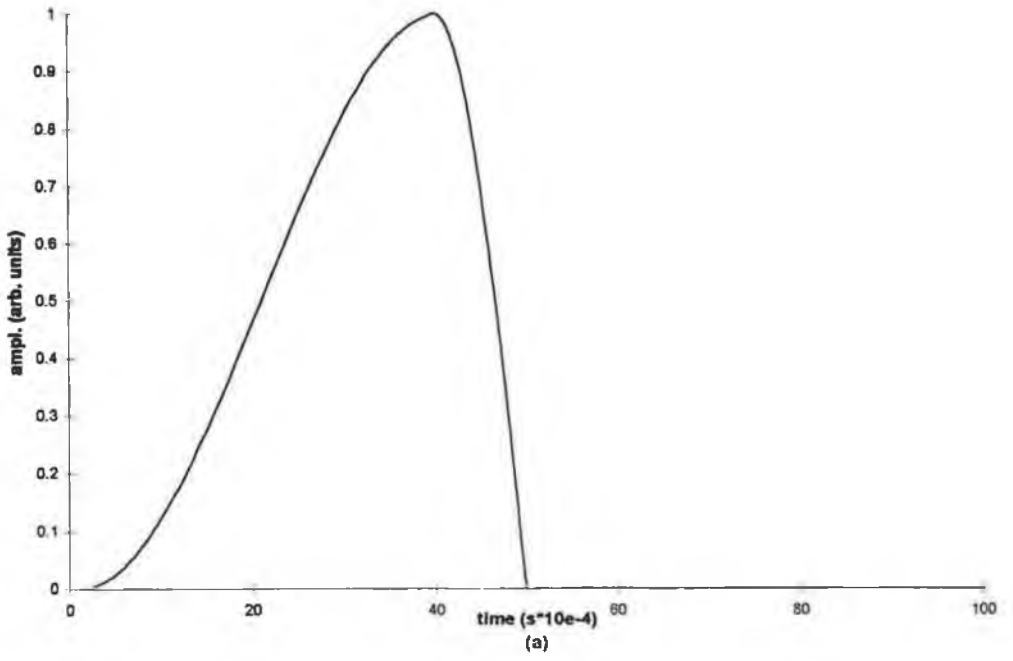
The Rosenberg glottal pulse model[19] incorporates most of the important features of glottal waves estimated by inverse filtering and by high speed motion pictures and takes the form

$$\begin{aligned} g_r(n) &= 1/2[\,1-\cos(\pi n / N_1)] && 0 \le n \le N_1 \\ &= \cos(\pi(n-N_1)/2N_2) && N_1 < n \le N_1 + N_2 \\ &= 0 && \text{otherwise} \end{aligned}$$

eqtn.2.1

The pulse wave shape and it's Fourier transform magnitude are shown in fig. 2.5 for typical values of N1 and N2. To create a sequence of such wave shapes an impulse train generator produces a sequence of unit impulses which are spaced by the desired fundamental period. This sequence is then convolved with the glottal pulse shape in order to produce the desired repetitive waveform. Since it is our goal to study abnormalities of the voicing source it is here at the glottal source that we introduce the perturbation measures. Three parameters which have received a lot of attention in the vocal pathology literature[20,21], namely, shimmer, additive noise and jitter were introduced. Firstly, shimmer, which can be defined in general terms as the variation in amplitude of the glottal source from period to period was introduced simply by adding a random variable gain factor to the impulse train prior to convolution with the glottal pulse, as shown marked 'A' in fig. 2.4 (b). This variation in amplitude was implemented using Matlab's random number generator 'randn.m' which produces a Gaussian distribution of random numbers with a mean of zero and a variance of one. Therefore, in order to introduce a standard deviation of a given percent, denoted by 'per', a calculation similar to the following was implemented

$$A' = A \times \left(100 + per \times randn(t)\right)/100 \qquad \text{eqtn.2.2}$$



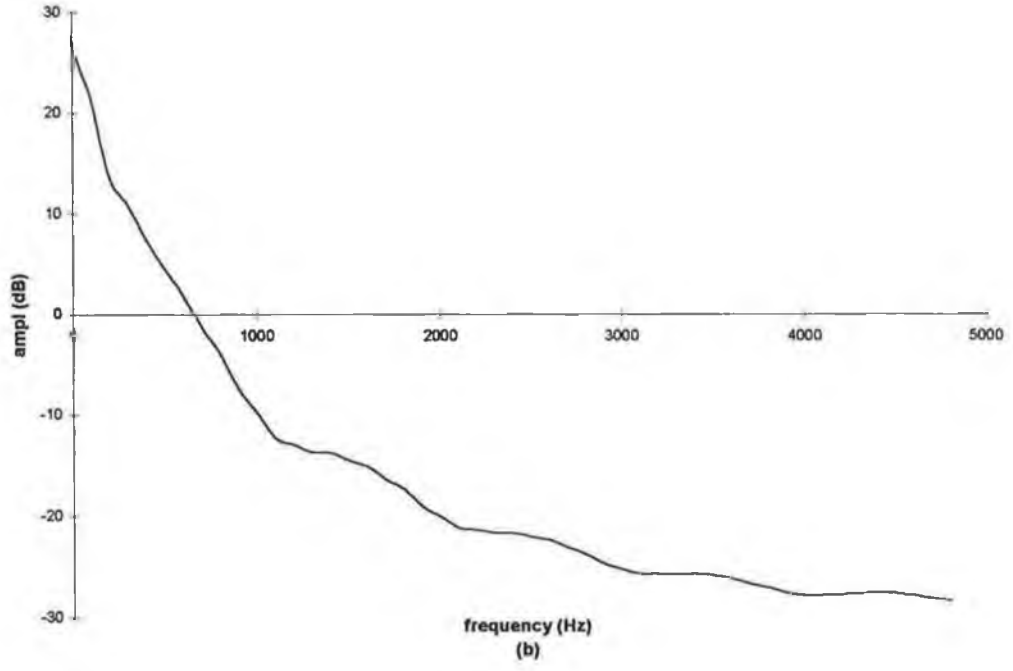


fig.2.5 (a) *Rosenberg glottal flow waveform and* (b) *it's Fourier power spectrum*

The random variation of the impulse amplitude with 'per' set at sixteen is shown in figure 2.6 (a) and a histogram of the variation is shown in fig. 2.6 (b). Three amplitude perturbed glottal waveforms are shown in fig. 2.7 for 'per' values of 4, 8 and 16. A 'per' value of 4 for example means that an originally constant amplitude of the glottal

source, 'A', now has a Gaussian distributed amplitude with a standard deviation equal to 4% of the original amplitude.

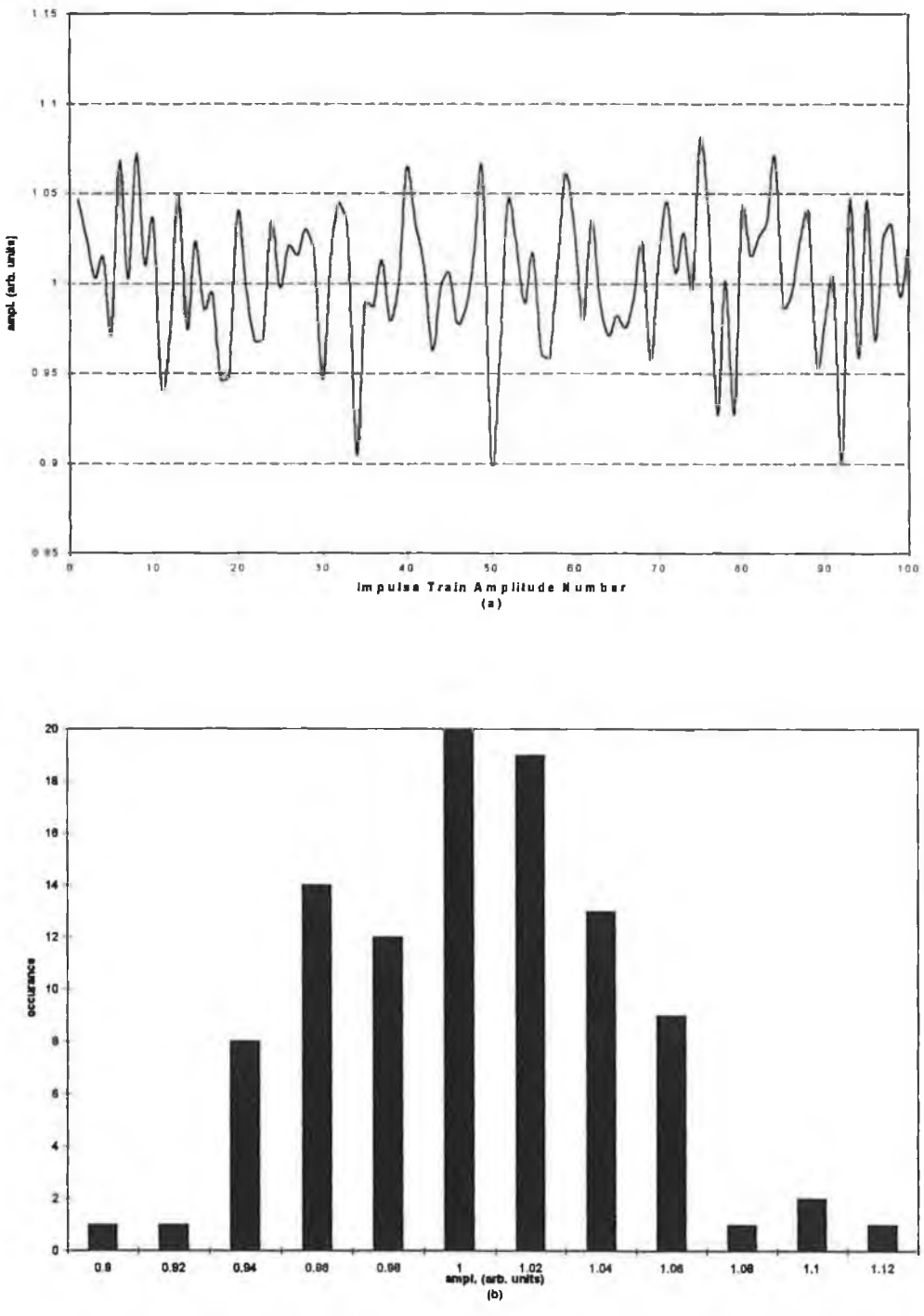
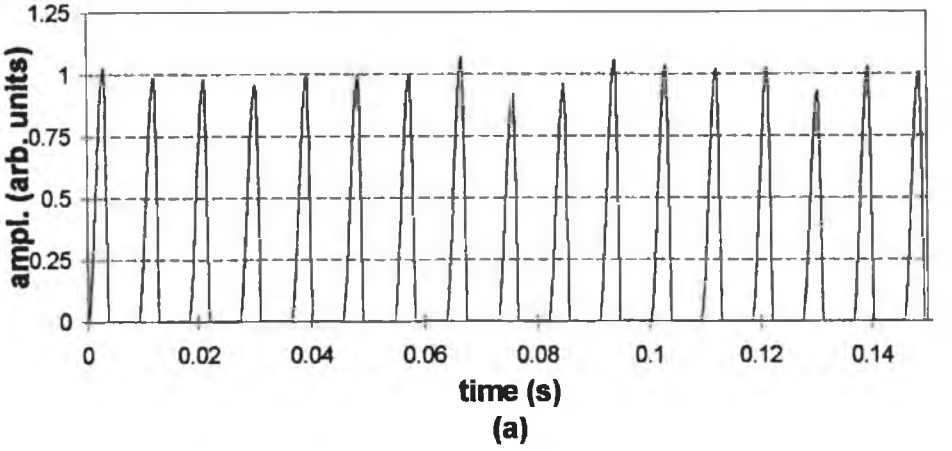
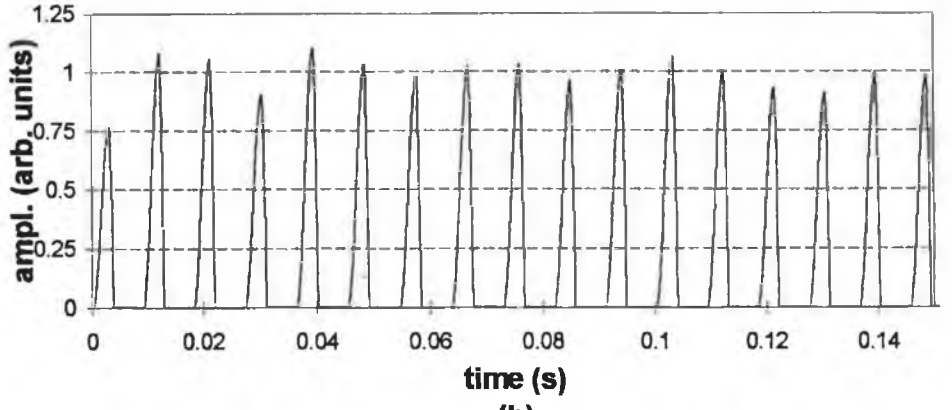


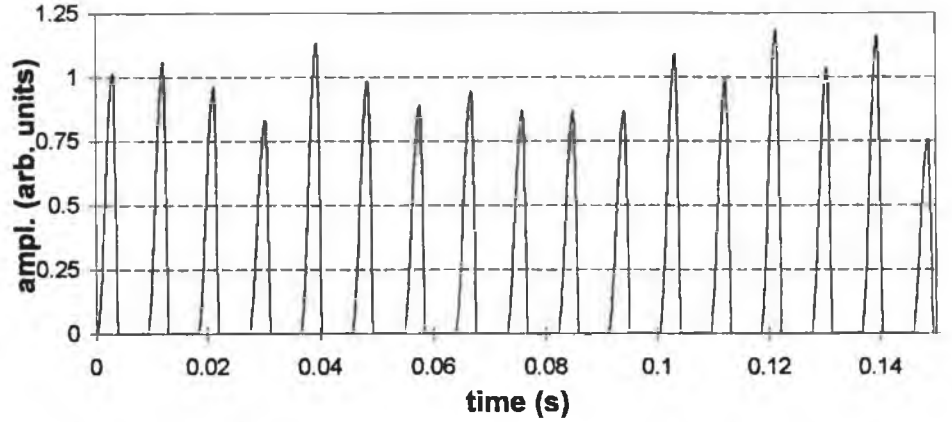
fig.2.6(a) *Random variation of amplitude of impulse train and* (b) *histogram of the variation*



(a)



(b)



(c)

fig.2.7 *Glottal source waveforms with* a) *std. dev. 4%* b) *std. dev. 8% and* c) *std. dev. 16 % amplitude perturbation*

The introduction of random noise and random pitch perturbation followed a similar strategy. Random additive noise was introduced by multiplication of the glottal pulse waveform by a random noise generator arranged to give signal dependent additive noise of a user specified variance, denoted 'per' in the Matlab program 'synadnoq.m'. The noise was added according to the following equation

$$g_r' = g_r \times \left(100 + \text{per} \times \text{randn}(n)\right) / 100 \qquad \text{eqtn.2.3}$$

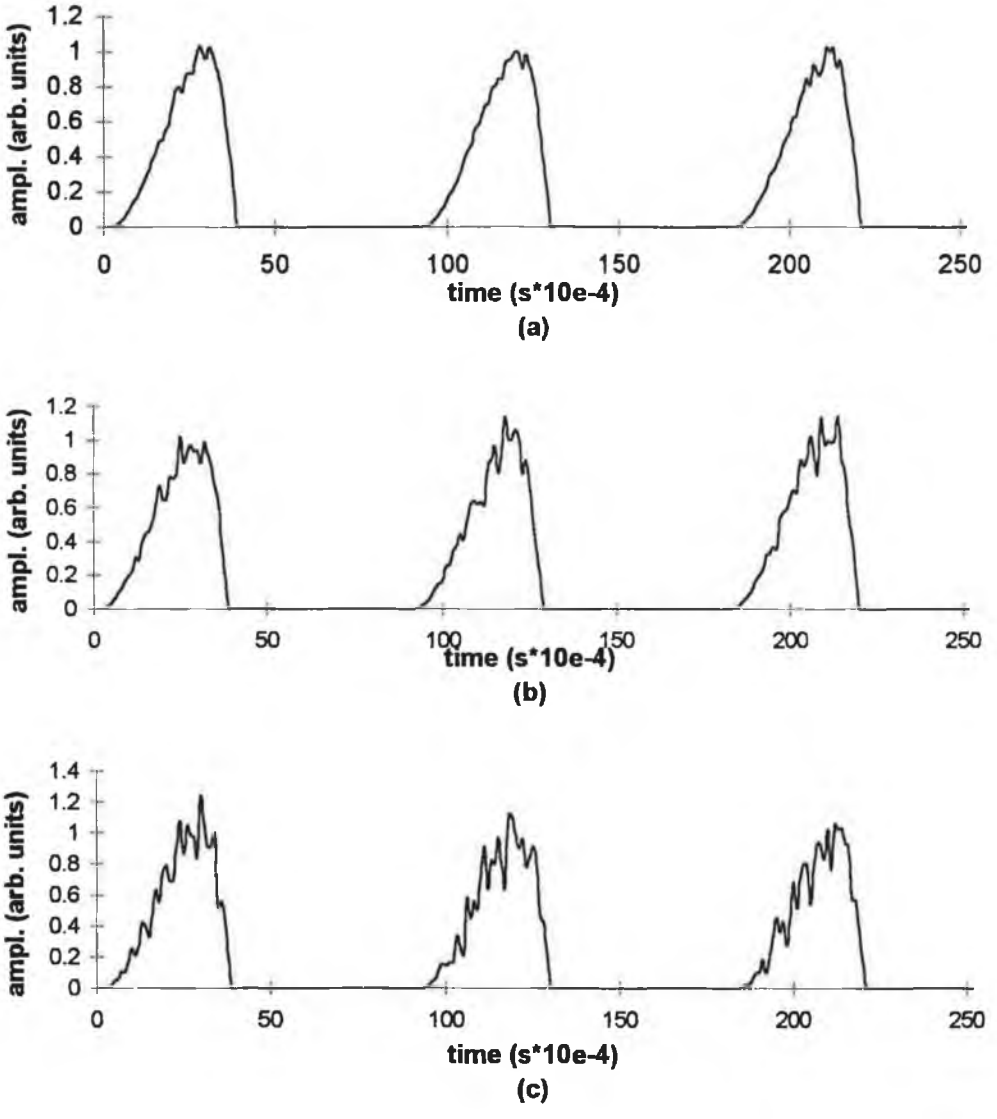


fig.2.8 *Signal dependent, random, additive, Gaussian noise* a) *std. dev. 4 %* b) *std. dev. 8 % and* c) *std. dev. 16 %*

As a result of this, greater noise occurs at peak flow but the signal to noise ratio remains constant at all points along the waveform during the open phase. Three additive noise levels of standard deviation 4, 8 and 16 percent are shown for the 110 Hz file in figure 2.8. Applying the noise in the above manner insures that the closed phase remains unaffected by the noise.

Finally, jitter, the variation in the pitch period from cycle to cycle was introduced. Two variations were implemented (fig.2.9). Firstly, cyclic variation of the pitch period, e.g. varying the period from say 100 Hz to 104 Hz to 100 Hz and repeating in this fashion. Secondly, the period was varied in the more usual random ordering e.g. 102 Hz, 98 Hz, 101 Hz, 103 Hz etc. The cyclic jitter was introduced in order to investigate a proposal by *Gauffin et al*[22] that conventional jitter measurements indicating the same value may arise from vocal pathologies with very different etiologies.

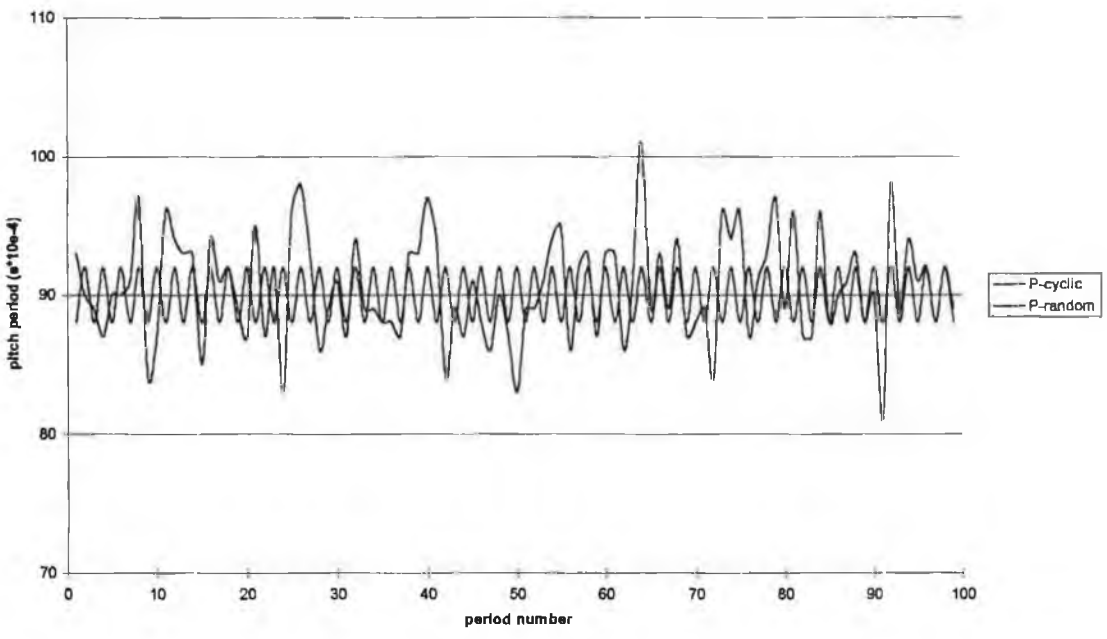


fig. 2.10 *Pitch period variation for conditions of cyclic and random jitter*

It should also be noted that the open quotient (OQ - the ratio of the glottal open period to total glottal period) which can affect the glottal spectrum was kept constant during

all the above perturbation variations, as opposed to simply truncating the period as in often done in multi-pulse resynthesis when introducing pitch perturbation. We avoid this approach as we want to vary f0 as an independent parameter of change, with a view to future studies that would focus on events that occur within the glottal cycle. This corresponds to the source of variation introduced at B in the schematic diagram of fig. 2.4 (b). Ananthapadmanabha[23] has shown that the open quotient is inversely proportional to the harmonic ratio, where the harmonic ratio (HR) is defined as the ratio of the amplitude of the second to the first harmonic. Therefore, had we simply truncated the closed phase we would have changed the spectral content of the signal as a result of a change in open quotient as opposed to simply an f0 increase.

2.3.2 Vocal Tract Model

These glottal pulses are now used to excite the vocal tract, the transmission properties of which in our digital model are based on the behaviour of a set of concatenated lossless acoustic tubes as shown in fig. 2.11.

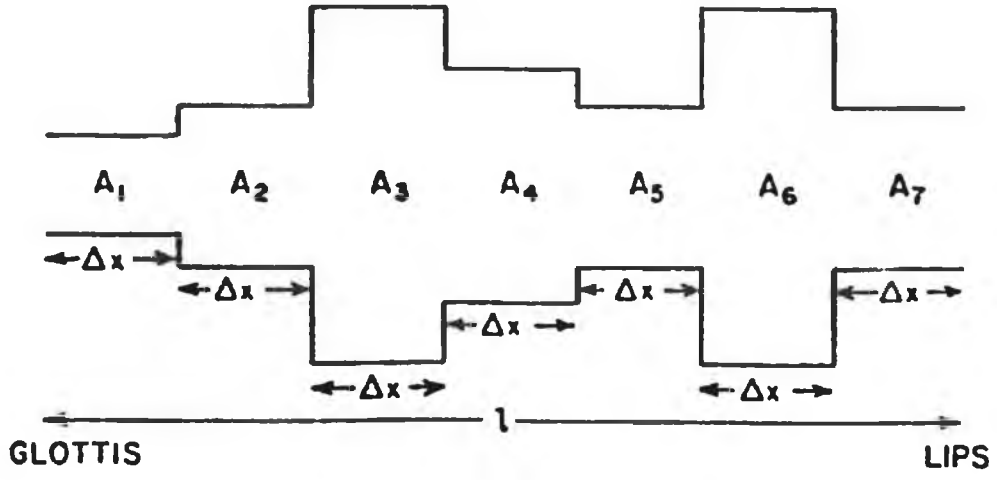


fig.2.11 *Concatenated Lossless Tube Model*

Portnoff[24] has shown that sound waves in a tube satisfy the pressure/volume velocity relationship

$$p_k(x,t) = \frac{\rho c}{A_k}\left[u_k^{+}(t-\frac{x}{c}) + u_k^{-}(t+\frac{x}{c})\right]$$

eqtn.2.4

$$u_k(x,t) = u_k(t-\frac{x}{c}) - u_k(t+\frac{x}{c})$$

p_k = pressure at k^{th} tube

u_k = volume velocity at k^{th} tube

ρ = density of air

where x is the distance measured from the left hand end of the k^{th} tube ($0<x<l_k$) and u_k+() and u_k-() are positive going and negative going travelling waves in the k^{th} tube.

Lossless and plane wave propagation assumptions, along with boundary conditions at the tube junctions obtained by applying the physical principle that pressure and volume velocity must be continuous in both time and space everywhere in the system, give rise to relatively straight forward solutions of the resulting equations, known as the Kelly/Lochbaum equations.[25] These equations can be usefully depicted using signal flow graph conventions[16] where $\tau = \Delta x/c$ is the one way propagation of the sections (fig.2.12a). This representation (or equivalently from fig.2.11) implies that the lossless tube models have properties in common with digital filters. An equivalent discrete time lattice filter is shown in part b) of the figure.

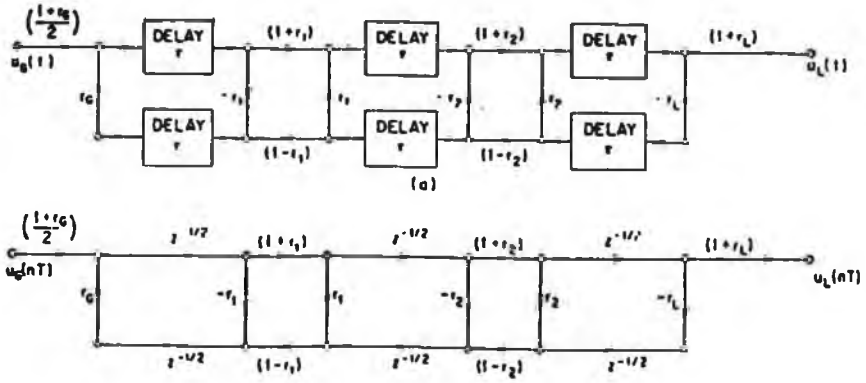


fig.2.12 a) *Signal flow graph for lossless tube model of the vocal tract;* b) *equivalent discrete time system*

For a discrete time vocal tract model consisting of a concatenation of N lossless tubes of equal length the system function is

For a discrete time vocal tract model consisting of a concatenation of N lossless tubes of equal length the system function is

$$V(z) = \frac{\prod_{k=1}^{N} (1 + r_k) z^{-N/2}}{D(z)} \qquad \text{eqtn.2.5}$$

where the denominator D(z) is obtained from the polynomial recursion

$$D_0(z) = 1$$
$$D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}) \qquad k = 1,2, \ldots, N$$
$$D(z) = D_N(z)$$

where the r_k's are the reflection coefficients at the tube junctions,

$$rk = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}$$

and it is assumed that there are no losses at the glottis and that all losses are introduced at the lip end through the reflection coefficient

$$r_N = r_L = \frac{A_{N+1} - A_k}{A_{N+1} + A_N}$$

The system function can also be written in the form of a direct-form difference equation as

$$V(z) = \frac{G}{D(z)} = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}} \qquad \text{eqtn.2.6}$$

Hence, given a set of area data the system function can be obtained. Fant[26] has supplied such data obtained from x-ray images of the phonation of the Russian vowel AA (Table 2.4).

SECTION	1	2	3	4	5	6	7	8	9	10
vowel AA	1.6	2.6	0.65	1.6	2.6	4	6.	8	7	5

Table 2.4 Vocal tract area data for Russian vowel AA (cm^2)

Radiation at the lips is simply modelled by the first order difference equation $R(z) = (1-z^{-1})$ to supply the final ingredient in our model. The waveform for the vowel AA at 110 Hz is shown in fig.2.13.

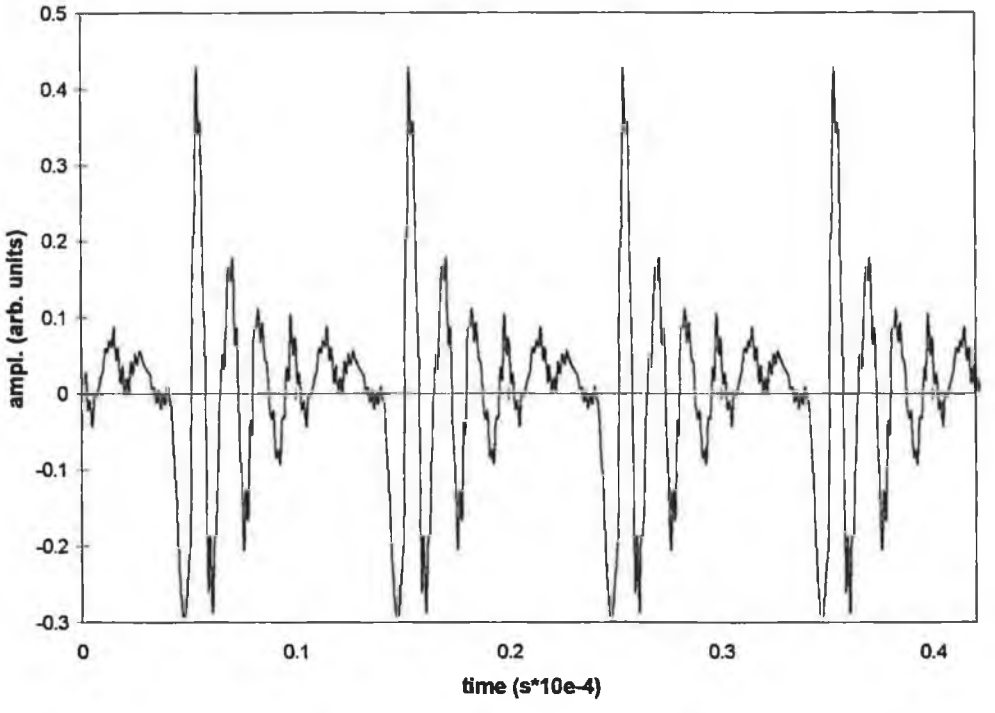


fig.2.13 *Synthesis vowel AA - Glottal pulse of fig.2.5 filtered using the digital model of the vocal tract transfer function.*

2.4.2 Vowel Data

The actual implementation of the above synthesis model was performed using a user written program (synthOQ.m) with calls to the signal exercises library for the AtoV function and Matlab's filter function. The program listings are given in appendix A. Tables 2.5 and 2.6 give a complete list of the data files produced to provide a means of testing and calibrating the subsequent analysis programs.

Table 2.5 *List of synthesis data for 110 Hz signal*

RANDOM JITTER (STD DEV.)	PERIODIC (OR CYCLIC) JITTER (%)	ADDITIVE NOISE (STD DEV.)	SHIMMER (STD DEV.)
1	1	1	1
2	2	2	2
3	3	4	4
4	4	8	8
5	5	16	16
6	6	32	32

Table 2.5 *List of synthesis data for 220 Hz signal*

RANDOM JITTER (STD DEV.)	PERIODIC (OR CYCLIC) JITTER (%)	ADDITIVE NOISE (STD DEV.)	SHIMMER (STD DEV.)
1	1	4	1
3	3	8	4
5	5	16	16

Further files were also created for three levels of noise for signals beginning at 80 Hz and increasing in six, approximately equi-spaced steps of 60 Hz up to 350 Hz.

2.5 Bibliography

1. Curtis, J.F. An introduction to microcomputers in speech, language and hearing, Little Brown and Co., 1987
2. Oppenheim, A.V. and Schafer, R.W. Discrete-time signal processing. Englewood Cliffs, N.J.: Prentice Hall, 1989
3. Lancaster, D. "Low-pass and high-pass filter responses", in The active filter cookbook. Indianapolis: Howard W. Sams & Co., 1987
4. Horowitz, P. and Hill, W. The Art of Electronics. Cambridge Mass.:Cambridge University Press, 1986
5. Oppenheim, A.V. Applications of digital signal processing. Englewood Cliffs, N.J.: Prentice Hall, 1978
6. Press, W.H. et al. Numerical Recipes in C. Cambridge, USA: Cambridge University Press, 1992
7. Sorensen, H.V. et al. Real valued fast Fourier transform algorithms. IEEE Trans. on Acoustics, Speech and Signal Processing 1987; VOL. ASSP.35. N0. 6
8. Koike, Y. and Imaizumi, S. Objective Evaluation of Laryngostroboscopic Findings. In O. Fujimura (Ed.) Part VII Vocal fold physiology, Vol 2. NY:Raven Press, 1988
9. Valencia-Naranjo, N. Diagnostic voice disorders, Current Opinion in Otolaryngology and Head and Neck Surgery, 1995 **3**:164-168
10. Fex, B. et al. Acoustic Analysis of Functional Dysphonia: Before and After Voice Therapy (Accent Method). J. Voice **8**:163-187
11. Sulter, A. et al. Standardised laryngeal videostroboscopic rating: Differences between untrained and trained male and female subjects, and effects of varying sound intensity, fundamental frequency, and age" J. Voice **10**:2:175-189
12. Woo, P. et al. "Aerodynamic and stroboscopic findings before and after microlaryngeal phonosurgery. J. Voice **8**:2:186-194
13. Hammarberg, B. and Gauffin, G. Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects. In O. Fujimura and M. Hirano (Eds.), Chapter17 Vocal Fold Physiology: Voice Quality Control San Diego:Singular Publ. Group, 1995

14. Lee, C.K. and Childers, D.G. Some acoustical, perceptual and physiological aspects of vocal quality", In J. Gauffin and B. Hammerberg (Eds.), Vocal fold physiology: acoustic, perceptual and physiological aspects of voice mechanisms. San Diego:Singular Publishing, 1991

15. Imaizumi, S. Acoustic measurement of pathological voice qualities for medical purposes, ICASSP, Tokyo, IEEE, 1986

16. Rabiner, L. and Schafer, R. Digitial processing of speech signals, Englewood Cliffs, N.J.: Prentice Hall, 1978

17. Deller, J. Proakis J. and Hansen, J. Discrete time processing of speech signals, NY: Macmillan, 1989

18. Burrus, C. et al. Computer-based exercises for signal processing using Matlab, Englewood Cliffs, NJ:Prentice Hall, 1994.

19. Rosenberg, A. Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Amer. 1971, **49**:583-590

20. Hillenbrand, J. "A Methotological Study of Perturbation and Additive Noise in Synthetically Generated Voice Signals", J. Speech Hear. Res., **30**:448-461, 1987

21. Askenfelt, A. and Hammerberg, B. Speech waveform perturbation analysis: A perceptual-acoustic comparison of seven measures, J. Speech Hearing Res. 1986, **29**:50-64

22. Gauffin, J. et al, Irregularities in the voice: A perceptual experiment using synthetic voices with subharmonics in P. Davis and N. Fletcher (Eds.), Vocal fold physiology : controlling complexity and chaos, San Diego: Singular Publishing, 1996

23. Ananthapadmanabha, TV. Spectral parameters of a voice source pulse J. Acoust. Soc. Am. 1991, **90**:2345

24. Portnoff, MR. and Schafer RW. Mathematical considerations in digital simulations of the vocal tract, J. Acoust. Soc. Am. 1973, **53**:294

25. Wu, H. et al. Vocal tract simulation : Implementation of continuous variations of the length in a Kelly-Lochbaum model, effects of area function spatial sampling" IEEE,1987

26. Fant, G. Acoustic theory of speech production. Mouton, The Hague, 1970.

Chapter 3

Investigation into Speaker Identification Using Digital Speech Spectrograms

3.1 Introduction

The advance of modern telecommunications in major industrialised nations has been paralleled by an increase in the use of human speech as an instrument in committing crimes. The would be assailant has taken advantage of the fact that the use of the telephone provides a means of maintaining anonymity whilst committing a variety of offences such as kidnappings, terrorist attacks, obscene phone calls and hoax bomb threats.

Where live recordings exist of an actual crime an expert witness is called upon to decide whether (based upon scientific principles) the recorded voice is the same or different from that of the suspect or a list of suspects. Forensic speaker identification has presented many difficulties and much controversy has surrounded it's use due the serious nature of the implications of a false identification. In order to appreciate some of the difficulties involved in forensic speaker identification, a discussion is given of the problem in the context of the broader field of speaker recognition[1].

Speaker recognition is a generic term which refers to any task which discriminates people based upon their speech characteristics. The potential applications of speaker recognition have increased with developments in telecommunications and automatic information processing. Technological research has not been slow in developing such applications, providing a number of solid state devices for access control to high security installations such as military facilities, nuclear power stations and research laboratories. More recently, automatic telephone transaction control (e.g. telephone banking) and tele-monitoring of individuals on probation have been introduced with considerable success. The basis for each of these recognition strategies is generally the same. A person identifies himself/herself as a 'customer' by entering a personal code number and is then required to pronounce a test phrase taken from a limited combination of words. Following some sort of acoustic analysis, usually involving the use of the Long Term Averaged Spectrum, a feature vector (i.e. a vector whose elements consist of acoustic parameters that have been shown to carry speaker identifying features) is derived from the test signal and matched to vectors gained from earlier access claims of the person in question. A similarity index is then calculated and recognition is affirmed if a certain threshold is exceeded : if not the procedure is repeated or the person is regarded as an impostor. Typical error rates for such systems are less than one per cent for both false rejection and false identification so can we apply this technology to the forensic case ?

There are several factors which separate the above, so-called speaker verification task from the more difficult task of speaker identification, so, at present the answer is in the negative. Firstly, the verification task involves a co-operative speaker whereas in the forensic situation, one reason for oral communication is to conceal identity.[2,3] Secondly, there are no pre-selected phrases or vowels which are known to contain highly speaker-specific information in the forensic case. Thirdly, the vast majority of forensic cases involve telephone transmitted speech[4] where there exist several possibilities for degradation along the transmission path and the signal is bandlimited between 300-3400 Hz. Finally, the verification task typically involves close set comparisons (i.e. the unknown speaker is contained within the test set), whereas in the forensic situation the set of potential speakers is open. So, under forensic investigation conditions the use of automatic methods has not seriously been considered.

Three further approaches to the problem have been developed however, and are currently in use. The first is speaker identification as performed by a phonetician or speech scientist through auditory recognition, the second is based on a semi-automatic computer analysis approach and the third is based upon the visual comparison of speech spectrograms which will now be described in some detail.

3.2 The Speech Spectrogram

The speech spectrograph is a device for displaying how the acoustic patterns of speech vary with time. It was developed by Koenig, Dunn and Lacey[12] during the forties as part of the war effort in the US. A rich source of information on speech spectrograms is a book by Potter, Kopp and Green, entitled Visible Speech.[13] The spectrograph is used extensively in speech research today, providing useful information in areas such as acoustic phonetics and speech pathology. A spectrogram for the phrase "You and I have to go today" is shown in fig.3.1 with time plotted on the horizontal axis, frequency on the vertical and the speech energy is plotted as a grey scale.

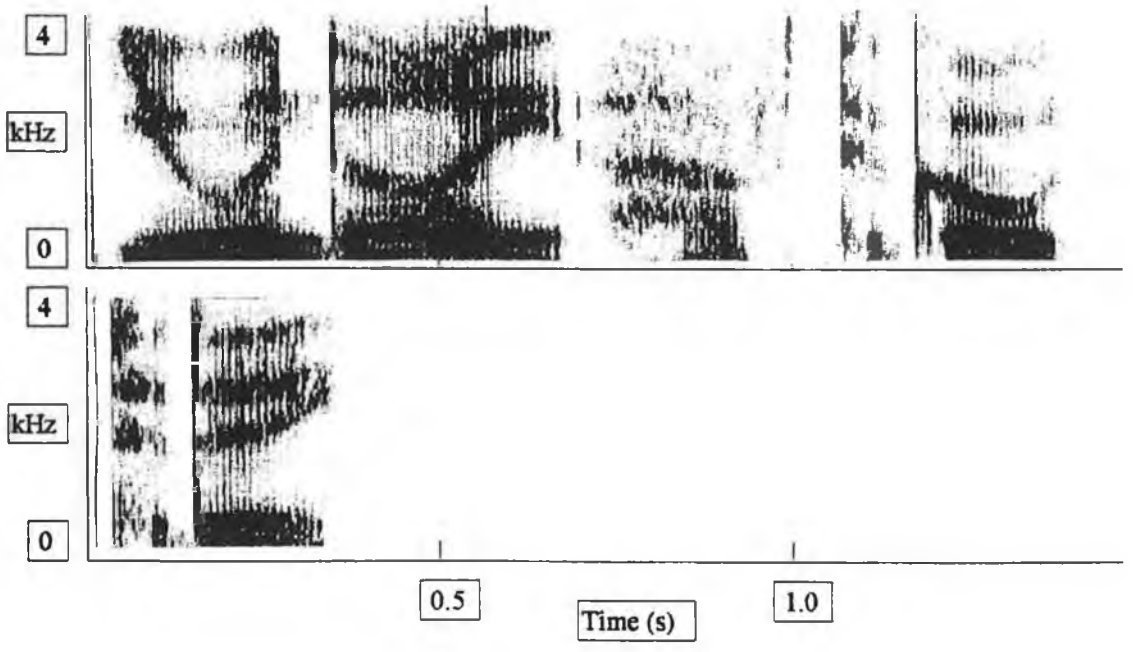


fig.3.1 *Spectrogram of the sentence "You and I have to go today".*

3.2.1 Spectrogram Production

a) Analog

The original speech spectrographic device consisted of a band-pass filter and a rotating drum as shown in fig.3.2. The acoustic speech waveform which was stored on a magnetic drum was played back repetitively into the heterodyning filter as the filter spanned the frequency range of interest. The output voltage or amplitude from the filter was burned onto teledeltos paper rotating on the drum. The amplitude was depicted rather crudely, due to the resolution limitations of the electro-sensitive paper. It took several minutes to produce a two second duration spectrogram in this manner.

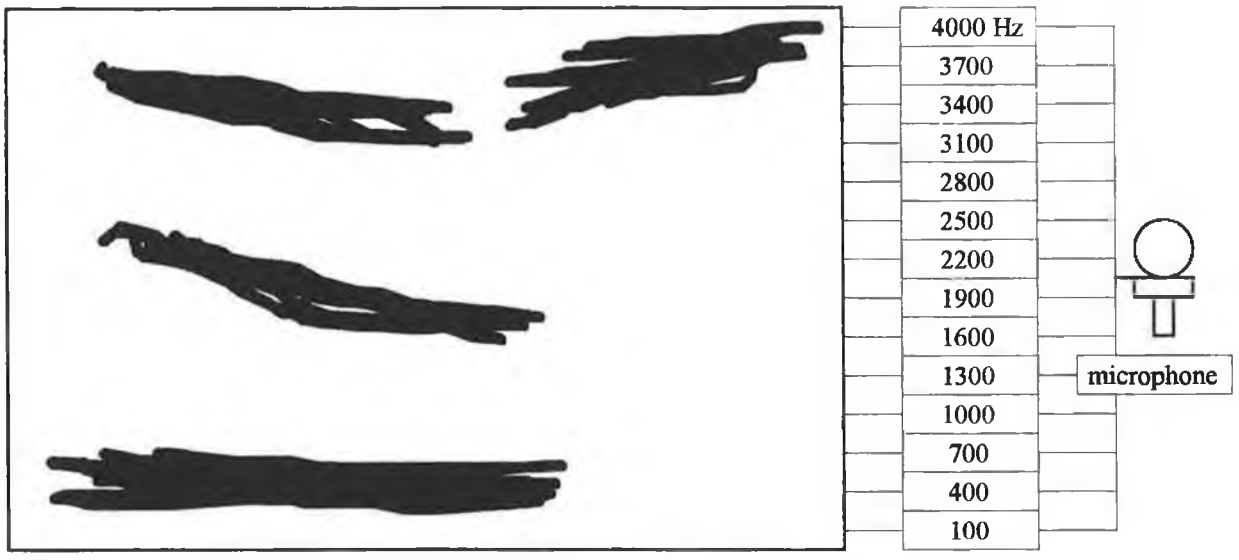


fig.3.2 *Schematic diagram of original spectrographic machine developed at Bell Laboratories.*

b) Digital

An alternative approach to the analog method using the Fourier transform became a computationally efficient alternative with the advent of the fast Fourier transform

(FFT) algorithm. Oppenheim[7] and Melmerstein[8] in 1970, were the first to suggest and implement such a digitally based spectrographic system. As technology developed, several improvements' were made to these original digital speech spectrograms and with the development of faster processors and dedicated DSP chips the real time voice spectrogram became a reality[9]. A brief review of the spectral analysis of speech production serves to illustrate the motivation behind the digital spectrogram.

c) Spectral Analysis of Speech

The process of speech production can be modelled as a linear system with either a quasi-periodic (voiced) excitation or a random noise (unvoiced) source as input. The system function is the response of the vocal tract to this input. For the production of a given phoneme e.g. a vowel sound, the vocal tract can be considered to remain stationary, giving rise to a given resonant condition. The system function can then be viewed in terms of the impulse response or the frequency response of the vocal tract to this quasi-periodic input (fig. 3.3).

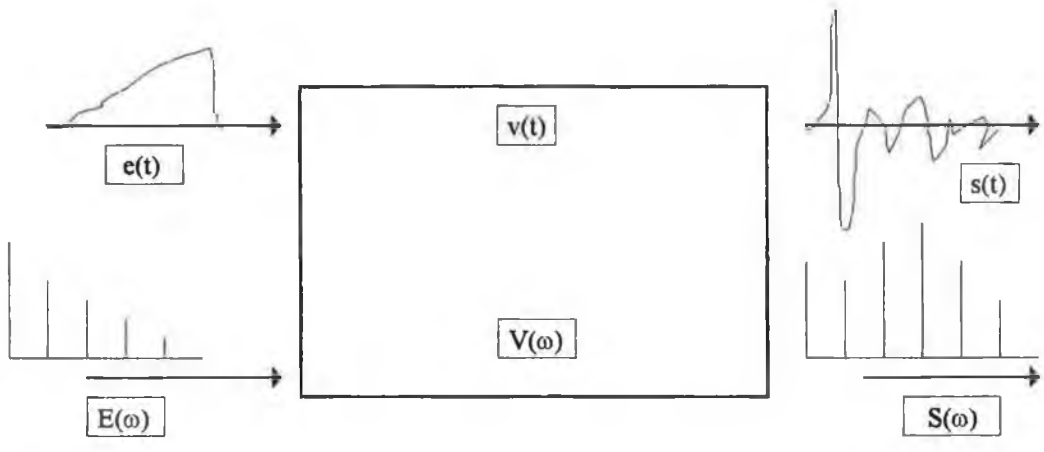


fig.3.3 *Source/filter model of speech production illustrating system characterisation via the impulse (s(t)=v(t)*e(t)) (* denotes convolution) and frequency response (S(ω)=E(ω)×V(ω)), where s(t) is the output waveform, e(t) the excitation source and v(t) the vocal tract filtering. Capital letters denote Fourier transform of time domain counterparts.*

The input corresponds to the glottal waveform or the volume velocity at the vocal cords and the output corresponds to the volume velocity at the lips. If, for example, the vowel produced was the vowel /i/ as in bee, a typical set of values for the first three resonances of the vocal cavity are 270, 2290 and 3010 Hz. Since the glottal waveform is periodic it's spectrum consists of a discrete set of harmonic frequencies. On passing through the vocal tract these frequency components are modified (i.e. multiplied) by the vocal tract response, providing the output spectrum. During the production of fricative sounds such as /s/, the excitation consists of a noise-like waveform produced as the egressive airflow is constricted between the tongue and the teeth (labio-dental) causing turbulent flow. Therefore, for fricative sounds, the output is noise-like and has no line spectrum.

In ongoing speech, the vocal tract changes shape relatively slowly due to physiological constraints on the articulators. Therefore, speech can be modelled as a linear system as in fig.3.3 but now the filter function is seen to vary relatively slowly over time with the resonance conditions of the vocal tract remaining stationary over a period of 30 to 40 ms. It is appropriate, therefore, to view the speech waveform using a short time spectral analysis. A short time window (~5ms) exhibits the resonance peaks of the system function along with the pitch period of excitation. A longer window (~25ms) reveals the harmonic frequency components in voiced speech as well as the spectral envelope. The trade off is that it cannot follow rapid changes as accurately. Figure 3.4 shows the case for both the narrowband and wideband analysis, both of which are commonly used in speech analysis. In practice we have seen very briefly how this analysis can be achieved via the sound spectrograph. Now, we take a look at an alternative approach (which we will show to be equivalent) using the Fourier transform.

d) Time Dependent Fourier Analysis[10]

The motivation for a short-time spectral representation which reflects the time varying properties of the speech waveform leads us to define the time dependent Fourier transform

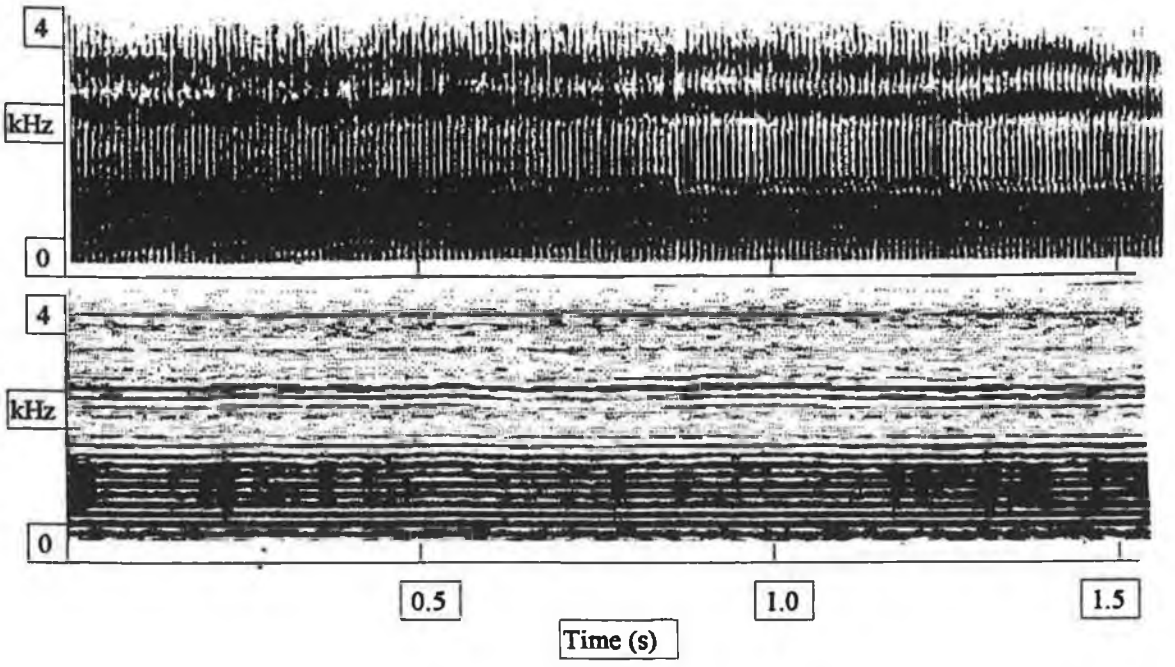


fig.3.4 *Narrowband and broadband spectrograms for the sustained phonation of the vowel a/.*

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{m=+\infty} w(n-m)x(m)e^{-j\omega m} \qquad \text{eqtn.3.1}$$

where w(n-m) is the window function that determines the portion of the input signal that receives emphasis at a particular time index, n.

The time dependent Fourier transform is a function of two variables: the time index n, which is discrete and the frequency variable, ω, which is continuous. The equation can therefore be interpreted in two distinct ways. Firstly, assuming n fixed leads to the normal Fourier transform of the sequence w(n-m)x(m), $-\infty < m < +\infty$. The second interpretation comes from considering $X_n(e^{j\omega})$ as a function of the time index n, with ω fixed. In this case the equation is clearly in the form of a convolution which leads naturally to considering the time dependent Fourier transform in terms of linear filtering. We shall consider the analysis in terms of the Fourier transform having noted it's equivalent linear filtering interpretation. From equation 3.1 we can see that the function takes on all integer values 'n' so as to slide the window along the sequence x(m). This is shown in figure 3.5 for three values of 'n'. in practice this sliding isn't

necessary as we have seen that the speech signal remains essentially stationary for time durations on the order of ten milliseconds. A computationally more efficient approach is to 'hop' the window along the sequence. This can be considered as a sampling of the equivalent linear filtering implementation.

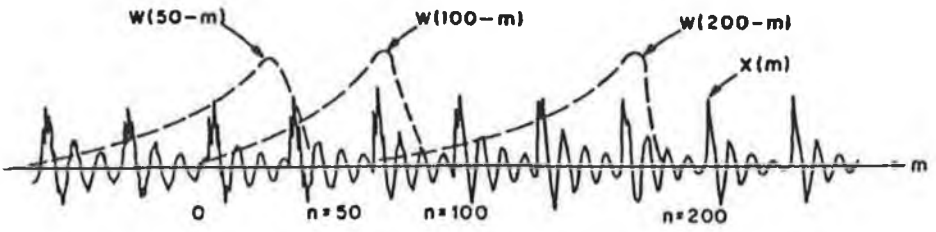


fig.3.5 *Sliding of window function along the speech sample. Three values of 'n' shown.*

The window has another important role besides simply selecting the segment for analysis. The rectangular window which Fourier transforms to

$$W(\omega) = \exp(+j\frac{\omega N}{2})\frac{\sin\left[\frac{N}{2}\omega\right]}{\sin\left[\frac{1}{2}\omega\right]} \qquad \text{eqtn.3.2}$$

where N is the total sample length.

It has the narrowest main lobe but has the highest side lobes of any of the commonly used windows. The result of windowing on the spectral estimates can be explained by viewing the time dependent Fourier transform (equation 3.1) in terms of the Fourier transform of a product. As illustrated in the following equation, for fixed 'n', this is equivalent to the convolution of the two individual transforms.

$$Xn(e^{j\omega}) = \frac{1}{2\Pi}\int_{-\Pi}^{\Pi} W(e^{j\theta})e^{j\theta n}X(e^{(\omega+\theta)})d\theta \qquad \text{eqtn.3.}$$

Hence a spectral estimate at frequency ω, gathers it's spectral energy contributions from within the mainlobe centred at ω, and also from the sidelode contributions. Therefore it is desirable to keep the sidelobes of the window function as low as possible in order to avoid 'leakage' in the spectral estimates. For this reason rectangular windows are rarely used in spectral analysis. Typical windows[11] generally have the property of coming smoothly to zero at their boundaries and therefore eliminating the discontinuities at the function edges. In speech analysis, commonly used windows include the Hamming and Hanning window functions. The Hamming window shown in fig.3.6 has the form

$$w(n) = 0.54-0.46\cos(2\Pi n/N) \qquad 0 < n < N \qquad \text{eqtn.3.4}$$
$$= 0 \qquad \text{otherwise.}$$

The sidelobes for this window are down by 40 dB, which is sufficient for most speech processing applications.

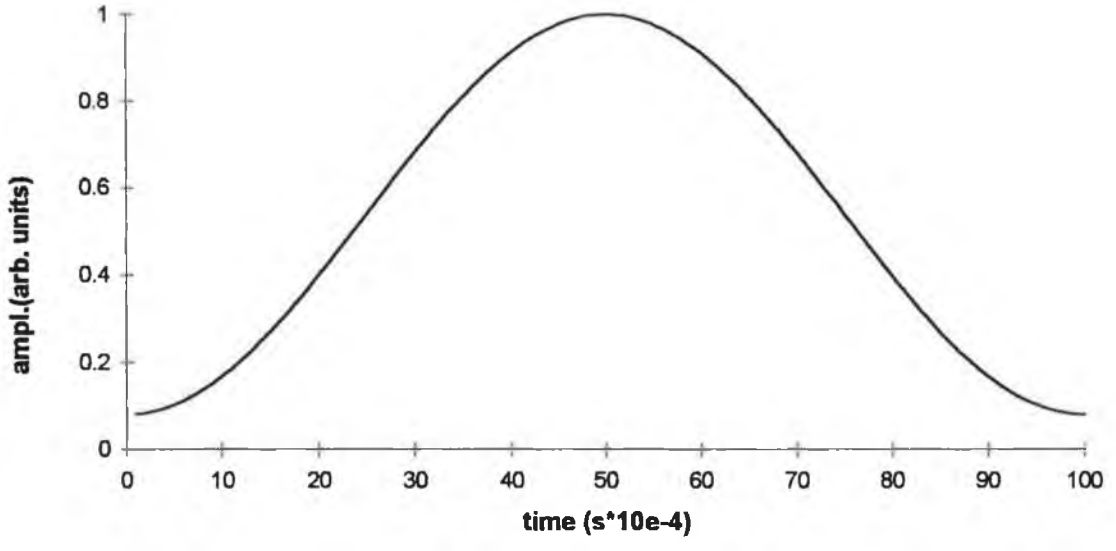


fig.3.6 *Hamming window function*

3.2.2 Implementation

The actual programming of the speech spectrogram was carried out in a Turbo C++ (Borland, Scotts Valley CA, USA) project file environment on a LEO 486DX PC. The complete time record is broken up into segments in order to calculate the short time Fourier transform providing a spectral cross section for each segment. Overlap segmentation is used in order to obtain more consistent spectral estimates. The number of points (N) required for each spectral cross section is greater in the narrowband case but since the time resolution is worse than in the wideband case the number of cross sections (M) to be computed is less. M was chosen to be a fixed percentage of N (seventy-five per cent overlap). Since the computation of the FFT is proportional to $N\log_2 N$, the analysis time for an utterance is essentially the same for both narrowband and wideband analyses. For a 6.5 second spectrogram the analysis time including display was approximately 30 seconds. On a Pentium PC this becomes real time. Hardcopy was attained using a Pizzaz++ screen dump routine (Application Techniques Inc., Pepperell, MA, USA) downloading to an inkjet printer (HP550C, Singapore). The input speech was low passed filtered at 3.8 kHz, digitised at 10 kHz and pre-emphasised using a first order difference equation in order to compensate for the -12 dB/octave falloff in amplitude of the source harmonics. The time window w(n), chosen to be a Hamming window (eqtn.3.4) was then applied. Many combinations of window length and overlap lengths were investigated in order to obtain the optimum display for both the narrowband and wideband spectrograms. For the narrowband analysis the window length was chosen to be 256 giving a 6 dB filter bandwidth of 70 Hz. In the wideband case the length was chosen to be 75, giving 6 dB filter bandwidths of 240 Hz. The FFT algorithm used was Numerical Recipes in C radix-4 algorithm, four.c, incorporated into the TC++ project file. The subsequent dB power spectrum amplitudes obtained were coded with a sixteen bit grey scale and the spectra were plotted (using TC++ graphics functions) with respect to time in order to produce the spectrogram display. Contour spectrograms were also produced using fine temporal hopping of 1 ms. It should be noted that the 6 dB bandwidths determine the frequency resolution of the FFT as point out by Harris[11] and not the classical

resolution criterion of 3 dB (or half power) as stated by Morris[9]. This is due to the fact that an FFT estimate is derived from the coherent addition of spectral components weighted through the window function. Also, the square root operation mentioned by Morris is not necessary as all that is required for the spectrogram display is the power, or rather the dB power spectrum.

All the literature pertaining to speaker identification based on spectrographic evidence refers to spectrograms that have been produced via the original analog device. However, in recent years the digital spectrogram which is commercially available in many forms has superseded the analog device. Although, it is clear that good quality spectrograms are available via the FFT approach, it is pertinent to compare the speaker identification ability based on visual comparison of these spectrograms, which have been produced using a completely different methodology and displayed via an entirely different hardware arrangement, with the original analog based spectrograms. An experiment (not previously reported in the literature) was set out in order to investigate this proposition but first we will take a brief glimpse at some of the extensive literature regarding the controversy surrounding the use of the original speech spectrogram for speaker identification.

3.3 Speaker Identification Based on Visual Inspection of Spectrograms

The dispute over the use of spectrograms for forensic speaker identification is well illustrated by the exchange between Koenig and Shipp et al and subsequently clarified by Nakasone et al, which took the form of letters to the editor of *The Journal of the Acoustical Society of America*. The initial letter by Koenig[12] (June 1986) highlights the main findings of the 1979 National Research Council[13] report on the reliability of spectrographic speaker identification under forensic conditions which were, in part:

(1) Estimates are available only for a few situations and they "do not constitute a generally adequate basis for a judicial or legislative body to use in making judgments

concerning the reliability and acceptability of aural-visual voice identification in forensic applications".

(2) Examiners should use all available knowledge and techniques that could improve the voice identification method.

(3) Spectrographic voice identification assumes that intra-speaker variability (differences in the same utterance repeated by the same speaker) is discernible from inter-speaker variability (differences in the same utterances by different speakers): however, that "assumption is not adequately supported by scientific theory and data." Viewpoints on actual error rates are presently based on "various professional judgments and fragmentary experimental results rather than from objective data representative of results in forensic applications".

In order to supply such data Koenig then outlines the work undertaken by the FBI in providing investigative support in over 2000 cases of voice identification over a period of fifteen years up until October, 1985. The error rates obtained for these investigations were 0.53 % false elimination and 0.31 % false identification. The Shipp et al reply[14] (April 1987) complains about examiner training, lack of detailed information on the methods used and results of investigation which were considered correct or incorrect based on whether a conviction was obtained or not (and presumably, in the light of other evidence relevant to the case). These objections were then countered by Koenig et al[15] who cites other literature for comparison practices. Further clarification on the above exchanges is provided by Melvin et al[16] in support of the spectrographic method, stating that the scientific community is divided on, rather than opposed to using the technique and they refer to Tosi's list of over seventy scientists in support of the method if practiced by trained examiners who follow the International Association for Identification norms.

Much of the conflict regarding the use of the spectrographic method can be attributed to the initial paper by Kersta[17], which appeared in *Nature* (December, 1962), in which he introduced the method, likening the technique to fingerprinting and coining the word 'voiceprint'. Subsequently, Kersta became convinced of the infallibility of the technique which he felt was robust with regard to aging, removal of adenoids etc.. We simply take note of the experiment type and error rates he reported by examinations

carried out by high school students with one week's training. For a group of nine talkers, using isolated cue words, the error rates ranged from 0 to 3 %. This provocative paper resulted in, among other comments, calls for more extensive research into the method and identification trials more relevant to forensic situations. A two year experiment on voice identification was undertaken by Tosi et al in order to test Kersta's claims and also to introduce models relevant to the forensic task. The reader is referred to *Experiment on Voice Identification*[19] for the details of the report and again we simply state the results. The experiment confirmed Kersta's experimental data for closed trials of identification and gave error rates of 2 per cent false identification and 5 per cent false elimination for the forensic tasks. From this bellicose background on speaker id via the spectrographic method we simply note the error rates and experimental conditions and techniques reported by Kersta and Tosi in order to compare the error rates obtained using digital spectrograms.

3.4 Experiment on Speaker Identification using Digital Spectrograms

Test Format:

The test format followed that of Kersta and later copied and extended by Tosi. This consisted of sorting and matching experiments in closed trials of identification on samples recorded contemporaneously. Each speaker was asked to utter the following words four times each - 'it', 'is', 'on', 'the', 'you', 'and', 'I', 'to', 'me', 'a'. also, he/she repeated the following sentences three times each -

'It is on the table'

'You and I have to go today'

'He gave me a card'

'He told me to put the kettle on.'

'And I told you a few minutes ago.'

Forty subjects (20 male, 20 female) who were free from any form of speech pathology supplied the above speech samples. All subjects were Irish Pre-Medical students in the age range of 17 to 33 years (average = 19.5, std. dev. = 3.5 years). The speakers were divided into four categories; 1) Male, Dublin accent, 2) Male, non-Dublin accent, 3) Female, Dublin accent and 4) Female, non-Dublin accent.

Digital spectrograms were produced for all the above utterences. Closed set, single utterance tests were constructed by randomly selecting six speakers from the same speaker category. Four spectrograms per isolated word existed for each speaker. Two spectrograms were selected as "known speaker" for each speaker for all ten words, which left two samples per speaker to be sorted for each word as shown in fig.3.7. The sentence procedure involved a matching experiment as opposed to the sorting experiment for the isolated utterances. In this case two references were taken for each speaker and the third sample, which had been chosen at random by the experimental coordinator, was required to be matched to one of these six (fig. 3.8).

All of the comparison tests were carried out by six examiners in the age range 22 - 31 years (average = 25.2 yrs, std. dev. = 4.5 yrs). All examiners were either academic staff or post-graduate research students from the RCSI Dept. of Chemistry and Physics. The examiners were required to make a definite decision based solely on the visual inspection of the spectrograms. Since none of the examiners had any prior experience with matching spectrograms, the category of Female-Dublin-Accent was set aside to act as a preliminary "training" set. After completing each experiment from this training set, the experimental coordinator would tell the examiner how many incorrect matches had been made, would identify which matches were incorrect and point out any spectrographic features from these that might have aided in a correct identification. Once the training set had been completed the experimental coordinator did not give any further retrospective assistance to the examiners but did tell them their number of incorrect matches.

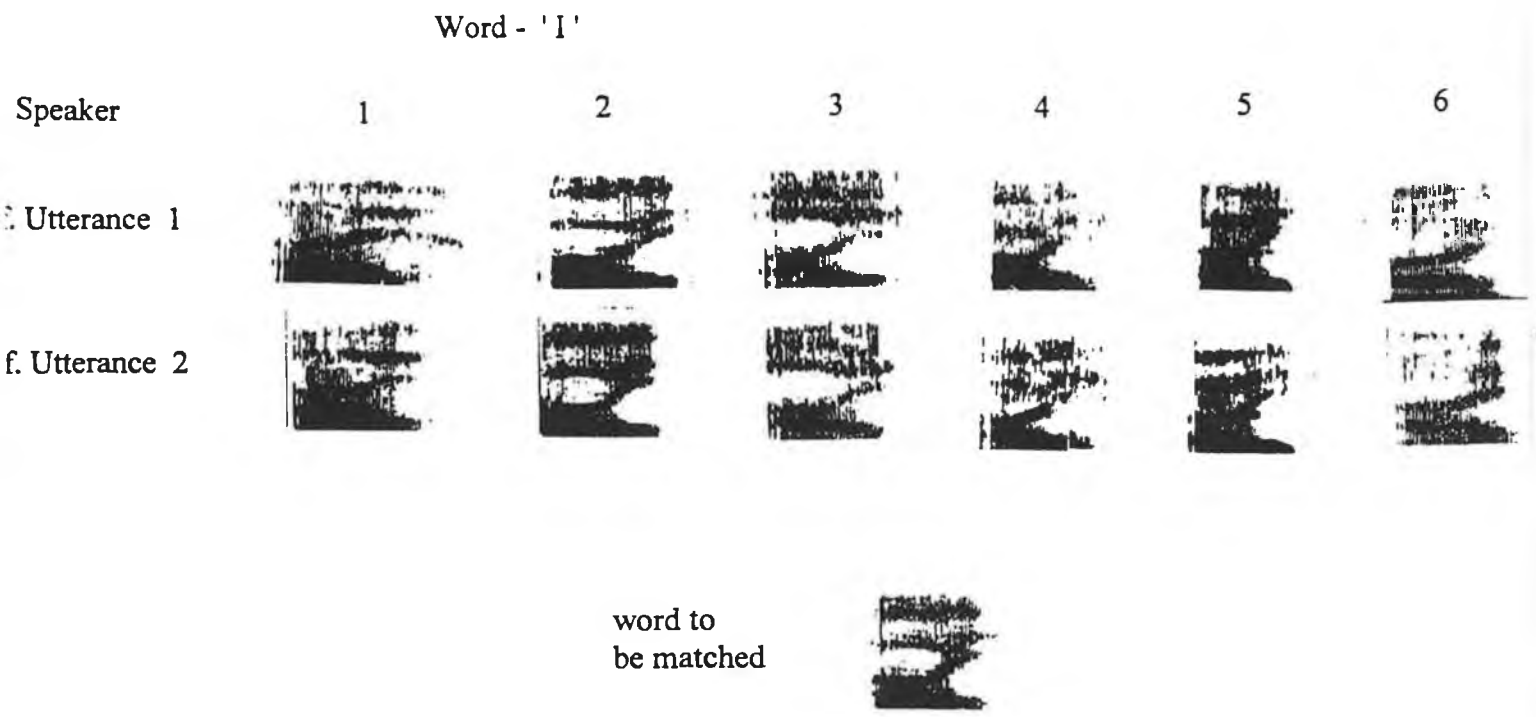


fig.3.7 *An example of the isolated word utterance sorting experiment for a six speaker test. There are a further eleven utterances to be matched.*

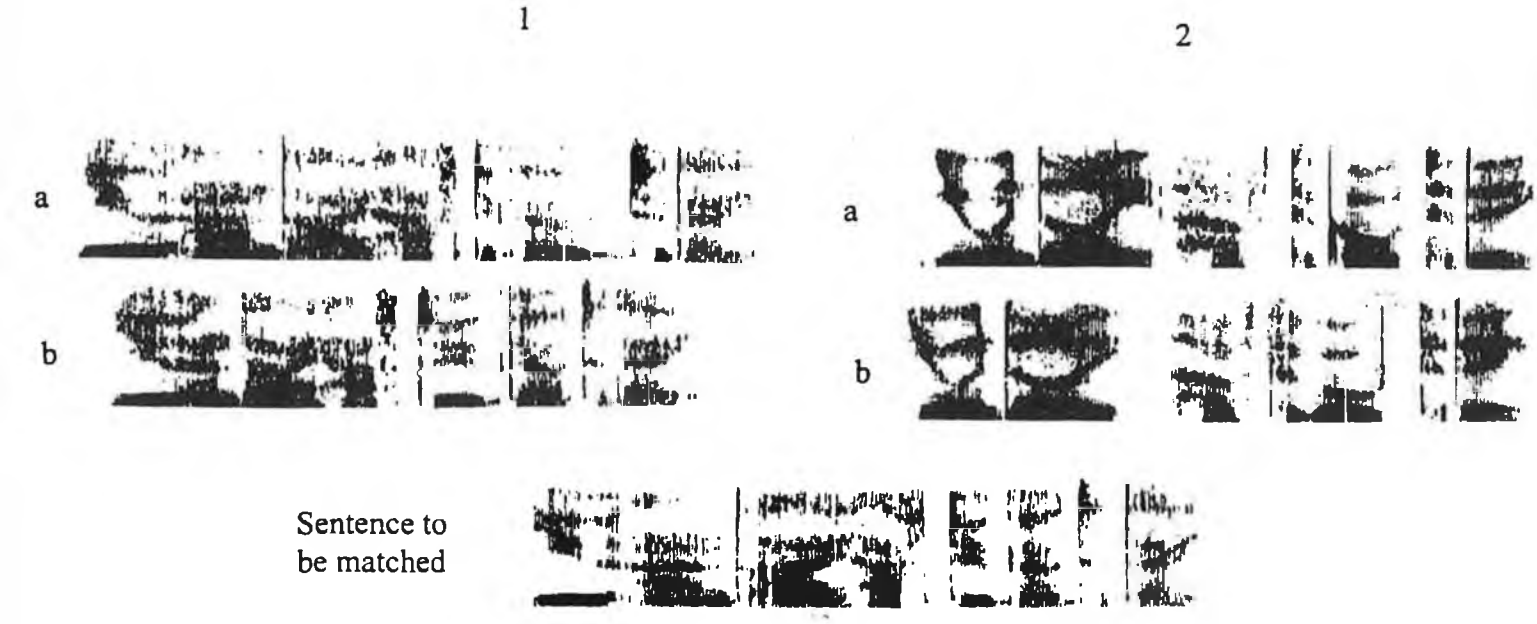


fig.3.8 *An example of the sentence matching. For purposes of illustration, only two utterances are shown (six were used in the actual test).*

3.5 Results:

Identification error rates for the single utterance sorting experiment for all six examiners are shown in Table 3.1. Each examiner carried out 40 single word utterance tests with 12 comparisons to be made in each. The first 10 tests were the training tests and these data are not included in Table 3.1. The mean identification error rate is 3.3 % (std. dev. = 2.9%) with individual values ranging from 0 to 6.7 %. The same data are shown in Table 3.2, however they are organised according to speaker category and word uttered and also includes the training set data. The single utterance data are displayed in descending order of total identification error rate for all speaker categories. The results from the 24 full sentence matching experiments are shown in Table 3.3.

EXAMINER NO.	NO. OF INCORRECT MATCHES (OUT OF 360)	PERCENTAGE INCORRECT (%)
1	0	0.0
2	14	3.9
3	12	3.3
4	22	6.1
5	24	6.7
6	0	0.0
Total	72 (out of 2160)	3.3

Table 3.1 *Single utterance identification error rates for each examiner*

Table 3.2 Mean identification % error rates for single utterance comparison tests calculated for all examiners and cathegorised according to speaker accent and word uttered.

UTTERANCE	SPEAKER TYPE I FEMALE, DUBLIN ACCENT (TRAINING SET)	SPEAKER TYPE II MALE DUBLIN ACCENT	SPEAKER TYPE III FEMALE, NON-DUBLIN ACCENT	SPEAKER TYPE IV MALE, NON-DUBLIN ACCENT	FAILURE RATE FOR ALL ACCENT TYPES (%)
YOU	19.4	5.6	13.9	5.6	8.3
THE	11.1	11.1	2.8	5.6	5..5
TO	2.8	2.8	13.9	0.0	5.5
ME	2.8	0.0	5.6	9.7	5.1
A	2.8	2.8	0.0	6.9	3.2
AND	2.8	0.0	0.0	5.6	11.8
IT	6.9	0.0	0.0	2.8	0.9
IS	15.3	0.0	0.0	2.8	0.9
I	6.9	0.0	2.8	0.0	0.9
ON	2.8	0.0	0.0	0.0	0.0
Mean (%)	7.4 (std. dev. = 5.7)	2.2 (std. dev. = 3.5)	3.9 (std. dev. = 5.3)	3.9 (std. dev. = 3.1)	3.3*

* Excluding training set data

EXAMINER NO.	NO. OF INCORRECT MATCHES (OUT OF 24)	PERCENTAGE INCORRECT (%)
1	0	0.0
2	2	8.3
3	0	0.0
4	2	8.3
5	0	0.0
6	0	0.0
Total	4 out of 144	2.8

Table 3.3 *Full-sentence line up test identification percentage error rate for each examiner.*

3.6 Discussion

The error rates obtained using digital spectrograms were in agreement with the error rates obtained by Kersta and Tosi in experiments that used the original spectrographic device. As expected, sentence matching, where multiple cues are available to the examiner, reduces the error rate (four examiners made zero identification errors using the sentence matching). Better performance may have been expected for the analog device since the digital version consists of a sampling of the analog filtering. Alternatively, better results may have been expected with the digital spectrogram since the resolution of the greyscale is much higher for the PC graphics card than the teledeltos paper. Other advantages of the digital version include near real time production, greater speech sample lengths, expanded time scales (zoom feature) and the ability to overlay spectrograms for comparison purposes. Furthermore, noise

artifacts are more conveniently removed from the digital signal. Other computer based speaker identification techniques exist and therefore having a computer based spectrogram is of further convenience.

3.7 Conclusion

The results reported in tables 3.1 and 3.2 are in good agreement with those reported by Tosi et al in their validation of Kersta's original experiment. Therefore, it can be assumed that there is no degradation in processing the spectrograms in a PC based environment via the FFT. This is an important result (independent of any controversy surrounding the method) if we consider that most, if not all persons currently using the spectrogram would now be using the digital version. Furthermore, no previous experiments on spectrographic speaker identification based on digital spectrograms have been reported in the literature.

Based on these preliminary speaker identification results we do not advocate for or against the use of the spectrogram for speaker identification but simply state that it seems a very useful investigative tool that should supplement a battery of assessment procedures. It is an excellent tool for transcription purposes and can also be useful for examining non-speech material of forensic interest such as gunshot sounds. In conclusion, digital spectrograms offer no deterioration (or improvement) in identifying power over their analog counterparts and hence provides a very important tool for use in investigations into speaker identification.

3.8 Bibliography

1. Doddington, GR. Speaker recognition-identifying people by their voices. Proc. IEEE 1985; **73**:1651-1664
2. Reich, AR, Moll, KL and Curtis, JF Effects of selected vocal disguises upon spectrographic speaker identification. J. Acoust. Soc. Am. 1976; **60**:919-925
3. Endres, W. Voice spectrograms as a function of age, voice disguise and voice imitation. J. Acoust. Soc. Am. 1971; **49**:1842-1848
4. Federico, A. A new automated method for reliable speaker identification over telephone channels. 1987; IEEE 1457-1460
5. Koenig, W., Dunn, HK and Lacey, LY The sound spectrograph, J. Acoust. Soc. Am. 1946; **18**:19-49
6. Potter, RK, Kopp, GA and Green, HG Visible speech. Princeton, NJ: Van Nostrand, 1947
7. Oppenheim, AV. Speech spectrograms using the fast Fourier transform. IEEE Spectrum, August 1970
8. Mermelstein, P. Computer generated spectrogram displays for on-line speech research. IEEE Trans. on Audio and Electroacoustics 1970; **19**:44-47
9. Morris, LR. Software engineering for an IBM-PC/Ti-speech realtime digital spectrogram production system. Speech Tech. 1986; 82-86
10. Rabiner, L. and Schafer, R. Digital processing of speech signals. Englewood Cliffs, N.J.: Prentice Hall, 1978
11. Harris, FJ. On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. of the IEEE, 1978 **66**:51:142
12. Koenig, BE. Spectrographic voice identification: A forensic survey. J. Acoust. Soc. Am. 1986, **79**:2088-2090
13. Bolt, RH, Cooper, FS, David, EE, Denes, PB, Pickett, JM and Stevens, KN On the theory and practice of voice identification. National Academy of Sciences, Washington DC 1979
14. Shipp, T., Doherty, ET and Hollien, H. Some fundamental considerations regarding voice identification. J. Acoust. Soc. Am. 1987; **82**:687-688

15. Koenig, BE, Ritenour, DV, Kohus, BA and Kelly, AS Reply to "Some fundamental considerations regarding voice identification". J. Acoust. Soc. Am. 1987; **82**:688-689

16. Melvin, C., Nakasone, H. and Tosi, O. More fundamental considerations regarding voice identification. J. Acoust. Soc. Am. 1988; **84**:1943-1944

17. Kersta, LG. Voiceprint identification. Nature 1962; **196**:1253-1257

18. Tosi, O., Oyer, H., Lashbrook, W., Pedrey, C., Nicol, J. and Nash, E. Experiment on voice identification. J. Acoust. Soc. Am. 1972; **51**:2030-2043

Chapter 4

Time Domain Analysis

4.1 Introduction

Time domain analysis forms a very important branch of speech processing, not just for the analysis results in themselves but also as a first stage for further processing of the speech signal. By *time domain* methods we mean simply that the processing involves the waveform of the speech signal directly in contrast to the techniques described in chapters 5 and 6 which we classify as frequency domain methods since they involve some form of spectral representation. Many important features of the speech signal can be simply specified through implementation of these techniques. Zero-crossing rate, short time energy, autocorrelation, voiced/unvoiced classification and pitch can all be conveniently extracted using straight forward time domain processing techniques. It is not our intention to give an exhaustive survey of these techniques but rather to show that time domain methods provide many useful, and indeed essential strategies for processing and pre-processing the speech waveform.

In the vocal pathology literature time domain methods, primarily involving the extraction of pitch and pitch perturbation measures have received considerable attention[1]. Practically all early attempts at providing quantifiable measures from the acoustic analyses of pathological voices were based on the estimation of pitch perturbation. Perturbation, as used in the speech pathology literature, can be considered as a generic term used to describe some form of variation in the speech waveform from period to period. This is usually simply the variation in the fundamental frequency from period to period (jitter) or the variation in the amplitude of the peak in the output radiated speech waveform from period to period (shimmer), although many other variations exist. These two measures of jitter and shimmer have received extensive attention in the literature and will be examined in detail in section 4.3. The reason for the comprehensive evaluation of these measures is that laryngeal pathology generally alters the normal vibratory pattern of the vocal cords and therefore analysis of the subsequent pitch variation in the output radiated speech waveform should reveal these source anomalies. For the same reason the autocorrelation function is taken of the voiced speech signal in order to measure the relative similarity between adjacent cycles of the waveform. This issue is addressed in section 4.4.

4.2 Pitch Extraction

A display of a time domain waveform which is called a sonogram is shown in fig.4.1(a),(b) for a normal and pathologic speaker for the sustained vowel phonation a/, taken at comfortable pitch and loudness level for each speaker. The difference between normal and pathological is readily evident from the display, therefore what is required is some means of quantifying this difference or perturbation. In order to provide consistent and reliable measurements of this difference, a reliable method of pitch extraction is pre-requisite. Many methods exist in both the frequency and time (and quefrency) domain for extracting the pitch period from the speech signal[2]. These methods can be divided on the grounds of whether they are short term average methods or single cycle detectors. In the former class of methods a window is applied in order to limit the signal to a few cycles and the pitch estimate can be updated by

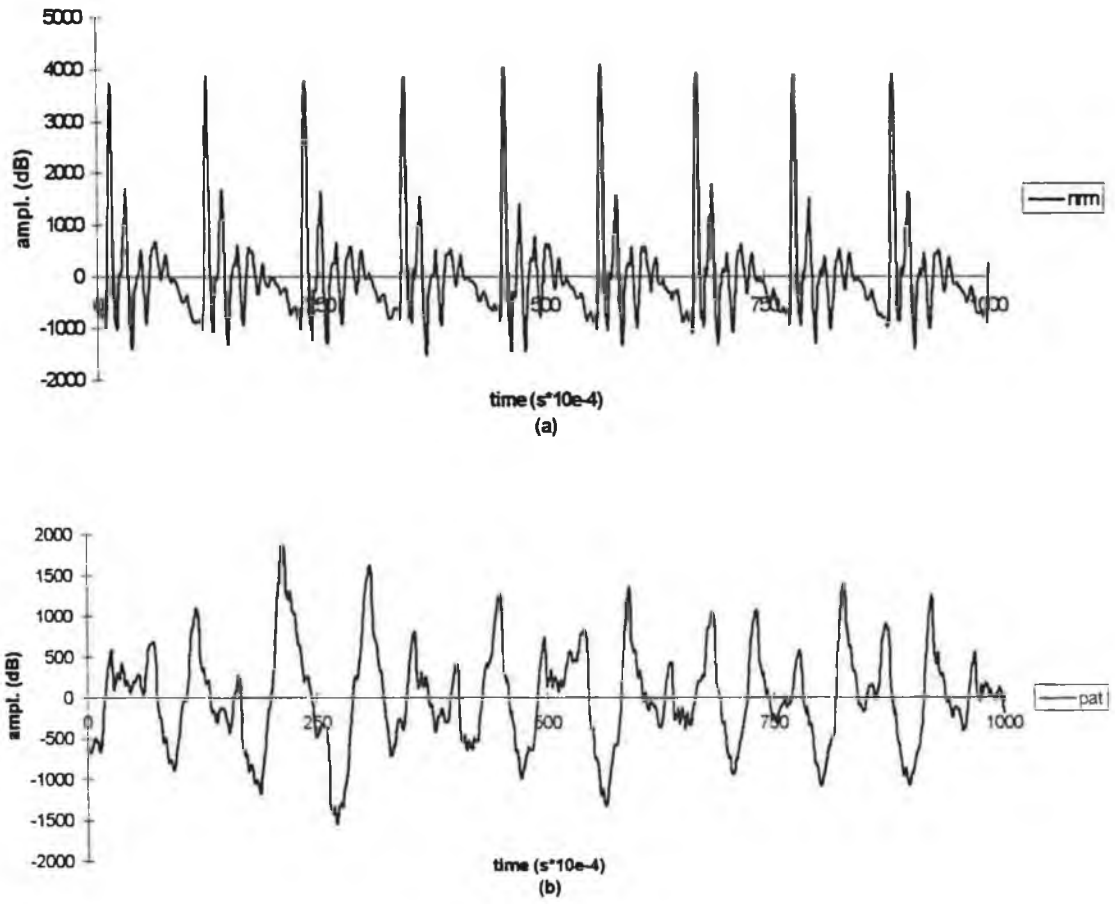


fig.4.1 *Sonograms of the phonation of the sustained vowel a/ for* (a) *a normal (nrm) and* (b) *patient (pat) of the present study*

successively hopping the window over the required range. Examples of such methods are the autocorrelation function[3] (with many processing variations), the log harmonic spectrum[4] and the cepstrum[5]. Although these methods, particularly the latter, have proven to be robust pitch detectors, even in the presence of competing noise, they are not applicable for perturbation analysis as we require a pitch synchronous estimation of the period. These single cycle detectors, as we have called them, are all time domain techniques that focus on a single event (or many events) within a cycle of voiced speech. The most prominent feature of the radiated speech waveform is often the peak amplitude that occurs every period as shown in figure 4.1(a). These positive peaks therefore provide a convenient means for estimating the pitch period. Other strategies

involve taking the negative peaks from the waveform. Low pass filtering the waveform at a frequency between the fundamental frequency (f0) and twice the fundamental and calculating the zero-crossing rate is another popular method. An output radiated speech waveform and it's low pass filtered version are shown in figure 4.2. Another strategy involves matching the waveform from cycle to cycle, usually implemented through some form of least squares estimation procedure[6]. The issue as to which method gives the most reliable results has been addressed in detail by several researchers along with other methological concerns[7,8,9].

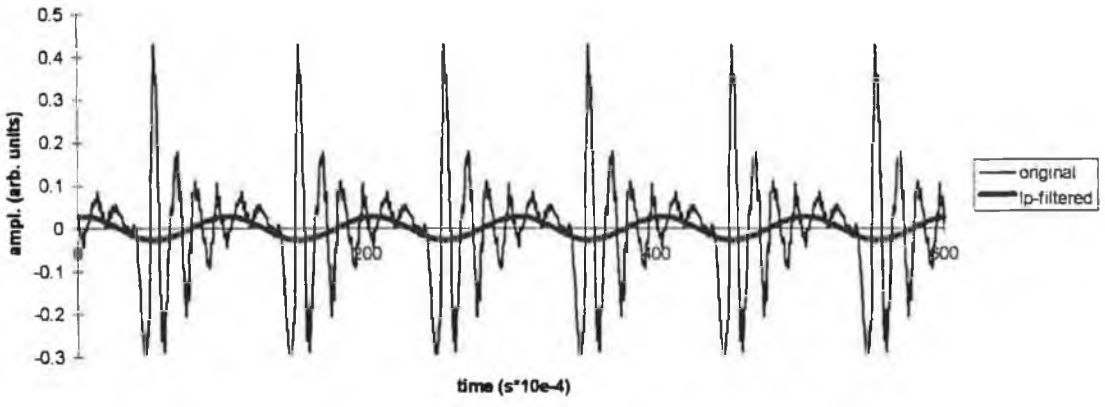


fig.4.2 *110 Hz synthesis file (original) and a filtered version which has been low passed at 1.5 times an initial f0 estimate*

Although Titze et al[8] have shown waveform matching to be a robust method of pitch extraction even in the presence of additive noise and low level frequency and amplitude modulations, the method becomes sensitive to frequency modulations exceeding six percent and hence, in anticipation of fluctuations of this magnitude in respect of pathological voices the method was not implemented. However, in recognition of the robustness of the method i.e. many estimates are obtained per cycle, a low pass version of waveform matching was developed. Three further methods using the positive peaks from the original and low passed waveforms and the positive zero crossings from the low passed waveform were also implemented. The analysis details for the three methods based on the low passed waveform are shown in A, B and C.

A. Waveform Matching of the Low Passed Filtered Waveform

In this method an initial estimate of the fundamental frequency (f0) is obtained using any convenient short term f0 extraction method, the cepstrum being chosen here. Following this initial f0 estimate, the waveform is low passed filtered at 1.5×f0 using a 250th order, low pass, finite impulse response (FIR) filter (Matlab's signal processing toolbox). As a consequence of this filtering the low passed waveform will only cross the x-axis twice per cycle, giving one positive (PZC) and one negative zero crossing (NZC) per cycle as shown in fig. 4.3. The negative zero crossings (NZC) are then used as rough period markers with respect to which the search for cyclic events begins. Any prominent event within the pitch markers is taken as the starting point for the waveform matching. The negative peak location 'NP(1)' was chosen in this implementation (fig.4.3). A point 'NP(2)' between the second and third rough markers is then found such that the mean squared error between the adjacent wavefroms is minimal. The pitch period is hence calculated to be 'NP(1)-NP(2)'.

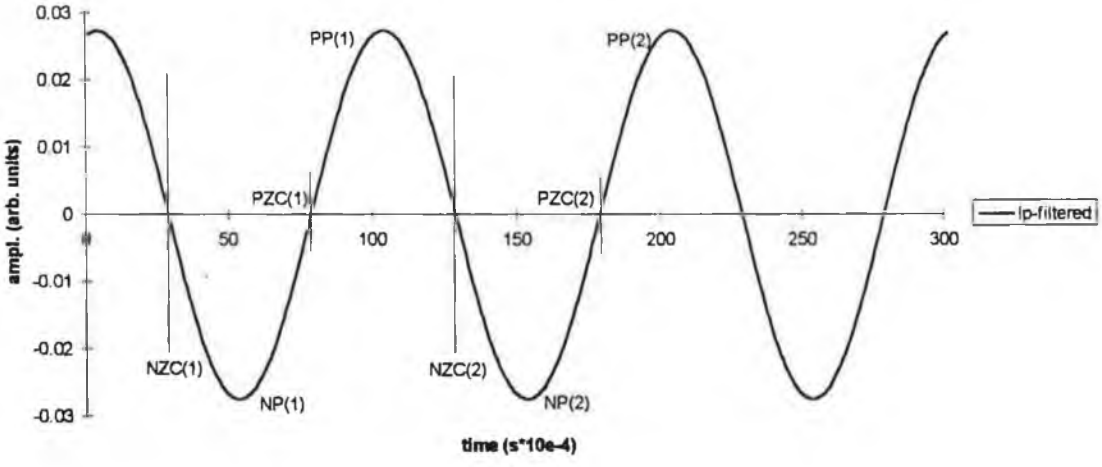


fig. 4.3 *Low passed (1.5×f0) filtered waveform with negative peak (NP), positive peak (PP), positive zero crossing (PZC) and negative zero crossing (NZC) pitch markers.*

The exact method can be followed from the five steps below with reference to figure 4.3, which is essentially the same approach that Titze et al used on the original waveform.

1. The rough markers are set using the negative zero crossings of the low passed filtered waveform (NZC(1) and NZC(2)).

2. The negative peak (NP(1)) between NZC(1) and NZC(2) is located.

3. An initial guess is made of the negative peak location between the second and third rough markers NP(2) = NZC(2)+(NP(1)-I(1)) (see fig.4.3).

4. A search limit of a user given percentage, PERC (typically 15-30%) is set such that

J1 = NP(1) - PERC*(NP(2)-NP(1))

J2 = NP(1) + PERC*(NP(2)-NP(1))

and a point Jm between J1 and J2 is found so that ERR(Jm) is minimal, where

$$\mathrm{ERR(Jm)} = \frac{1}{j - \mathrm{NP}(j-1)} \times \sum_{k=P(i-1)}^{j-1} (\mathrm{buff}(k + [j - \mathrm{NP}(i-1)]) - \mathrm{buff}(k))^2 \qquad \text{eqtn.4.1}$$

where 'buff' is the data buffer.

5. The resulting estimate of pitch period is limited by the sampling frequency and hence interpolation is used to improve the estimate. A second order polynomial is fitted to the points ERR(Jm-1), ERR(Jm) and ERR(Jm+1) to find the minimal location J'.

6. The period is hence given as NP(2)= NP(1)-J'

7. The process is similarly repeated for all periods of the waveform.

B. Positive Peaks from the Original and Low Passed Waveforms

Again, rough period boundary markers are set as for the waveform matching method. Then the positive peak locations (PP(i)) are found within these markers using a simple peak picking algorithm. Interpolation is then used to obtain a more accurate estimate of the peak location using a second order polynomial as follows

$$\mathrm{PPr(i)} = \mathrm{PP_i} + \frac{-0.5 \times (\mathrm{buff}(\mathrm{PP_i}+1) - \mathrm{buff}(\mathrm{PP_i}-1))}{\mathrm{buff}(\mathrm{PP_i}+1) - 2 \times \mathrm{buff}(\mathrm{PP_i}) + \mathrm{buff}(\mathrm{PP_i}-1)} \qquad \text{eqtn.4.2}$$

where 'buff' specifies the signal array.

C. The Zero Crossing Method

Zero crossing can only be used on a low pass filtered version of the output radiated speech waveform unless of course a contact microphone or electroglottography (EGG) was used to record the signal in which case the original waveform can be analysed directly. The rough period markers are set as before using the negative zero crossing locations. The positive zero crossings (PZC) are then located between these markers (fig.4.3) and a first order polynomial (i.e. straight line interpolation - equation 4.3) is fitted in order to improve the estimate.

$$PZC_r(i) = PZC_i + \frac{-buff(PZC_i)}{buff(PZC_i + 1) - buff(PZC_i)} \qquad \text{eqtn.4.3}$$

where again 'buff' indicates the signal.
The i^{th} fundamental frequency is hence calculated as follows

$$f0(i) = \frac{fsam}{PZC(i+1) - PZC(i)} \qquad \text{eqtn.4.4}$$

where fsam is the sampling frequency.

4.2.1 Test Stimuli

In order to test the accuracy of the pitch extraction and subsequent perturbation analysis programs it is important to have precise knowledge of the accuracy obtainable with the synthesis data. Therefore, some further comments on the jittered synthesis data which were introduced in chapter 2 are given. In the synthesis we have simulated two very different conditions of jitter, random and cyclic. The former case, which is

perhaps the more common simulation, is produced using a random number generator as shown

$$P1 = \frac{fsam}{f0},$$

$$P = \frac{P1 \times (100 + per \times randn)}{100} \qquad \text{eqtn.4.5}$$

P1 = pitch period , fsam = sampling frequency, f0 = fundamental frequency
randn = normally distributed random numbers, mean zero, variance one
per = percentage perturbation

The user inputs f0, from which the pitch period is simply obtained as shown in equation 4.4. Pitch perturbations of variance 'per' are then introduced via Matlab's random number generator 'randn.m' which produces normally distributed random numbers with mean zero and variance one. For example, setting per = 4, gives a jitter set with a standard deviation of 4%. This pitch period variation is plotted with respect to time in fig. 4.4. However, our data are also constrained by the fact that they must be integer valued. The function round.m is used to round the data to the nearest integer value. This must be imposed because alternatively the data takes the 'floor' integer value by default.

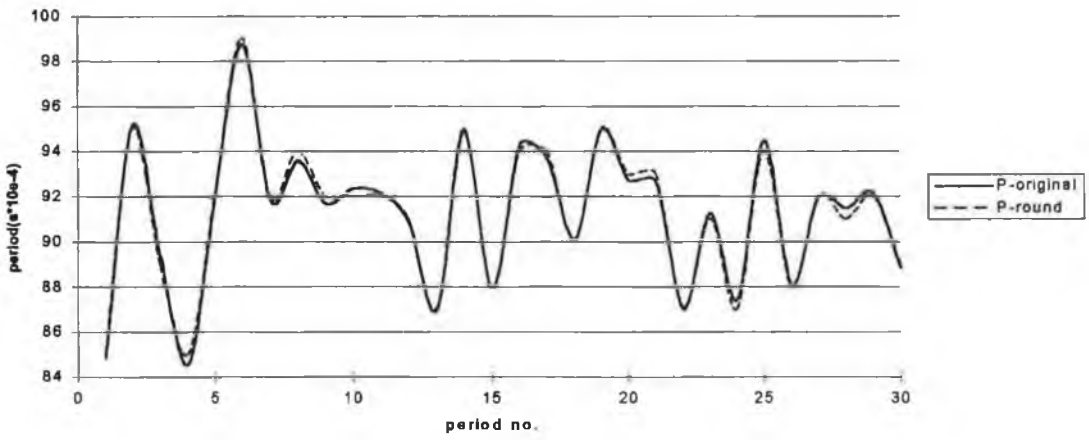


fig.4.4 *Variation of pitch period for the 4% std. dev. random jitter signal with and without rounding to the nearest integer.*

The glottal period was varied so as to keep the relative harmonic levels equal i.e. all components were scaled in such a way so as to match a given period and this put further restrictions on the data, requiring integer values for the rise time (N1) and closing time (N2) of the glottal pulse. The synthesis data were produced in such a way so as to represent a 10 kHz sampling frequency and the jitter set signals were investigated using the 110 Hz files. The equivalent pitch period is ~9.1 ms or 91 sample points when rounded. It can be seen therefore (fig.4.4) that when low level pitch perturbations are introduced through the random number generator, they are only very crudely approximated when rounded to the nearest integer.

The cyclic jitter values were produced by simply alternating successive periods between P1 and P2 where P1 and P2 are fixed. The spectral consequence of this is to produce an harmonic peak at an octave lower in the frequency spectrum (fig. 4.5).

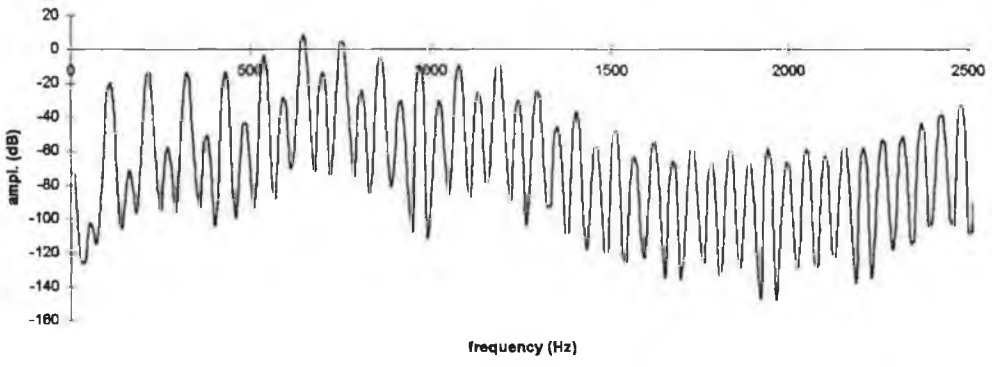


fig.4.5 *Sub-harmonic regime introduced as a result of the cyclic jitter perturbation*

This may seem like a very unnatural way in which to introduce jitter but in consideration of how one glottal cycle can influence the next it may in fact be a more realistic manner in which to simulate jitter in certain cases of vocal pathology e.g. fry phonation. Some work has been done in respect to what the distribution of jitter actually follows (Pinto et al[10]) in cases of normal and pathological speakers. The data set for normals produce a somewhat Gaussian distribution, although some degree of skewness is apparent. The distribution involving cases with vocal pathology depends on the type of the pathology with diplophonia, for example giving a bimodal distribution.

4.2.2 Results of the Various Extraction Procedures

Having detailed the jitter signals in the previous section we are now in a position to appreciate the results that are returned from the f0 extraction procedures. The next section gives a detailed account of the various pitch perturbation measurements that have been developed in order to assess the periodicity of the speech signal. We shall use one of these, the perturbation index defined as

$$PF1 = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{f0(i+1)-f0(i)}{0.5\times[(f0(i+1)+f0(i))]}\times 100 \qquad \text{eqtn.4.6}$$

(where N is the total number of periods), in order to compare the different extraction procedures and, after choosing the most appropriate extraction method, we compare the usefulness of the different perturbation measurements. Equation 4.6 is called the 'Pitch Perturbation Factor One' (PF1). This value (for the output radiated speech waveform) is plotted in per cent versus random pitch perturbation and cyclic pitch perturbation of the glottal source in fig.4.6 and fig.4.7.

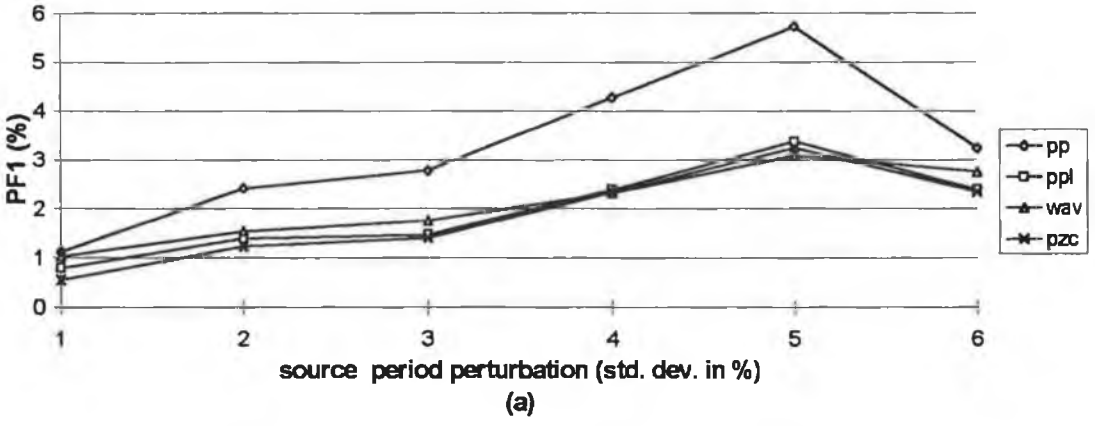


fig.4.6 *PF1(%) values for random jitter using the four extraction methods where PP-positive peaks, PPl-positive peaks of low passed filtered waveform, wav-waveform matching and PZC-positive zero crossing*

It can be seen from fig.4.6 that all methods show an increased PF1 with source random jitter as expected. The methods based on the low passed filtered waveform show somewhat lower values than the positive peak picking method that was applied to the original waveform. The slight reduction in PF1 when the jitter goes from 5% to 6% std. dev. of glottal source is a consequence of rounding to the nearest integer as explained above. The cyclic jitter values only show two levels of perturbation when examined using PF1. Except in the case of waveform matching the general trend is that as the cyclic jitter increases the PF1 factor increases (fig.4.7) .

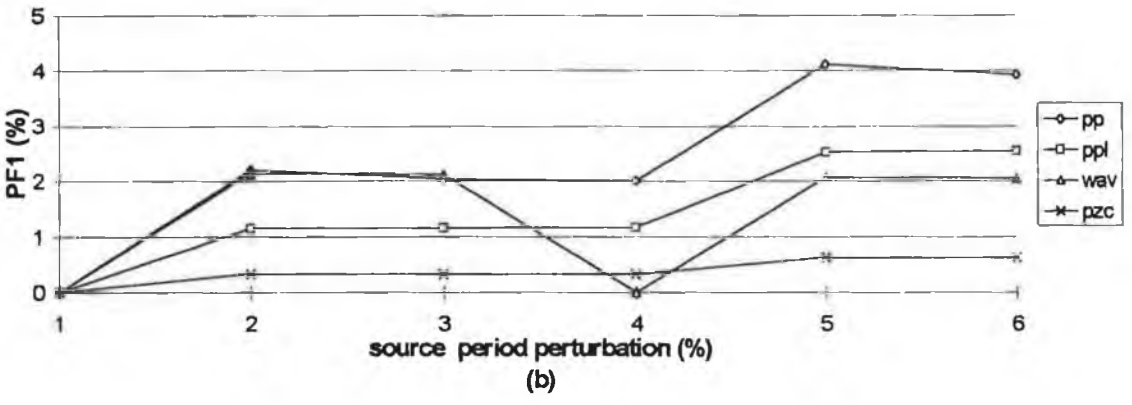


fig.4.7 *PF1(%) values for cyclic jitter using the four extraction methods where PP-positive peaks, PPl-positive peaks of low passed filtered waveform, wav-waveform matching and PZC-positive zero crossings*

The performance of the extraction methods in the presence of additive noise is shown in figure 4.8. As expected the positive peaks (PP) taken from the original waveform perform give high jitter values in the presence of additive noise and their values are not shown in fig.4.8 as they are a scale factor higher than the values returned by the other three methods. Figure 4.9 illustrates the problem encountered here. The waveform matching method is very robust against additive noise until Gaussian noise of standard deviation 8 is exceeded. At std. dev. 32 the method performs as badly as the PP of the low passed waveform. The PZC method gives the lowest jitter values for the high noise levels, giving a PF1 value of only 0.5% for noise of standard deviation 32 of the

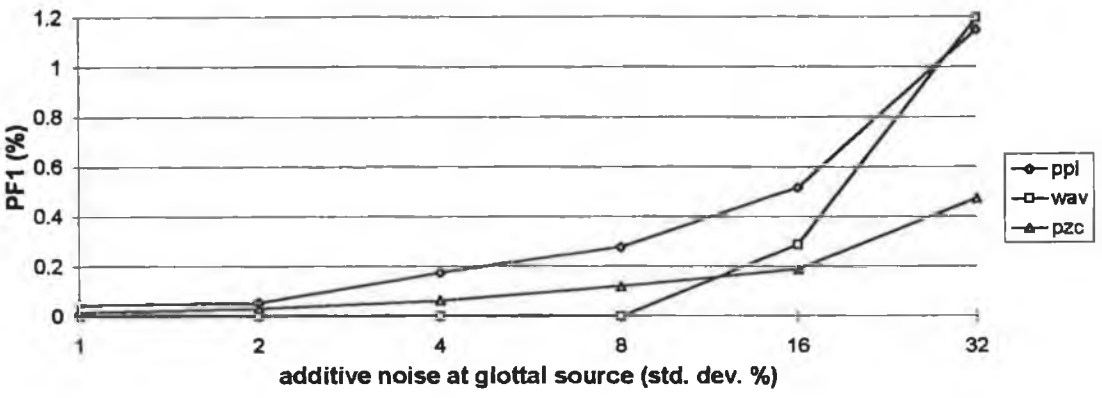


fig.4.8 *PF1(%) values for additive noise using the four extraction methods where PP-positive peaks (off scale), PPl-positive peaks of low passed filtered waveform, wav-waveform matching and PZC-positive zero crossings*

fig.4.9 *Speech waveform with additive noise (std. dev. 16%) illustrating why positive peak picking gives high jitter scores*

glottal source. However the jitter values obtained by this method are not as low as the waveform matching values for lower levels of additive noise.

Figure 4.10 shows the effect of shimmer on the jitter values. All methods based on low pass filtering the waveform perform extremely poorly for the shimmer signals with jitter values of over 5 % for glottal source amplitude perturbation signals with a variance of 32. The PZC method performs somewhat better but still gives a PF1 value

of 2 % for a glottal source amplitude perturbation of std. dev. 32%. In contrast the PP of the original waveform are relatively unaffected even by large levels of shimmer. Figure 4.11 illustrates why the low passed versions perform so badly under conditions of high shimmer. When the amplitude changes in the original signal, the low passed version skews somewhat, offsetting the zero crossing markers.

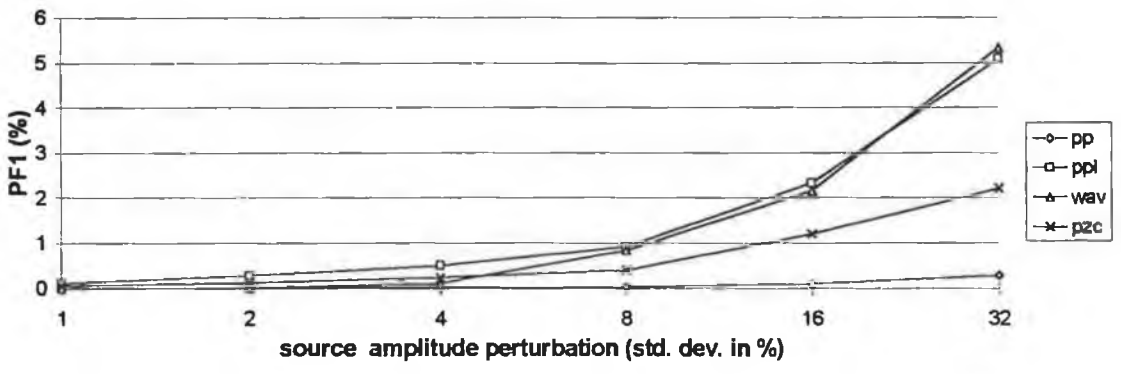


fig.4.10 *PF1(%) values for shimmer signal set using the four extraction methods where PP-positive peaks, PPl-positive peaks of low passed filtered waveform, wav-waveform matching and PZC-positive zero crossings*

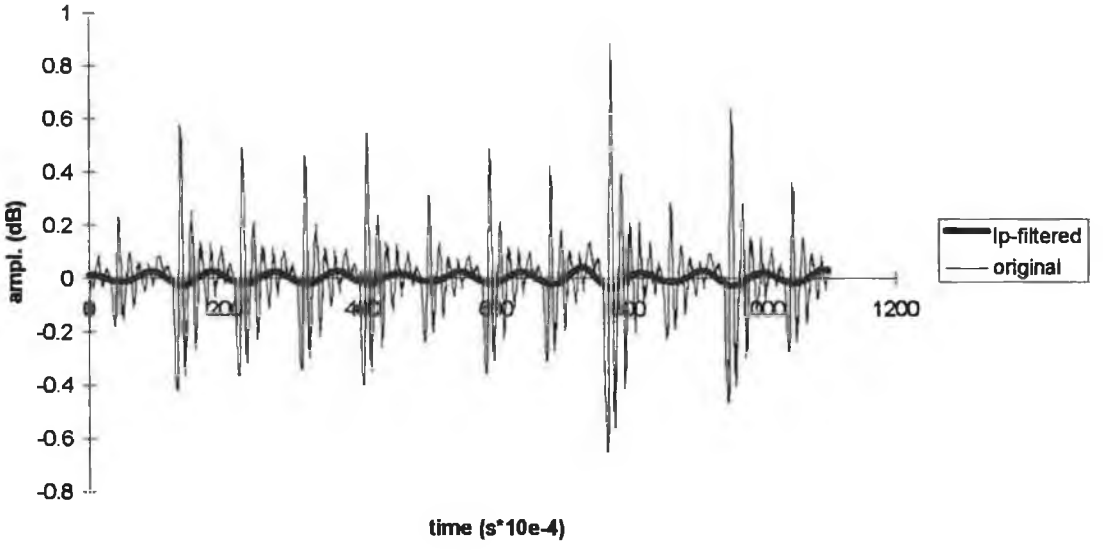


fig.4.11 *Low passed filtered and original waveforms for a shimmer signal with std. dev. 16 % random shimmer introduced at the glottal source*

Also of interest are the jitter values obtained for the radiated output as compared to the jitter values obtained for the glottal source signals. Figure 4.12 shows this relation between source and output jitter values for the extraction method that gave the lowest level of jitter for that perturbation. For the shimmer signal set, the increased levels of jitter (where the fundamental frequency was obtained using any of the low pass filtered extraction procedures) for the output radiated waveforms are simply due to the overall multiplicative effect of the vocal tract filter function.

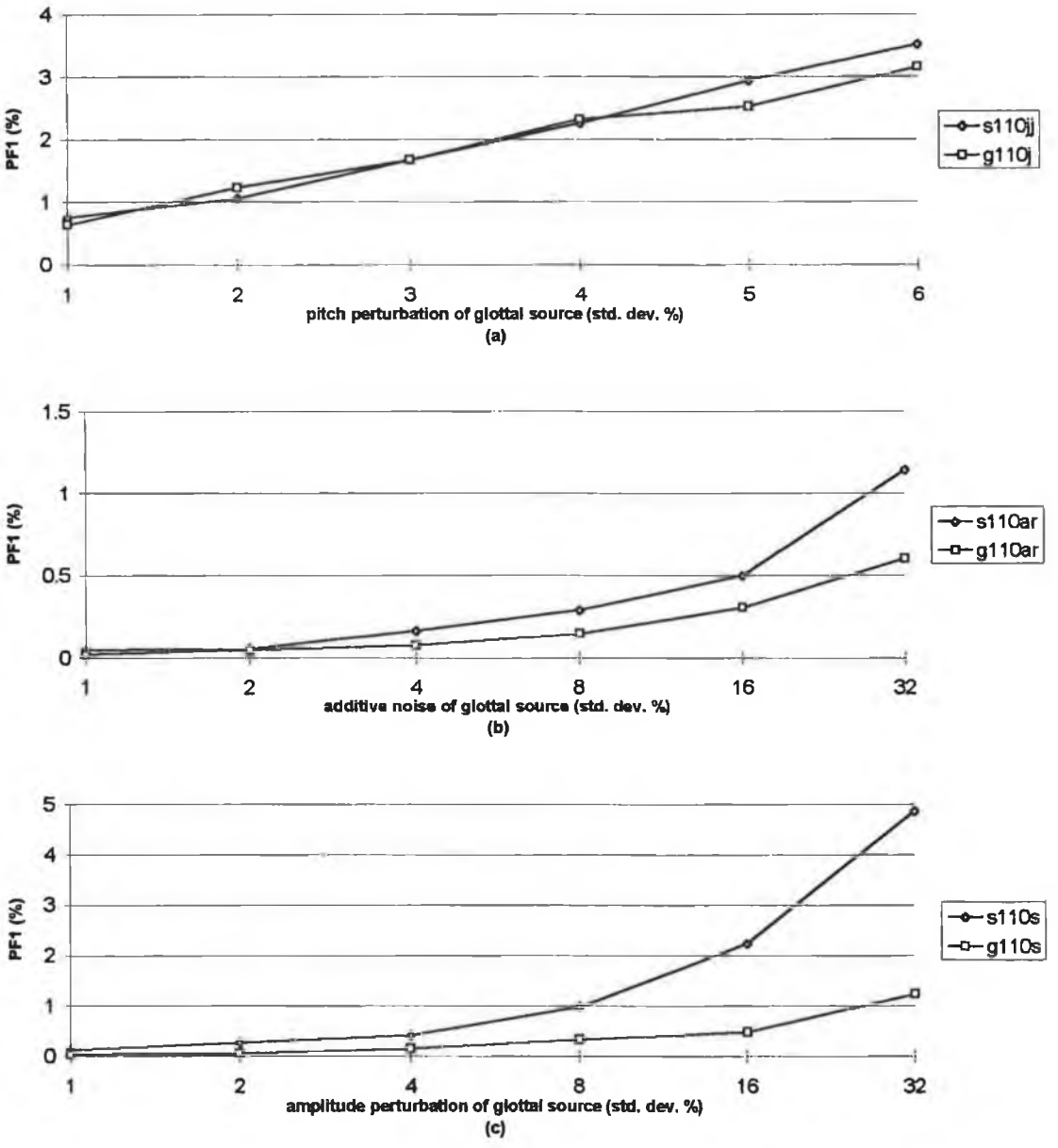


fig.4.12 *PF1 values for source and radiated waveforms for (a) random jitter (b) additive noise and (c) amplitude perturbation (g-glottal s-output radiated speech)*

4.3 Measurement of Pitch Perturbation

Several measures exist that give an indication of the level of perturbation present in the speech waveform. Lieberman's[11] pitch perturbation coefficient, which specifies the number of period perturbations that exceed 0.5ms for a given number of cycles of the waveform was the first measure introduced. It was significantly improved through the introduction of the relative average pitch perturbation measure by Koike[12]. This measure is defined as

$$RAPP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|\frac{f0(i-1)+f0(i)+f0(i+1)}{3}\right| - f0(i)}{\frac{1}{N}\sum_{i=1}^{N} f0(i)} \qquad \text{eqtn.4.7}$$

where $f0(i) = i^{th}$ fundamental and N = number of periods

the three point moving average was introduced to exclude the effects of the slow and smooth changes that occur in the melodic contour (i.e. the graph of fundamental frequency plotted with respect to time). These slow changes (tremor) are due a combination of neurological and physical mechanisms. Experimental evidence for neurological causes of tremor in the 1-5 Hz range has emerged, while tremor in the 20-30 Hz range has been shown to be due to beat frequencies produced due to the oscillation of folds with slightly different masses and hence slightly different fundamental frequencies.

Table 4.1 lists all the perturbation methods that are used at present and were calculated in the present study. Despite the number of methods shown, the use of some of the indices are of questionable importance and others are not strictly independent of one another. Generally, the measures are some slight variation of PF1 as given in equation 4.6 (e.g. RAPP, as shown in equation 4.7). The source code (pperb.m) for the formulae is given in appendix x. In a comprehensive report of such methods Zyski et al have shown that the APPP was the best predictor of vocal pathology followed by

RAPP. Askenfelt et al[13] also carried out a test comparing seven different perturbation measures and found that the standard deviation of the distribution of the relative frequency differences was the

PERTURBATION MEASURE	ABBREVIATION
Average pitch perturbation	APP
Relative Average pitch perturbation	RAPP
Average percentage pitch perturbation	APPP
Normalised std.dev. pitch	stdndf0
Mean 1st order perturbation	PF1
Mean 2nd order perturbation	PF2
Directional perturbation factor	DPF
Normalised std.dev. of 2nd order pert.	stdnd2f0
Std.dev. of pitch perturbation	stddf0

Table 4.1 *Pitch Perturbation Measures (source code - Appendix A (pperb.m))*

most useful acoustic measure for use in clinical applications. This standard deviation is a measure of the DF0 distribution, where DF0 = (Fn+1 - Fn)/Fn . An example of such a distribution is shown in figure 4.13 for a normal and pathologic speaker.

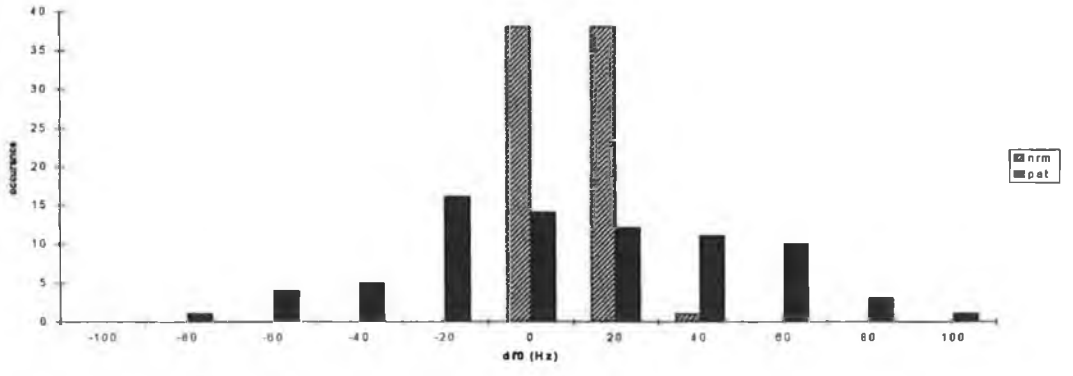
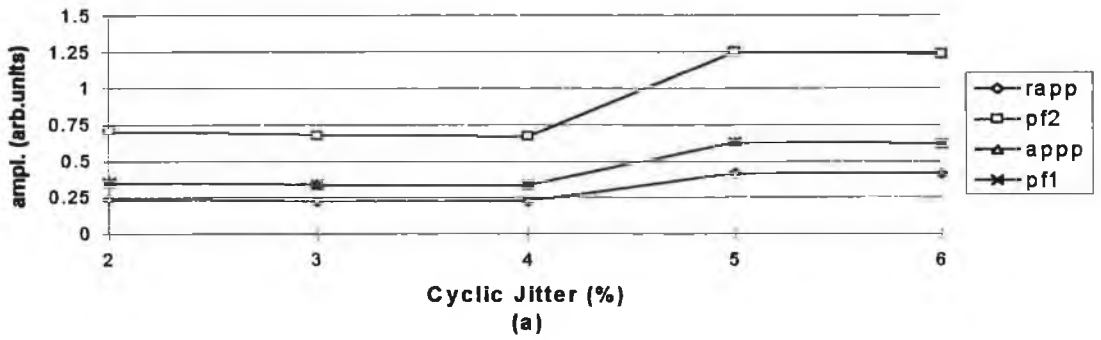
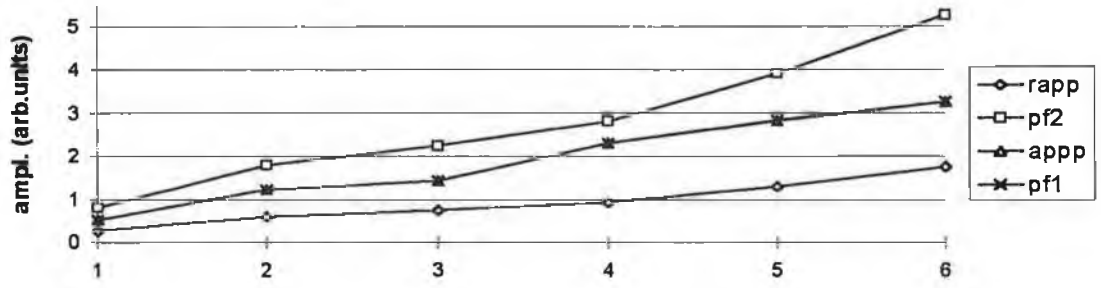


fig.4.13 *Histogram of the difference in f0 (df0) distribution for a patient and normal of the present study*

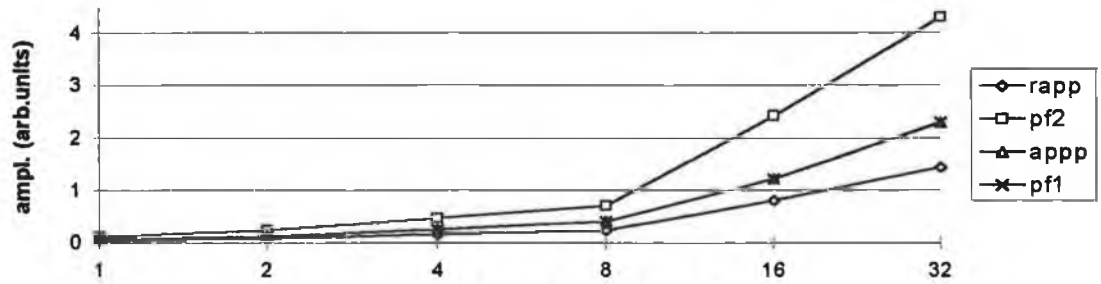
As stated in section 4.2.2 Pinto et al have also investigated the distribution of perturbation measures and have attempted to unify the classification of perturbation measures through the use of forward and backward difference equations. We have followed their recommendation that fundamental frequency as opposed to pitch period be used when investigating jitter and all the jitter perturbations used in this study were calculated using the fundamental frequency as opposed to the pitch period. The variation of four pitch perturbation measures with respect to the four different source perturbations are shown in fig.4.14(a,b,c,d). All measures are shown to give essentially the same information with PF1 and APPP being essentially the same index.



(a)



(b)



(c)

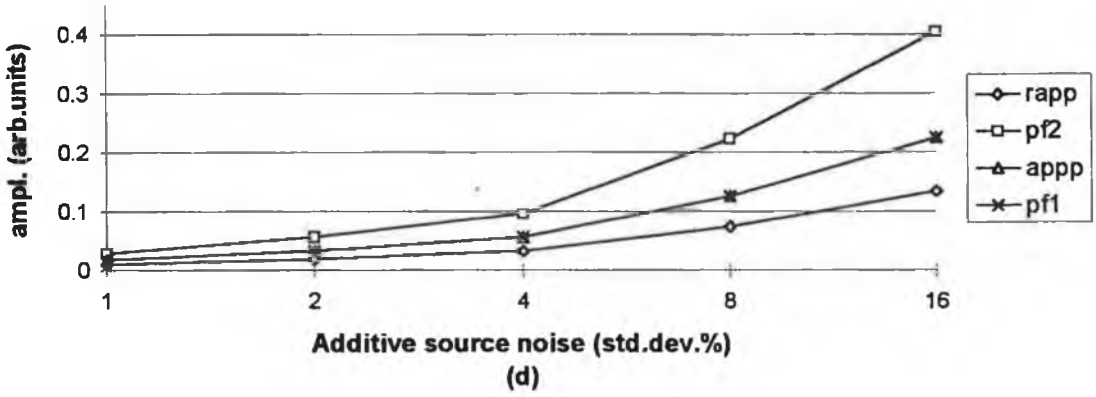


fig.4.14 *Variation of four pitch perturbation measures (RAPP, PF2, APPP and PF1) with respect to the four source perturabtion measures of (a) cyclic jitter (b) random jitter (c) shimmer and (d) additive noise*

Up until now we have only considered the indices as showing general trends. However the idea of using the indices is to introduce accurate, quantifiable measures of aperiodicity. Therefore, if our signal is synthesized with 6% jitter then a first order measure should show 6% jitter or else there is an error in the tracking or the index. The first requirement of our pitch tracker is that it should follow the changes in f0 perturbation accurately. However, as we have seen in section 4.2, no method follows the jitter increase in a linear fashion as we might have expected. The reason for this is not so much due to any limitations of the pitch trackers or indices but is more a consequence of the scaling and integer requirements of the glottal pulses as mentioned above. This point is illustrated with the cyclic jitter values. Fig.4.7 shows 3 levels of jitter which are given here in tabular form (Table.4.2). Therefore caution must be taken when interpreting the indices returned from a given extraction procedure for the synthesis files of given jitter levels (i.e. the discrepancy between values has arisen due to the constrains of the synthesis files as opposed to any limitations of the pitch trackers or perturbation indices). Keeping all parameters in the glottal model scaled accurately can only be achieved by increasing the sampling frequency. A simpler approach would be to simply truncate the closed phase. Although, this is not the desired approach from the spectral characterisation viewpoint, it does provide a simple means of assessing the accuracy of pitch extraction procedures in a consistent manner.

EXTRACTION METHOD/ SYNTHESIS FILE	POSITIVE PEAKS (PP) (% JITTER)	POSITIVE PEAKS LOW PASSED-PPL (% JITTER)	POSITIVE ZERO CROSSINGS (PZC) (% JITTER)	ACTUAL SYNTHESIS VALUE (% JITTER)
s110jp6	3.93	2.55	0.62	6.39
g110jp6	4.12	1.90	0.89	
s110jp5	4.12	2.17	0.67	5.35
g110jp5	4.17	2.17	0.79	
s110jp4	2.01	1.17	0.32	4.3
g110jp4	2.10	0.94	0.41	
s110jp3	2.04	1.15	0.33	3.24
g110jp3	2.13	1.01	0.41	
g110jp2	2.20	1.05	0.34	2.17
s110jp2	2.15	0.95	0.41	
g110jp1	0	0	0	1.09
s110jp1	0	0	0	

Table 4.2 *Cyclic jitter values for the source and radiated waveforms for all extraction procedures compared to actual synthesis value for cyclic jitter (right hand column).*

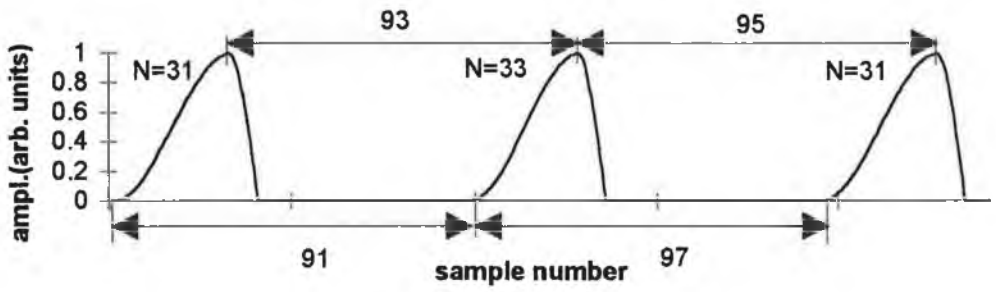


fig.4.15 *Integer constraint on scaling of glottal pulse model for 6 % cyclic jitter*

The integer constraints only apply to the jitter set synthesis data and therefore the measurement of jitter in the presence of noise or shimmer can be taken as accurate.

4.4 Measurement of Amplitude Perturbation

The measurement of amplitude perturbation has also received considerable attention in the literature. The Zyski et al review is based on measurements that were taken from speech samples recorded using a contact microphone in which the amplitude of the resultant waveform is somewhat more directly related to the glottal waveform than when simply using the standard audio microphone. Gauffin and Sundberg[14] have shown that the peak amplitude of the flow glottogram waveform is in fact related more closely to the amplitude of the fundamental in the glottal source spectrum than to the overall intensity of the speech waveform and that the negative peak amplitude of the differentiated flow glottogram shows a high correlation with sound pressure level. Therefore the peak (or rms, if we consider the waveform to be quasi-periodic) amplitude in the output radiated speech waveform can be considered to relate more closely to the peak to peak of the differentiated flow glottogram signal than the peak flow of the glottogram signal. Furthermore, Hillenbrand has shown that the same shimmer levels result for synthesis data with jitter, shimmer and additive noise perturbations, regardless of whether peak or rms amplitudes of the output radiated speech waveform were used. As shown in figure 4.16(a) values for measured shimmer reflect well the source amplitude perturbation. In a) the HPF1 measure, defined as

$$\mathrm{HPF1} = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{\mathrm{Hf}(i+1)-\mathrm{Hf}(i)}{0.5\times[(\mathrm{Hf}(i+1)+\mathrm{Hf}(i))]}\times 100 \qquad \text{eqtn.4.8}$$

Hf = peak amplitude of waveform within a cycle

N = total no. of periods

or in words, the variation in peak amplitude of the waveform from cycle to cycle, divided by the amplitude of the waveform shows good correlation with the source amplitude perturbation levels. The effect of the vocal tract filter function does not alter the amount of shimmer measured. This implies a direct correlation between the peak in

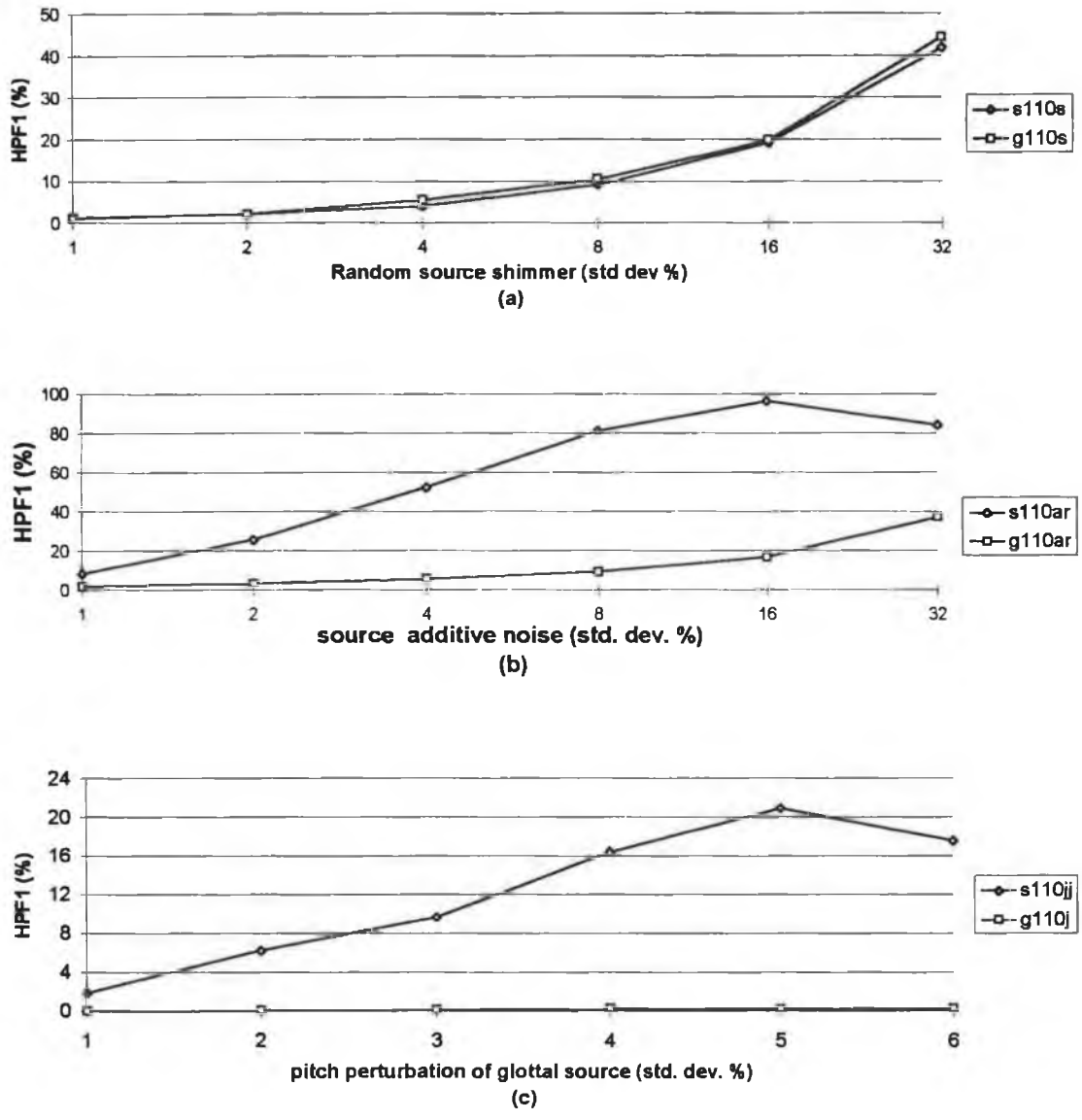


fig.4.16 *Amplitude perturbation factor one (HPF1) vs (a) shimmer, (b) additive noise and (c) random jitter*

the source model and the peak in the output radiated speech waveform. Part b) of the figure shows that shimmer measures (HPF1) are very sensitive to random noise introduced at the glottal source. Also note that the source amplitude perturbation increases in a regular fashion whereas the vocal tract filtered waveform's amplitude perturbation is somewhat less regular. Part c) of the figure shows the amplitude perturbation measure plotted with respect to jitter. For the source signal, jitter has

very little effect on amplitude perturbation. However, for the filtered signal there is a marked effect on the amplitude perturbation, and in contrast to the jitter measurements obtained in the presence of shimmer, which were due to the measurement technique rather than increased aperiodicity of the signal, the amplitude perturbation measurement reflects an actual increase in shimmer. This is due to the source filter or harmonic formant interaction as stated by Imaizumi[15]. As the source period changes, the vocal tract filter is excited with different harmonic frequencies, therefore receiving different resonance contributions and hence differences in the amplitude of the waveform from period to period. A list of shimmer values that were evaluated in the present study are shown in table 4.3. There is a one to one correspondence with the pitch perturbation values, except for a further dB measure included in table 4.3. The same comments noted for the jitter measures are also true for the shimmer measures listed here.

PERTURBATION MEASURE	ABBREVIATION
Average amplitude perturbation	AAP
Relative Average ampl. Perturbation	RAAP
Average percentage ampl. perturbation	APAP
Normalised std.dev. amplitude	stdndHf0
Mean 1st order perturbation	HPF1
Mean 2nd order perturbation	HPF2
Directional perturbation factor	DHPF
Normalised std.dev. of 2nd order pert.	stdndH2f0
Std.dev. of ampl. Perturbation	stddHf0
Average power difference (shimmer)	dBdHf0

Table 4.3 *Amplitude perturbation measures (source code- appendix A (amperb.m))*

In recognition of the fact that the various jitter and shimmer indices represent basically the same information, only the pitch perturbation factor one (PF1) and amplitude perturbation factor one (HPF1) are shown for the patient and normal data fig.4.17(a,b).

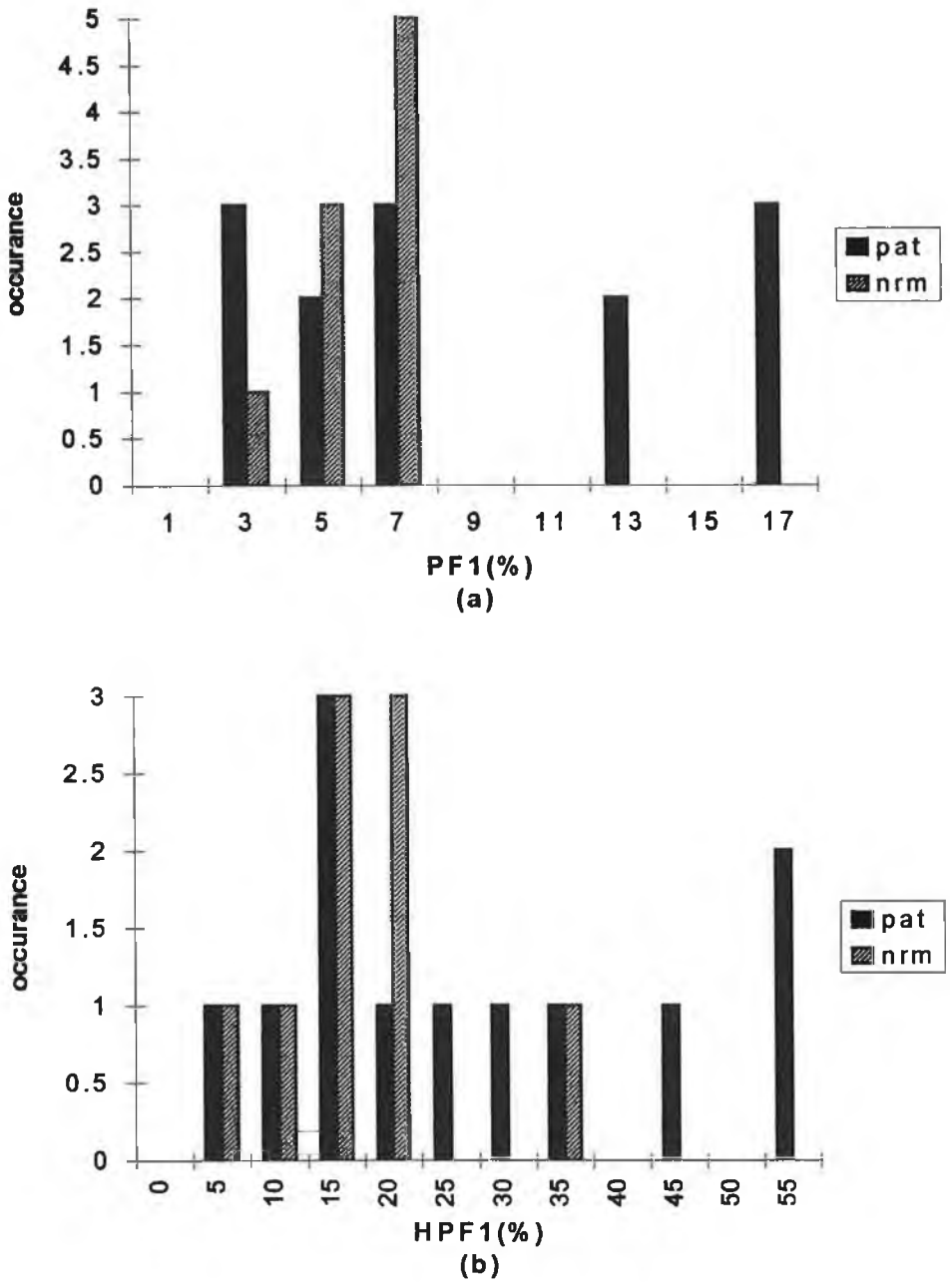


fig. 4.17 *Histogram of (a) PF1 and (b) HPF1 for the patient and normal data*

The poor separability of the patient/normal data set is not surprising for many reasons. Firstly, the poor recording conditions inflated all jitter measures therefore reducing the accuracy of the methods. The data set for the patients varies from mild functional dysphonia to severe vocal pathology. For functional dysphonias which have breathiness as a cardinal symptom, one could hypothesize that the aperiodicity may not

increase to any great extent. Due to the more sinusoidal nature of the waveform (although aspiration noise is also present), waveform matching methods may show lower levels of jitter than for the more complex 'normal' waveforms. However, five of the thirteen samples from the patient data set show a clear difference in PF1 to the normal data set and are from patients with severe organic vocal pathology. Note the values for PF1 range from 1 to 7 % for the normals in this study, whereas Horii[16] has stated that 0.1 to 1% perturbation are in the range of normal. Although we have chosen the variation of fundamental frequency as opposed to pitch period in the evaluation of jitter, the PF1 measure is a time-frequency invariant measurement as indicated by equation 4.9

$$\frac{\Delta f0}{f0} = \frac{\Delta T}{T} \qquad \text{eqtn.4.9}$$

and as such can not be considered to be a cause of the above discrepancy. The difference is due to poor quality tape recording as mentioned above, which has been shown to significantly increase jitter values (2 refs). The effect of the poor quality recording is particularly strong here as the recorder exhibited a strong "Watergate Buzz" (i.e. mains frequency and odd integer harmonics) in it's frequency response characteristic. In an effort to remove these unwanted artifacts, all speech signals were high pass filtered at 60 Hz using a Ramiz filter. More sophisticated methods, that remove the higher harmonic noise components using comb filtering have also been developed for this purpose.

4.5 Autocorrelation and Correlation Analysis

Many other time domain methods exist other than those mentioned above in respect to perturbation analyses. In fact in chapter 5 the harmonic to noise ratio is calculated using an adaptation of Yumoto's time domain signal to noise ratio estimate[17].

Kasuya's time domain filter[18], again providing an harmonic to noise ratio estimate was also programmed and the source code is given in appendix.x. The peak to mean of the output waveform value was also suggested as a useful measure by Klatt et al[19] and Fant[20] has shown that the bandwidth of the first formant can also be estimated from calculations performed on the output waveform.

Another popular time domain method is the correlation or autocorrelation function, defined as

$$\mathrm{Corr}(g,h)_j = \sum_{k=0}^{N-1} g_{j+k} h_k , \qquad \text{eqtn.4.10}$$

g,h are periodic with period N.

$g_k = h_k$ for autocorrelation.

This measures the similarity of the waveform from period to period. A property of the autocorrelation function is that if the function being correlated is periodic then the autocorrelation is also period with the same frequency. This property has been successfully used in order to provide accurate pitch estimation. Rabiner gives an extensive list of processing details in order to enhance the pitch estimate using the autocorrelation function. A program was written in order to implement this estimate (timepit.m) but our main concern here is the use of the correlation function for providing a similarity index as opposed to a pitch estimate. The basic idea, as introduced by Hillenbrand is that a waveform with good periodicity exhibits strong similarity between adjacent cycles and hence will have a high correlation index. In the program implemented here, the correlation index proposed by Hillenbrand[21] (XCP) and anew measure (MM2) were calculated for the original and low and band pass versions of the original waveforms. The method involves taking the correlation of the original waveform with a delayed copy of itself at delays between the maximum and minimum expected pitch period (3.3ms and 16.7 ms were used). For periodic signals a peak occurs in the correlation function at a delay corresponding to the fundamental period. As the correlation peak is dependent on the signal amplitude a normalisation scheme is required. Two normalised correlation indices were calculated every 10 ms using a 30 ms analysis frame. The first was the standard measure of the peak in the

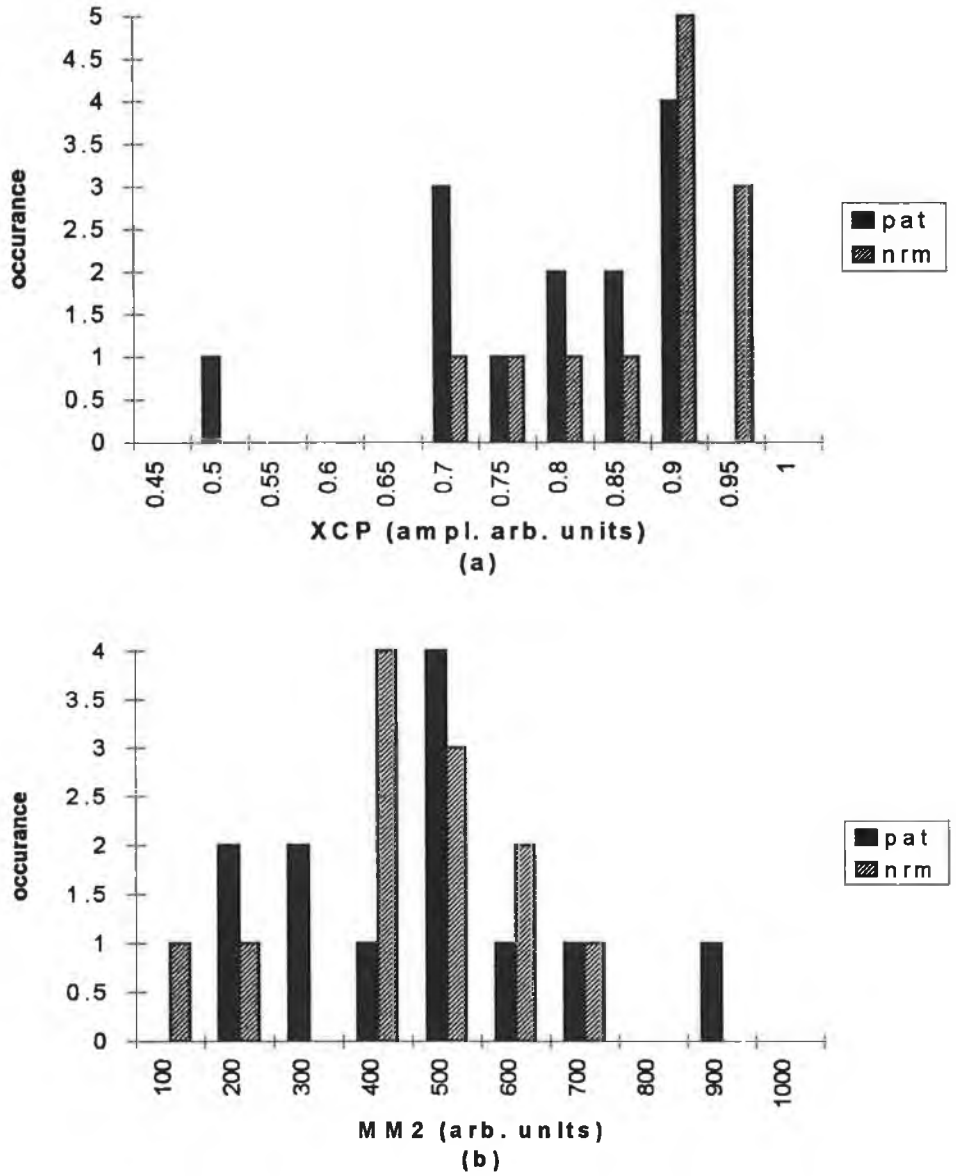


fig.4.18 *Histogram of correlation indices (a) XCP and (b) MM2 for normal and patient data sets*

autocorrelation function (XCP), the second was calculated by simply taking the standard deviation of the difference between waveforms when the overlap corresponded to the peak in the autocorrelation function (MM2). The performance of the two measures with respect to the patient/normal data set is shown in fig. 4.18. As can be seen in the figure, the measures show a strong overlap between the patient and normal data sets.

4.6 Discussion/Conclusion :

In attempting to provide objective acoustic indices of vocal pathology based on time domain analysis methods the main thrust has been towards extracting measures of pitch and amplitude perturbation. The original strategy was firstly to find a pitch extraction method that gives reliable pitch estimates in the presence of additive noise and shimmer and then to investigate the nature of the perturbation by choosing suitable perturbation indices from the list given in table 4.1. However, we have shown that different pitch extraction methods perform better depending on the source perturbation present. When shimmer is present the method giving the lowest jitter index is the positive peaks from the original waveform. When random noise is added to the source signal using the positive peaks to extract the pitch period gives very high jitter values. Conversely, the positive zero crossing of the low pass waveform give robust pitch estimates in the presence of additive noise and poor estimates in the presence of shimmer. Therefore a comparison of the jitter indices returned from different extraction methods could possibly provide information regarding the source of the perturbation. The results for the waveform matching (low passed) are of considerable interest (fig. 4.8). The least squares error estimate of waveform matching reported by Titze et al[9] only examines waveforms with a minimum of 10 dB signal to noise ratio. However, signal to noise ratios for pathological voices would be expected to fall below this value. Fig.4.8 shows that the waveform matching is very robust up until std. dev. 8 % (~ 15 dB - output file) additive noise but once this value is exceeded the method rapidly deteriorates with positive zero crossings showing considerably lower jitter values. This suggests that the waveform matching method of pitch extraction may not be the most applicable technique to use on pathological voice types.

Furthermore, it has been shown that although several pitch and amplitude perturbations measures exist many are redundant or offer no new information regarding the signal. Despite this fact, different (new) perturbation measures may be useful in differentiating perturbation types. For the cyclic and random jitter signals of the present study, a simple second order perturbation measure subtracting every second period reveals

which type of perturbation is present. The original intention of introducing these files was to show that for a given % pitch perturbation, the spectral characteristics can be quite different and therefore a perturbation of a given value could arise due to very different vibratory characteristics of the vocal folds. This index comparison could not be tested due to the limitations of the jitter synthesis data as stated. There are further considerations to bear in mind in considering cyclic jitter in pathological voice types (e.g. vocal creak). For the synthesis data it was known a priori what the actual fundamental was and filtering began at 1.5 times this value but for real data our pitch trackers could easily chose 'two cycles' as the fundamental and therefore give zero perturbation indices. Further work is required to produce useful perturbation indices under these conditions. In regard to jitter measurements in the presence of additive noise and shimmer we have shown that the increase in jitter is due to measurement error that arises due the presence of these perturbations. In respect of the low pass results for shimmer, a simple normalisation scheme would perhaps solve this problem. However in the case of shimmer in the presence of jitter we have shown the increase in shimmer to be a result of the source filter interaction as a result of different fundamental frequencies exciting the vocal tract. Also, we have shown that shimmer measurements are more strongly affected by additive noise than actual shimmer levels present in the signal. A more robust correlation measure would be obtained if it compared not only adjacent periods of the waveform but also the first (then second etc) with the third, fourth etc. until the complete time record is finished. Presently used indices based on commercially available speech software packages "cannot reliably be applied to voices that are even mildly aperiodic". The conclusion reached from this study is that present perturbation indices have some utility but that the measures could be greatly advanced through careful consideration of the extraction procedures used along with the implementation of new indices based on a knowledge of vibratory characteristics.

4.7 Bibliography

1. Zyski, BJ. et al Perturbation analyses of normal and pathologic larynges. Folia phoniat. 1984, **36**: 190-198
2. Hess, W. Pitch determination of speech signals, algorithms and devices. Berlin, Heidelberg: Springer Verlag, 1983
3. Rabiner, LR. On the use of autocorrelation for pitch detection. IEEE Trans. Acoust. Speech and Sig. Processing, 1977; ASSP-**25**: 24-33
4. Rabiner, L. and Schafer, R. Digitial processing of speech signals, Englewood Cliffs, N.J.: Prentice Hall, 1978
5. Deller, J. Proakis J. and Hansen, J. Discrete time processing of speech signals, NY: Macmillan, 1989
6. Milenkovic, P. Least mean square measures of voice perturbation. J. Speech and Hear. Res. 1987, **30**: 529-538
7. Deem, JF. et al The automatic extraction of pitch perturbation using microcomputers: Some methodological considerations. J. Speech and Hear. Res. **32**: 689-697
8. Hillenbrand, J. A methodological study of perturbation and additive noise in synthetically generated voice signals. J. Speech and Hear. Res. 1987; **30**: 448-461
9. Titze, IR. and Liang. H. Comparison of f0 extraction methods for high precision voice perturbation measurements. 1993; **36**: 1120-1133
10. Pinto, N. and Titze, IR. Unification of perturbation measures in speech analysis. J. Acoust. Soc. Am. 1990; **87**: 1278-1289
11. Lieberman, P. Some acoustic measures of the fundamental periodicity of normal and pathological larynges. J. Acoust. Soc. Am. 1963; **35**: 344-353
12. Koike. Y. application of some acoustic measures for evaluation of laryngeal dysfunction. Studia Phonologica 1973; **VII**: 17-23
13. Titze, IR. Coupling of neural and mechanical oscillators in control of pitch, vibrato and tremor. In PJ. Davis and NH. Fletcher (Eds.) Controlling complexity and chaos. San Diego: Singular Publ. Group, 1996

14. Askenfelt, AG. and Hammarberg, B. Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures. J. Speech and Hearing Res. 1986; **29**: 50-64

15. Gauffin, J. and Sundberg, J. spectral correlates of glottal voice source waveform characteristics. J. Speech and Hear. Res. 1989; **32**: 556-565

16. Imaizumi, S. Acoustic measurement of pathological voice qualities for medical purposes, ICASSP, Tokyo, IEEE, 1986

17. Horii, Y. Fundamental frequency perturbation observed in sustained phonation. J. Speech and Hearing Res. 1979; **18**: 19-201

18. Gelfer, MP. And Fendel, DM. Comparisons of jitter, shimmer and signal to noise ratio from directly digitised versus taped voice samples. J. Voice 1995; **9**: 378-382

19. Parsons, T. Voice and speech processing. NY,NY: McGraw-Hill, 1986

20. Yumoto, E. et al Harmonics to noise ratio as an index of the degree of hoarseness. J. Acoust. Soc. Am. 1982; **71**: 1544-1550

21. Kasuya, H. An adaptive comb filtering method as applied to acoustic analysis of pathological voice. ICASSP 1986, Tokyo, IEEE 669-672

22. Klatt, DH. and Klatt, LC. Analysis, synthesis and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 1990; **87**: 820-857

23. Fant, G. Acoustic theory of speech production. Mouton, The Hague, 1970.

24. Hillenbrand, J. et al Acoustic correlates of breathy vocal quality. J. Speech and Hear. Res. 1994; **37**: 769-777

Chapter 5

Harmonic Intensity Analysis

5.1 Introduction

A review of the literature on acoustic analysis of vocal pathology reveals that along with jitter and shimmer, the most commonly studied acoustic symptom of pathological voice has been the presence of noise in the acoustic speech waveform[1]. Commercial software packages have recently become available that provide indices for jitter, shimmer and harmonic to noise ratios. However a comparison study of the indices produced by the various packages led Bielamowicz[2] et al to "question their utility in quantifying vocal quality, especially in pathological voices". Other researchers have similarly found noise difficult to quantify or as stated by Hillenbrand[1], "the precise quantification of noise levels has not proven to be a simple matter". Although jitter, shimmer and noise levels are all readily observable on either sonographic or spectrographic displays, only the former two (i.e. jitter and shimmer) have been satisfactorily quantified (at least for normal voices).

Early attempts to quantify the level of noise in pathological voices were based on (subjective) visual inspection of voice spectrograms. Yanagihara[3] proposed a five point rating scale of the noise level in the spectrogram which correlated with pathological voice rating as assessed by three listeners. These ratings have since been used to calibrate more objective measures of noise levels and this raises a very important question about grading noise levels with respect to vocal pathology. Two approaches seem applicable: one is to grade the noise index with respect to a trained listener rating[4], or alternatively to rate the noise level with respect to the degree of the pathology (physiological/anatomical) present[5] or even whether a pathology is present or not[6]. The difficulties and variability surrounding each of these approaches explains in part why there is an absence of a database of rated pathological voices from which researchers can test new methods of analysis[7]. Other studies have focused on direct dynamic changes of the vocal cords as viewed using digital imagery, x-rays or stroboscopy and correlated these observations with acoustic findings including noise. Whatever the correlation procedure followed, objective classification of a given vocal quality requires a very high, multi-dimensional rating scheme. Nevertheless, broad terms are also useful for determining another important goal, which is to state whether a voice can be thought of as normal or pathologic. It should also be noted that an index of some significance to perceptual judgments may have little use as a correlate of physical characteristics (and vice versa). With these potential problems in mind we turn our attention to the quantification of noise levels in pathological voices.

Narrowband spectrograms for a patient and normal of the present study are shown in figure. 5.1 and broadband spectrograms for the same utterances are shown in fig.5.2. Rontal et al[8] reported positive objective analysis using broadband spectrograms and cited several advantages in using spectrography, such as the ability to keep a permanent record and the ability to analyse continuous speech. Disadvantages, are that it is a visual comparison and in that sense still subjective and some training or at least familiarity is required to be able to read the spectrographic images effectively and hence make useful diagnoses.

As pointed out by Rontal, clinicians have been slow in using the spectrographic ratings. What is preferable to the clinician is a simple index. Many possible solutions have been

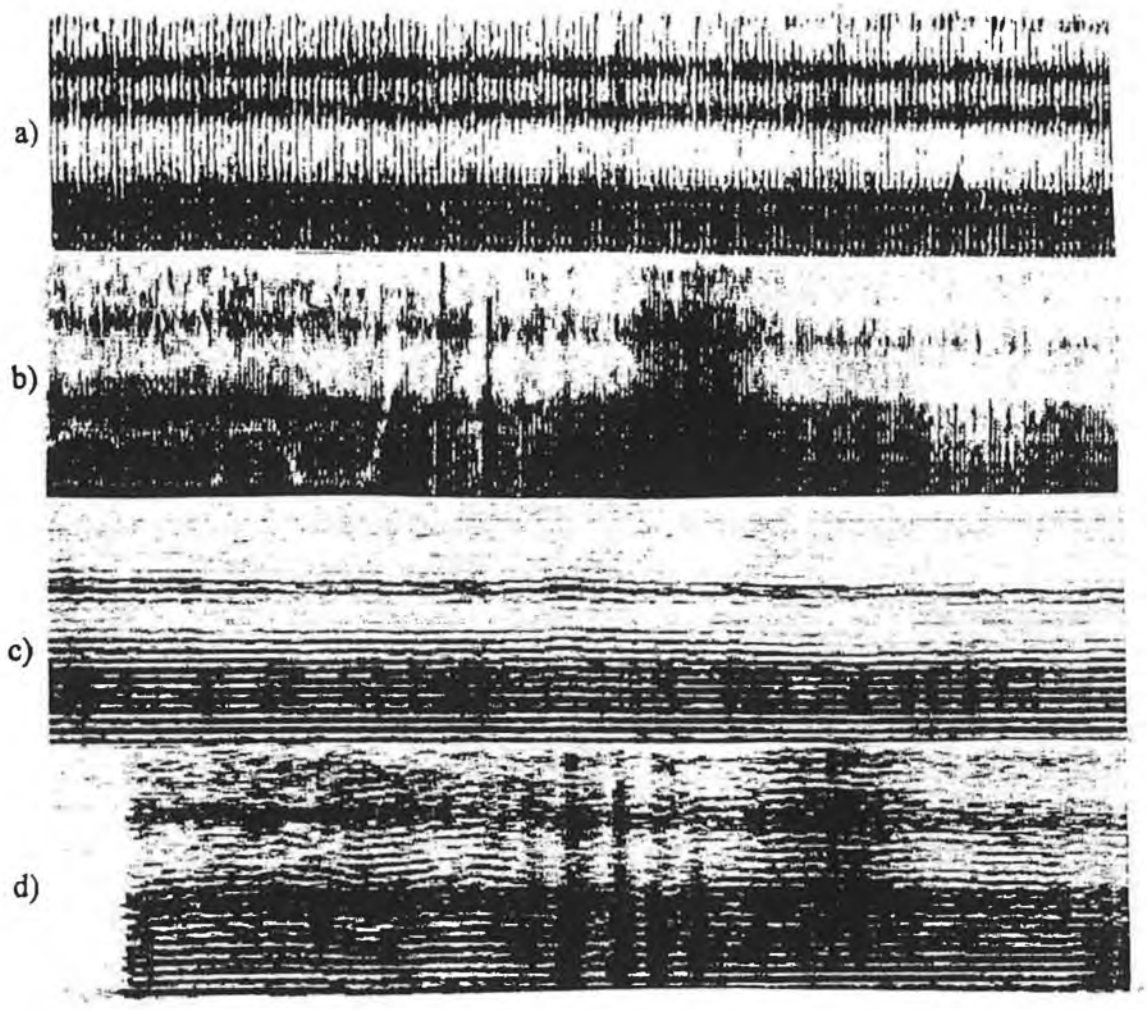


fig.5.1 *Broadband spectrograms for a (a) normal and (b) patient (nodule) from the present study for a sustained phonation of the vowel sound a/. Narrowband spectrograms of the same utterance (c) normal and (d) patient. Frequency is plotted on the y-axis (0-4 kHz) and the x-axis represents time(s)- each row represents 1.6 sec)*

put forward in an attempt to provide such a noise index. Indices such as the H/N (harmonic to noise) and S/N (signal to noise) ratios calculated from such diverse methodologies as spectral, cepstral, time domain averaging (as per traditional S/N ratio measurements) and wavelet analysis have been investigated. The authors when presenting the different measurement techniques often begin by citing weaknesses in other approaches with respect to their own and showing improvements made and sometimes a brief numerical comparison with another method.

An examination is given here of six of these methods. This followed a literature search which revealed a total of twelve methods. Three of the other methods are alluded to,

although not investigated. The basic idea behind each method is carefully developed and adjustments (and different interpretations) are given where it has been felt necessary. The details of the methods are given in depth in section 5.3 but firstly we turn our attention to three very important issues associated with determining and interpreting H/N ratios. The development motivates the need for new measurement techniques and new analysis ratios.

Firstly, to what extent do presently used indices represent the amount of 'noise' present in a signal as opposed to the presence of other perturbation measures such as jitter and shimmer ? This question is carefully addressed and the approaches by which to overcome the methological issues are clearly stated. The different spectral consequences are clearly illustrated, hence providing an approach for independent measurement of jitter, shimmer and additive noise.

The question (often overlooked) as to what is noise is addressed and defined. The H/N and S/N are clearly explained and the relationship between the H/N (S/N) at the source is related to the H/N at the output. Defining these relationships leads to a discussion on what is the most pertinent measurement to make on the acoustic speech waveform. Several possibilities exist:

1 level of harmonics
2 level of noise
3 magnitude of H/N (S/N)
4 level of H/N at a given frequency location
5 geometric ratios
6 limiting the frequency range
7. Ratio between various harmonic numbers (or harmonic regions)

A discussion is given on what inferences are to be made from these spectral measurements with respect to glottal flow and hence the vibratory pattern of the vocal cords.

5.2 Harmonic Intensity Analysis: Preliminary Considerations

5.2.1 Definition of Noise

Conspicuously absent from the literature relating to noise levels in pathological voices is a clear definition of what is meant by 'noise'. An operational definition generally given (or more often implied) is that noise constitutes the non-harmonic energy found in the speech signal. Thus, a perfectly periodic waveform exhibits an infinite harmonic to noise ratio: the unperturbed synthesis files used in the present study give harmonic to noise ratios of 300 dB. Contrary to this, the waveforms from real voices vary to a certain extent, due to such effects as flutter or tremor and therefore contain 'noise' energy with harmonic to noise ratios in the 20 to 30 dB range being typical (depends on ratio type and methodology). Also, jitter and shimmer artifacts contribute to a reduction in measured H/N ratios since by this definition they are properly labelled noise components. Consequently, for an overall measurement of 'noise' according to the above definition, all perturbation artifacts should be included.

Nonetheless, we are also concerned with characterising the vibratory pattern of the vocal folds based on the waveform analysis and in respect of this, we are required to differentiate, if possible the different origins of the noise. Four distinct possibilities exist.

1. Variation of pitch period (jitter)
2. Variation of peak amplitude from cycle to cycle (shimmer)
3. (Additive) Noise
4. Variation of waveform within a vibratory cycle

Number 3 refers to the turbulent flow produced at the glottis during phonation, perhaps due to lack of, or, incomplete closure or due to the presence of mass lesions.

Stevens et al[9] have carried out some investigations into the nature of the turbulent flow, although further studies relating to vocal pathologies are required. This noise source is generally modelled as random, mean zero, Gaussian noise. The use of the word 'noise' here leaves room for ambiguity. However, the context of the word usage usually suffices and we generally refer to the noise of turbulent origin as 'additive noise'. The more general usage of the word 'noise' in the vocal pathology literature implicitly means noise of some origin, other than 1,2 and 4 above i.e. noise of turbulent (or possibly other) origin. Except for the definition outline of noise given in this section we also use the word 'noise' in the narrow sense of the meaning.

Finally, 'noise', can also occur due to specific changes within the vibratory pattern from cycle to cycle which are not due to shimmer or jitter but due to a change in the shape of the glottal waveform (number 4 above). In order to be able to study this latter characteristic we need to have precise knowledge about the other three noise elements. Lastly, it should be pointed out that, in theory at least, that a noise free signal by the above definition could also show considerable pathology, i.e. the waveshape could be quite irregular, yet consistent from period to period. Conversely, we can imagine a situation in which the period markers are fixed for each cycle but the waveform behaves very erratically between the period markers. Spasmodic dysphonia is an example of a waveform containing irregular, unrelated waveshapes of similar period.

5.2.2 Spectral Consequences of Jitter, Shimmer and Additive Noise:

Many authors[1,10,11] have observed that H/N ratios may not simply reflect the amount of additive noise present in a voice signal or as reported by Muta et al[10] "glottal source perturbations distort the harmonic structure and thus affect both noise measures and harmonic strength measures." If our ultimate aim is to categorise or make direct inferences about specific vibratory or glottal events based on acoustic analysis of the output waveform, then it is of paramount importance to be able to make measurements of jitter, shimmer and additive noise that are independent of each other. The

interaction effects of these three parameters has been studied by Hillenbrand[1] who concluded that there are strong measurement interactions among the three variables and that "caution should be exercised in interpreting measures of perturbation and noise in terms of specific aspects of the laryngeal vibratory cycle". For example, adding increasing amounts of jitter not only affects the pitch perturbation but also reduces the harmonic to noise ratio of the signal. Alternatively, adding noise to the signal causes an increase in the measured jitter and shimmer as well as reducing the H/N ratio. It is the former problem that we are concerned with here, i.e. the effect of jitter (and shimmer) on the H/N ratio.

In an attempt to isolate the origin of the 'noise', the spectral consequences of adding different amounts of additive noise, jitter and shimmer are investigated. This is firstly investigated by referring directly to the Fourier series calculation, from which possible spectral characterisations of each perturbation measure are postulated. The spectra for the synthesis files are then examined in order to test the hypotheses.

In applying the Fourier series (eqtn.5.1), two periods of a perfectly periodic sine wave with a total time record length, T are considered, as shown in figure 5.2. (In considering shimmer, the glottal pulse is used in place of sine waves).

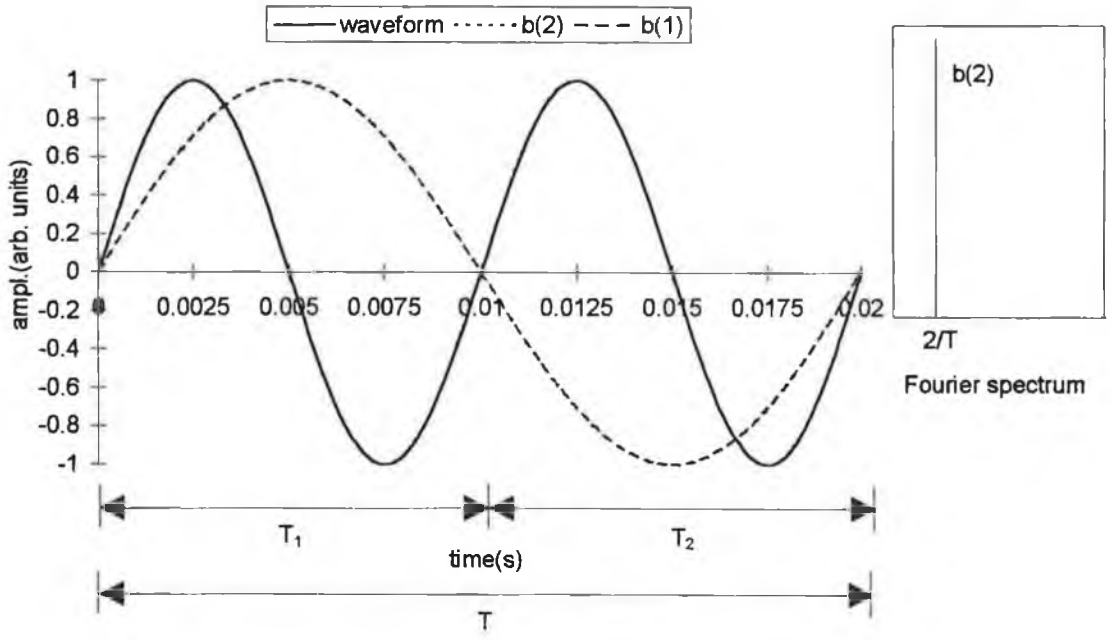


fig.5.2 *Two periods of a sine wave with $T_1 = T_2 = T/2$. The sine functions for the first two Fourier coefficients are shown (B_1 and B_2 (coincident with waveform)).*

The basis behind the method is that if each waveform of period, T_1, is exactly the same, then harmonics will only appear in the spectrum at integer multiples of $(1/T_1=2/T)$. However, if the waveforms differ in any respect, then energy appears in the spectrum at integer multiples of (1/T). It should be noted therefore, that in this development, odd harmonics signify some form of perturbation and even harmonics represent the unperturbered waveform. Of course, in this case, since the waveform is simply a sine wave we obtain a single spectral peak at $f=1/T_1$ for the unperturbered signal, the harmonic energy at all other locations being equal to zero.

$$f(t) = \frac{A_0}{2} + \sum_{n=1}^{\infty} (A_k \cos(\frac{2k\Pi t}{T}) + B_k \sin(\frac{2k\Pi t}{T}))$$

eqtn.5.1

where $A_k = \frac{2}{T}\int_0^T f(t)\cos(\frac{2k\Pi t}{T})dt$ eqtn.5.2

and

$$B_k = \frac{2}{T}\int_0^T f(t)\sin(\frac{2k\Pi t}{T})dt$$ eqtn.5.3

t = time (Fourier Series)

T = periodic time

A_0 = mean value of waveform

To see how the Fourier series arrives at this estimate we consider the sine terms of the series for each harmonic location k×(1/T) (cosine terms are zero for odd functions). Fitting the first harmonic, 1/T Hz to the waveform in fig. 5.3 and taking the sum for the Fourier coefficient, $B_1(1/T)$ we see that (in consideration of the equation 5.3 and fig. 5.3) what is obtained in the positive half of the cycle is also obtained in the negative half of the cycle with all contributions to the sum adding to zero. A similar result is obtained for all higher terms in the series, except for 2×(1/T) when the contributions add constructively. This is a completely general analysis not specific to sine waves. In

$$B_k = \frac{2}{T}\int_0^T \sin(\frac{2 \times 2\Pi t}{T})\sin(\frac{2k\Pi t}{T})dt \qquad \text{eqtn.5.4}$$

putting k=2, the integrand becomes

$$B_2 = \frac{2}{T}\int_0^T \sin^2(\frac{2 \times 2\Pi t}{T})\,dt = \frac{2}{T}\int_0^T 1/2(1-\cos(\frac{8\Pi t}{T})\,dt \qquad \text{eqtn.5.5}$$

$$\frac{2}{T}\left[\frac{t}{2} - \frac{T}{2 \times 8 \times \Pi}\sin(\frac{8\Pi t}{T})\right]_0^T = 1 \qquad \text{eqtn.5.6}$$

and for $k \neq 2$, eqtn5.4 becomes

$$B_k = \frac{2}{T}\int_0^T \frac{1}{2}\left[\cos(\frac{(2-k) \times 2\Pi t}{T}) - \cos(\frac{(2+k) \times 2\Pi t}{T})\right]dt \qquad \text{eqtn.5.7}$$

where we have used the trigonometric identity

$$\sin A \sin B = \frac{1}{2}\left[\cos(A-B) - \cos(A+B)\right] \qquad \text{eqtn.5.8}$$

$$\frac{2}{T}\left[\frac{1}{2}(\frac{T}{2\Pi(2-k)})\sin(\frac{(2-k)2\Pi t}{T}) - \frac{1}{2}(\frac{T}{2\Pi(2+k)})\sin(\frac{(2+k)2\Pi t}{T})\right]_0^T = 0 \qquad \text{eqtn.5.9}$$

Therefore showing that our heuristic development is correct i.e. energy is present at the '2×1/T' harmonic (B_2) and zero at all higher harmonics.

Now let us consider the case of shimmer shown in fig.5.3. In utilising the sine wave in this analysis the result returned is simply a sinewave with amplitude equal to the average of the sinewave amplitudes as we would expect from the analytical expression. It is interesting to see the result based on the graphical approach.

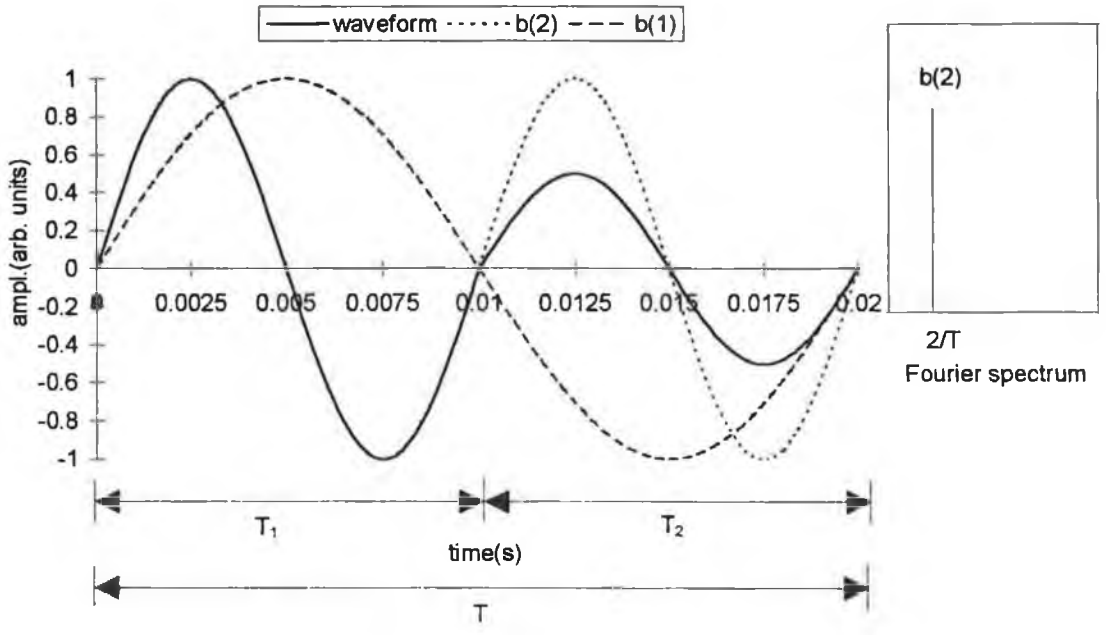


Fig.5.3 *Two periods of a sine wave (with $T_1 = T_2 = T/2$) with the amplitude of the second period reduced to illustrate shimmer. The sine functions for the first two Fourier coefficients are shown (B_1 and B_2).*

Fitting our 1/T1 (2/T) Hz waveform we see that the contributions to B_2 (2/T) are reduced (compare with fig.5.2) i.e. the first even harmonic component is reduced. All higher even harmonics (k×(2/T)) will still sum to zero. Now, considering 1/T, the contributions are not symmetrical but for the particular case of the sine wave this symmetry necessity is removed by the fact that there is a positive and negative contribution to the energy in each half of the cycle. All higher odd harmonic energy contributions sum to zero in similar fashion. The result is a sinusoid of reduced amplitude. It is interesting to note that the Fourier series (and equivalently, the Fourier transform) does not differentiate between two sinusoids with amplitudes of 1, and ½ respectively, that follow each other as in fig.5.3 and a sinusoid of constant amplitude ¾.

For waveforms that are non-symmetrical about the x-axis, energy contributions to higher harmonics will not cancel in the presence of shimmer. The glottal pulse model is examined in fig.5.4. The unperturbered waveform is examined first. Fitting our 1/T1 (2/T) Hz waveform we see that the contributions to B_2 (2/T) add constructively, giving amplitude of the 'fundamental', at the first even harmonic. All higher even harmonics (k×(2/T)) sum in similar fashion, giving spectral contributions dependent on the frequency characteristics of the waveshape - the glottal pulse having a low pass nature as illustrated in the caption in fig.5.4.

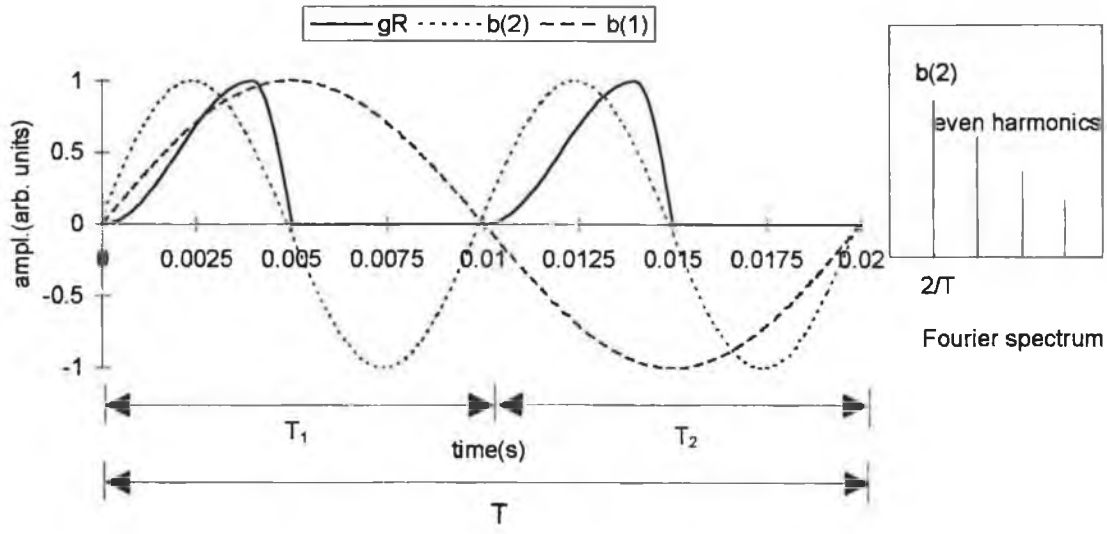


Fig.5.4 *Two periods of the Rosenberg glottal pulse (with T_1 = T_2 = T/2). The sine functions for the first two Fourier coefficients are shown (B_1 and B_2).*

Now, considering 1/T, the contributions are symmetrical but opposite and sum to zero. All higher odd harmonic energy contributions sum to zero in similar fashion. Shimmer is introduced as shown in fig.5.5. The contributions to B_n(2/T) are reduced due to the decreased amplitude of the second period of the waveform. All higher even harmonics are similarly reduced, and the reduction is in accordance with the spectral energy contributions for that frequency. The contributions to B_n(1/T) do not cancel due to the amplitude difference in the glottal waveform. A similar result occurs for all higher harmonics and again this is in accordance with the spectral energy contributions

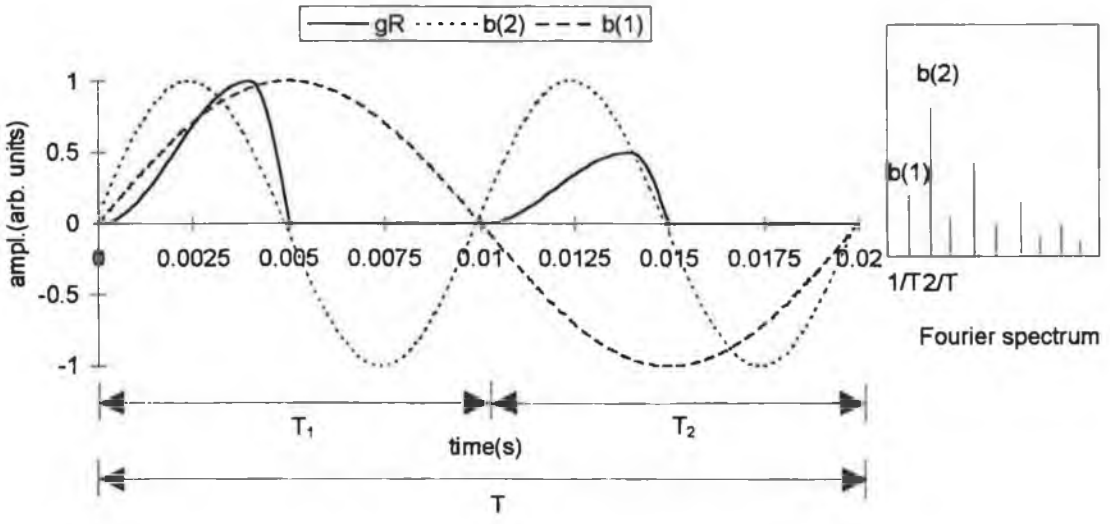


Fig.5.5 *Two periods of the Rosenberg glottal pulse (with $T_1 = T_2 = T/2$) with the amplitude of the second period reduced to illustrate shimmer. The sine functions for the first two Fourier coefficients are shown (B_1 and B_2).*

for that frequency. Because we cannot shift the waveform to be even or odd, the cosine terms of the series should also be considered. For the glottal pulse shown in fig.5.4 the same arguments that were used for the sine terms hold true for the cosine terms (fig.5.4). Shimmer was introduced with a reduction in the amplitude of the waveform in fig.5.5. This could just as easily have been an increase in amplitude giving rise to increased even harmonics and in the case of random shimmer we would expect the overall effect to sum to zero, leaving the amplitude of the even harmonics unperturbed. A similar argument can be put forward for the odd harmonic components. However, since this is a difference measurement we anticipate some variability, and that the variability increases as the variance of the shimmer signal increases.

In the case of jitter the analysis is somewhat different. In this instance the basis vectors are separated by 1/T′ Hz (fig.5.6). As the two periods are not equal, a contribution exists at 1/T′ Hz. Notice also that there is a reduction in the amplitude contributions at 2×1/T′ Hz as the basis vectors no longer match the "fundamental frequency", $1/T_1$. The situation is analogous to 'leakage' that occurs when a non-integer number of periods are present in the analysis frame[12]. This can also be viewed analytically by

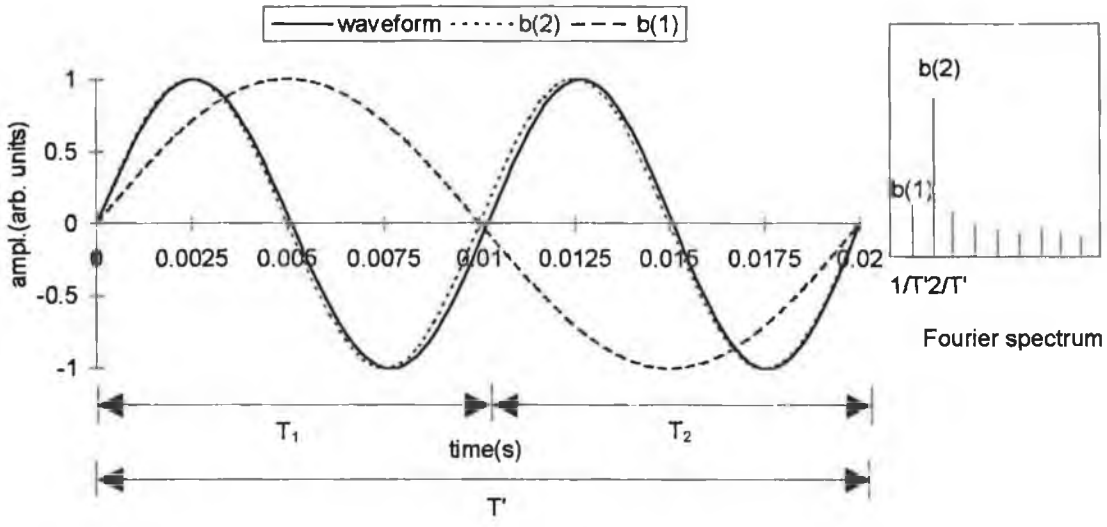


Fig.5.6 *Two periods of a sine wave (with $T_1 \neq T_2 \neq T/2$) illustrating jitter. The sine functions for the first two Fourier coefficients are shown (B_1 and B_2).*

substituting the sine waves of frequency $1/T_1$ and $1/T_2$ into eqtn.5.3 and noting that the calculation of the higher harmonic terms for B_n no longer integrate to zero. Hillenbrand observed that jitter leads to a more prominent smearing of the harmonic structure at higher frequencies. We see here that by virtue of the fact that the periods in question are not sub-multiples of the basis vectors, higher harmonics appear in the spectrum and in respect to the sine wave all of these are noise components. As the even harmonics (2k) increase, collecting estimates at even multiples of (1/T'), the difference to the actual harmonic locations of $2k\times1/T_1$ and $2k\times1/T_2$ increases, therefore reducing the contributions to higher even harmonics (1/T') as the frequency increases. However, at some upper frequency location we expect the even analysis harmonic $2k\times(1/T')$ to match or cross over the signal harmonic, $2k\times(1/T_1)$, therefore producing the reverse effect with $2k\times(1/T_1)$ gaining more of the energy contributions and $2k\times(1/T_2)$ obtaining less. The same process is simultaneously occurring for $k\times(1/T_2)$. Harmonic reinforcement will also occur when $k\times(1/T_1-1/T_2)$ matches either of the signal frequencies $1/T_1$ or $1/T_2$. Superimposed on this harmonic interplay between the analysis basis vectors and the signal frequencies, as the harmonic number increases, due to the frequency characteristics of the signal, the energy at higher harmonics decreases. Therefore the contributions at between harmonic locations are dependent on the jitter

artifact and signal characteristics with the signal harmonics ($1/T_1$ and $1/T_2$) becoming further separated at the upper partials with occasional reinforcement occurring when

$$k \times (\frac{1}{T_1} - \frac{1}{T_2}) = m \times \frac{1}{T_1} (\text{or } m \times \frac{1}{T_2}) \qquad \text{eqtn.5.10}$$

The odd analysis harmonics ($(2k+1)\times 1/T'$) also receive reinforcement at particular frequencies governed by

$$k \times (\frac{1}{T_1 + T_2} - \frac{1}{T_1}) = m \times \frac{1}{T_1} \qquad \text{eqtn.5.11}$$

Eqtn.5.11 can similarly be written for ($1/T_2$). Because the frequency differences in equation 5.11 may be large it may be more convenient to simply match the $1/(T_1+T_2)$ to the integer number on the right side of the equation. Also, equations for minima can be written by adding ½ to 'm' in the above equations. Note that in a two cycle analysis development it is impossible to differentiate cyclic and random jitter. In the case of cyclic jitter, the above mentioned 'odd analysis harmonic' is equivalent to the subharmonic frequency and the development is exactly as laid out above. For random jitter, the above mentioned trends are still valid but the random variability introduced will have a large bearing on the overall spectral characteristics.

Considering the addition of mean zero random noise (fig. 5.7), it can be seen that the basis vectors are correct but that the spectral estimates at even harmonics are more variable and energy appears at odd harmonic locations due to the random noise components. If the variance of the noise increases we would expect the variance of our spectral estimates to increase also. The sine wave signal plus noise may be represented by the following equation,

$$s(t) = \sin(\frac{2 \times 2\Pi t}{T}) + q(t) \qquad \text{eqtn.5.12}$$

where q(t) is the noise component

We have seen (eqtn5.4 to eqtn.5.9) that the Fourier series for the sine wave gives zero energy at all frequency locations except 2×1/T. Therefore for $n \neq 2$ we have

$$B_k = \frac{2}{T}\int_0^T q(t)\frac{\sin 2\Pi knt}{T}dt \qquad \text{eqtn.5.13}$$

Since q(t) is random it can be inferred that it provides energy contributions to all harmonics in the spectrum. In fact for truly random noise the integral in eqtn.5.11 cannot be evaluated and the autocorrelation of q(t) must be evaluated prior to integration[13]. Fourier estimates of noise are dealt with in more detail in section 5.3.4. Our development, simply through a direct implementation of the Fourier series and a brief reference to the Fourier transform has led to the above spectral characterisations of shimmer, jitter and additive noise respectively.

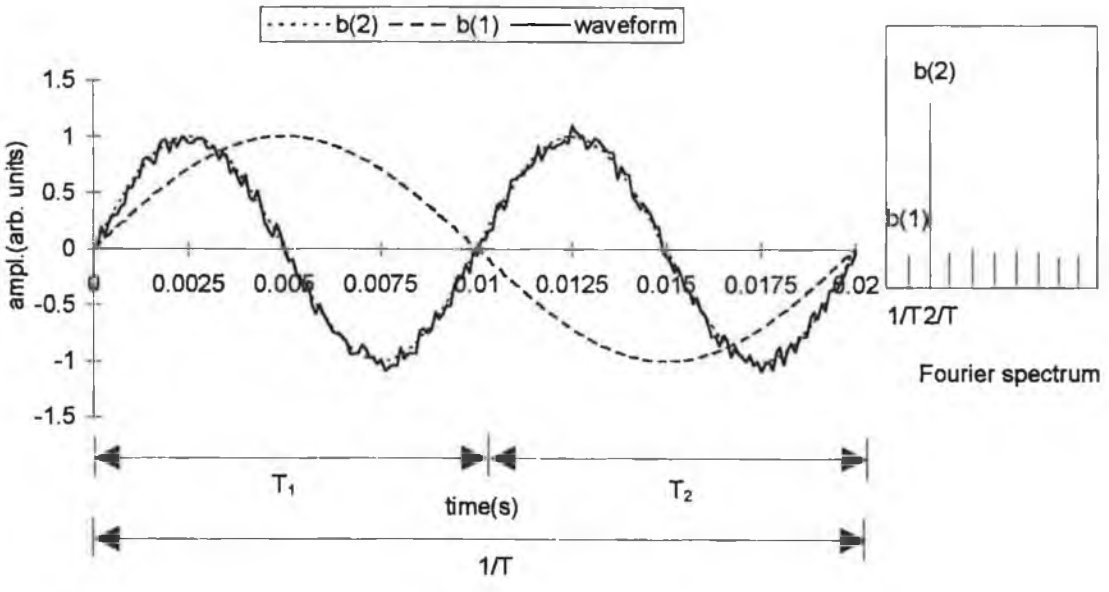


Fig.5.7 *Two periods of a sine wave (with $T_1 = T_2 = T/2$) in the presence of additive noise. The sine functions for the first two Fourier coefficients are shown (B_1 and B_2).*

In summary, for shimmer signals the level of the even harmonics is unchanged and the level of odd harmonics increases with the variance of the shimmer and the contributions to the noise at a given frequency are in direct relation to the contributions to the signal

at that frequency. Jitter gives reduced harmonic levels, a broadening or segmentation of spectral harmonics that increases with frequency and energy is also introduced between harmonics. Again, the contributions at a given between harmonic location are dependent on the frequency characteristics of the unperturbered signal. Finally, additive noise causes the energy at harmonic locations to be more variable and energy is introduced between harmonics with flat spectral characteristics. Therefore both the jitter and shimmer signals are dependent on the characteristics of the signal being perturbed and the additive noise is independent of the signal characteristics.

Our conclusions are in agreement with the jitter and shimmer results reported by Klingholtz et al[14] and with the additive noise and jitter observations reported by Hillenbrand. In the study undertaken by Klingholtz the harmonic level estimated in shimmer was found to remain constant, whereas Hillenbrand[1] found the harmonic levels to be significantly reduced with respect to jitter and additive noise. Our development supports the former observation. (Hillenbrand's report may have been a result of the particular synthesis used, where there appear to be unusually high noise levels at the formant locations in the spectra that he illustrated).

To test our hypotheses spectra for the glottal pulses with 6 levels of additive noise, jitter (2 types) and shimmer were examined. Some typical results are shown. A program was written (psha2.m) to carry out the two cycle Fourier series analysis. Another program (paha.m) provided periodogram estimates (paha.m), which are based on averaged Fourier transforms of longer time records. This therefore provided a second means of analysing the test signals. The program details are given in sections 5.3.7 and 5.3.4 respectively.

Figure 5.8 illustrates the two cycle Fourier series analysis for shimmer values of std. dev. 2% and 32 %. The Fourier series coefficients are computed every two cycles for an analysis interval of 1 second and the mean of the Fourier coefficients are plotted. It can be seen that increases in shimmer cause the noise floor to go up in a consistent manner for all frequencies (i.e. in accordance with the signal characteristic). The harmonic levels themselves remain unperturbed and the source spectrum envelope is maintained. The noise component has reached a higher level for the std. dev. 32% shimmer signal. Following from the two cycle development and calculating the energy (coefficients are squared prior to averaging) every two periods, this is the expected

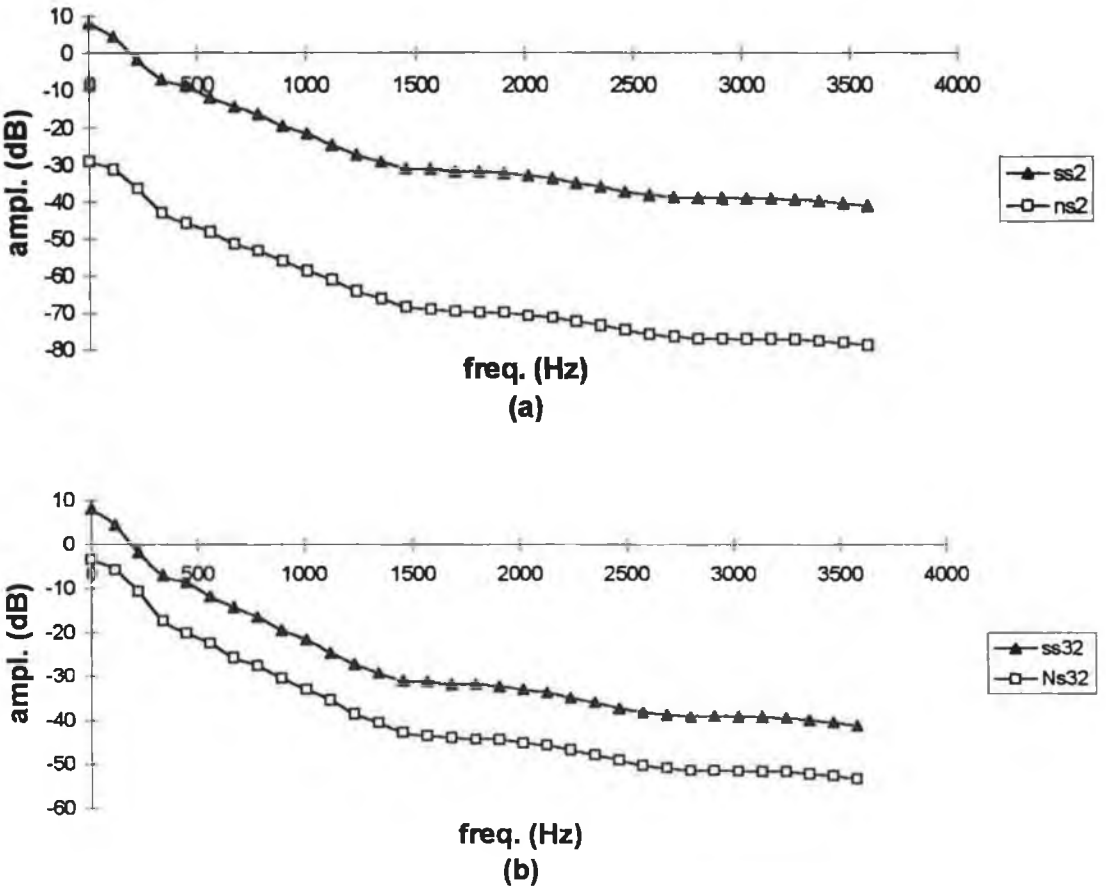


fig.5.8 *Two cycle Fourier series coefficients for glottal pulse signals with (a) 2% std. dev. random shimmer and (b) 32% std. dev. random shimmer*

result, with the harmonic levels on the average remaining unperturbed and energy introduced between harmonics which increases with increasing shimmer. In considering the periodogram estimate (Fourier amplitude is gathered over many cycles before calculating the energy) of the same signal (fig.5.8(b)), we again expect the harmonic variation to sum to zero as shown. A similar consideration of the between harmonics might lead one to conclude that a summing to zero also occurs here since the shimmer signal has a mean of zero. However, the periodogram graphs show exactly the same trend as the two cycle analysis plots. The periodogram plots are an average of six 4096 point spectra hopped 1024 points. This averaging has reduced the variance of the spectral estimates (section5.3.). Figure 5.10 shows the variance associated with a single Fourier transform spectrum of 4096 points for the signals with std. dev. 2% and 32% shimmer. The increased variance of the between harmonic

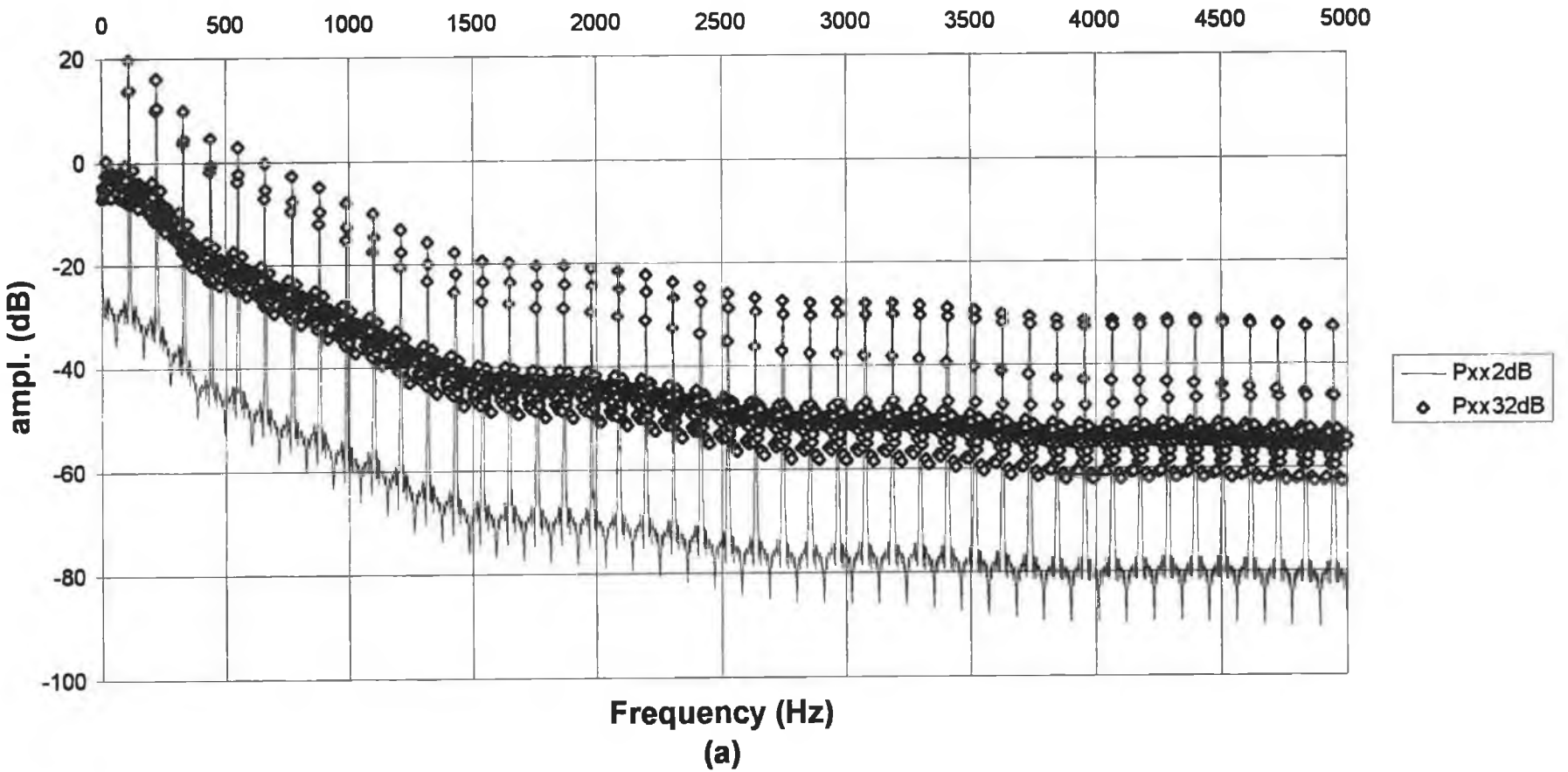


fig.5.9 (a) *Periodogram estimate or power spectral density (PPx2dB,PPx32dB) for the test shimmer signals of std. dev. 2% and 32 %.*

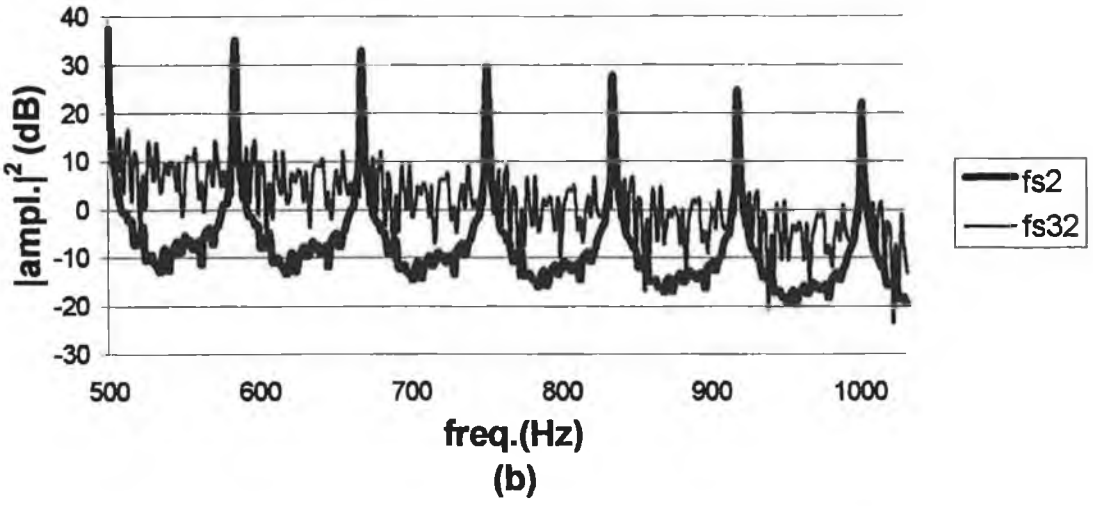


Fig.5.9 (b) *Fourier transform estimates (fs2 and fs32) of the same signals illustrating the increased variance of the std. dev. 32% signal's between harmonic energy.*

estimates is now very obvious. Therefore as the transform gathers it's frequency estimates the variance of the estimates increases according as the shimmer signal increases. This variance is still symmetrical about zero, however the square of the variance is not, therefore raising the noise level as shown. The variance is also present at harmonic peak locations but due to the fact that the signal is much larger than the variance the effect is very small when the squaring operation is applied in order to obtain the energy. Note the variance in question here is linearly related to the source variance but that it is of smaller magnitude.

For cyclic jitter (fig.5.10 (a)) we see that the main characteristic is that a strong subharmonic has been introduced at an octave lower than f0, as might have been expected. The amplitude in the subharmonic spectrum can be seen to follow an interesting trend, which is governed by eqtn.5.11. Substituting values gives matches the 23rd '106 Hz signal component' with the 50th (signal-subharmonic). This also occurs for the '109.9 Hz signal component' at about this frequency.

For random jitter (fig5.10 (b)) the spectrum is somewhat different. The harmonic structure is severely affected for the std. dev. 6% random jitter signal shown in the two cycle Fourier series spectrum. The spectral envelope is maintained, however, with

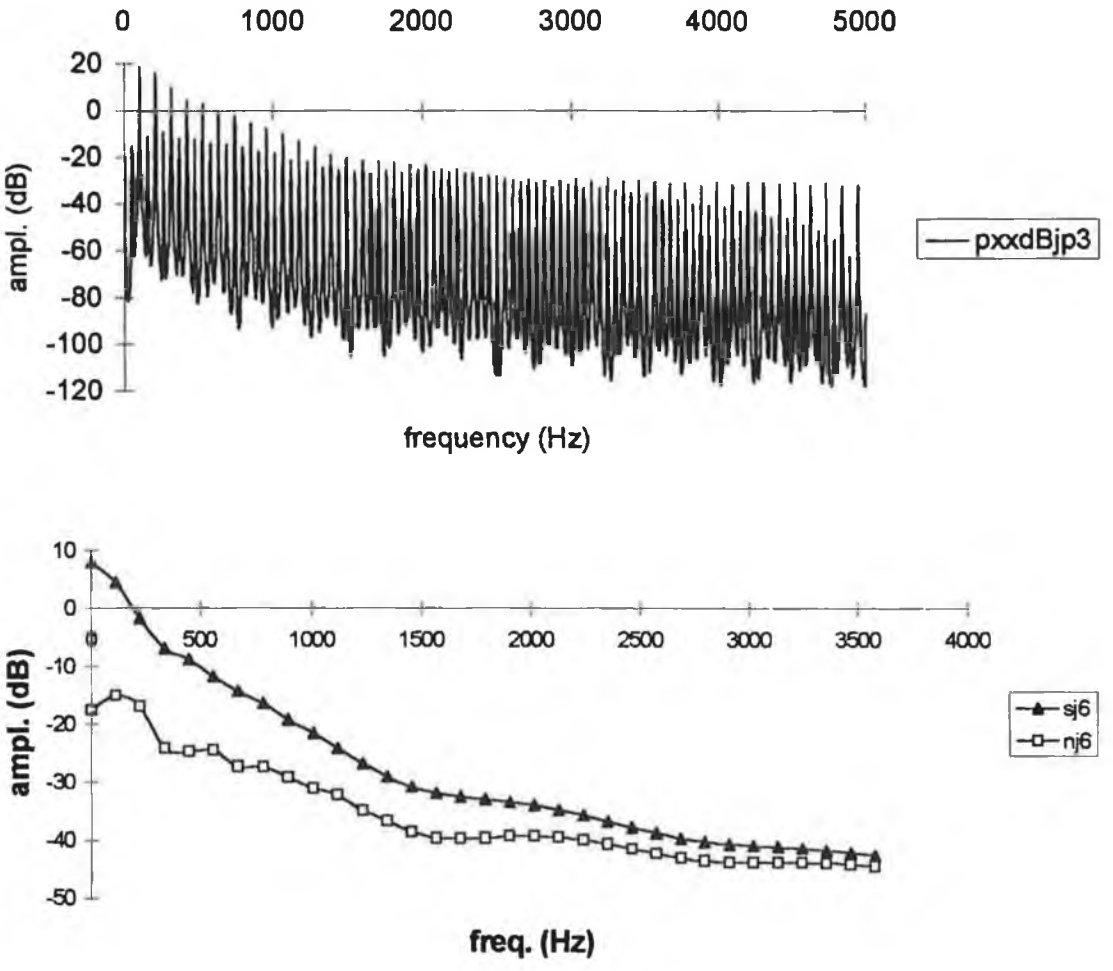
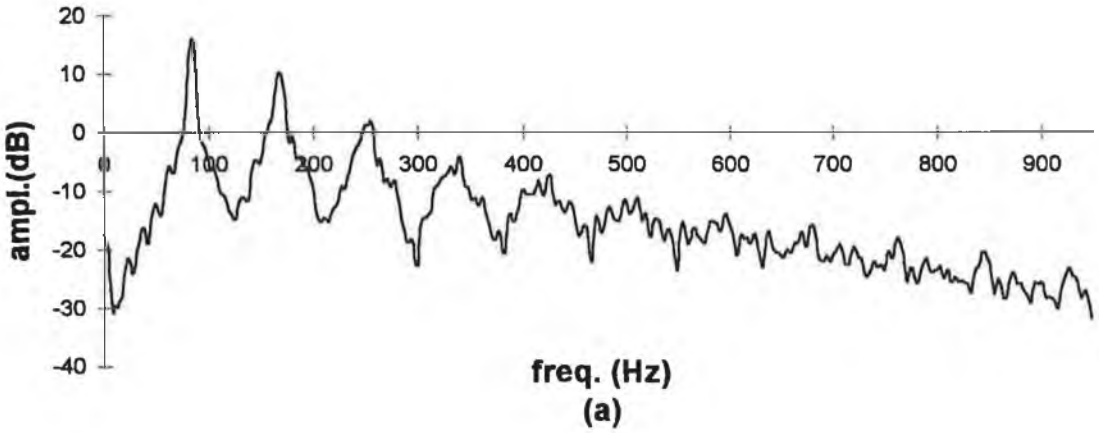
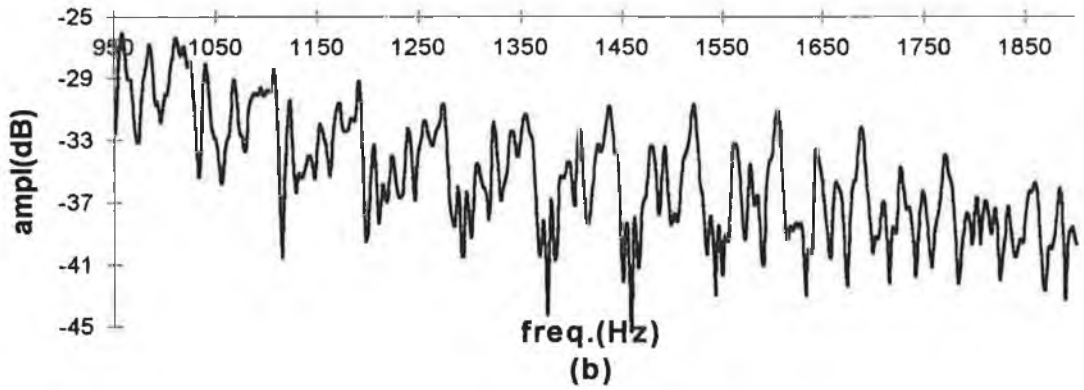


fig.5.10 (a) *Periodogram estimate for 3 % cyclic jitter of source signal* (b) *Two cycle Fourier Series averaged energy spectrum for random jitter signal with 6% std. dev.*

noise energy i.e. non harmonic energy, becoming nearer in magnitude to the harmonic energy as the frequency increases.

The periodogram provides further information. For a perfectly periodic signal the estimate at a given frequency location is the convolution of the Fourier transform of the window function with the Fourier transform of the signal at that frequency. This is shown in fig.5.11 for the 110 Hz glottal waveform with a random jitter component of std. dev. 6%. As the frequency increases the spread of the signal about the higher harmonics increases, reducing the amplitude at k×110 Hz locations. Cross over effects





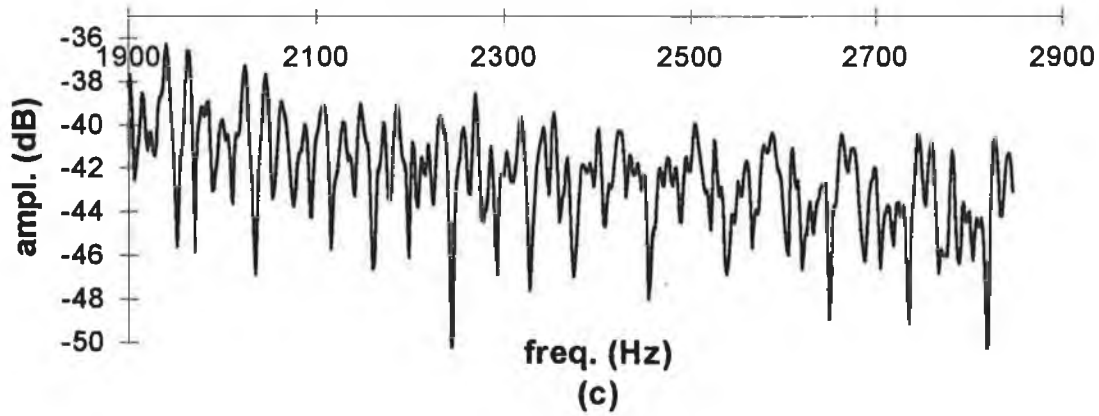


fig.5.11 *Periodogram estimates of the random jitter signal (a) 0-900Hz showing spectral broadening and harmonic decomposition, (b) in general harmonic structure is diminished with a reappearance in harmonic structure at ~1500Hz (c) as for (b) with harmonic reappearance ~2650 Hz*

also occur. Hence there is a broadening of the spectral peak due to the window function. For the higher harmonics the difference between the two frequency contributions increases causing further broadening of the spectral peaks. However if the jitter is large the contributions may become individually resolved as the harmonic number increases giving rise to a more irregular looking spectrum. However, some reappearance of harmonic structure is also evidenced due to the cross over in jittered frequencies as mentioned earlier (5.11(b)) (5.11(c)).

Figure 5.12 shows the effect of adding random Gaussian noise to the glottal source. The level of the harmonics themselves are unaffected except as the noise floor moves upwards, in a sense consuming the lower level harmonics as it moves. The noise spectrum is white as expected. The periodogram reveals no extra information (not shown).

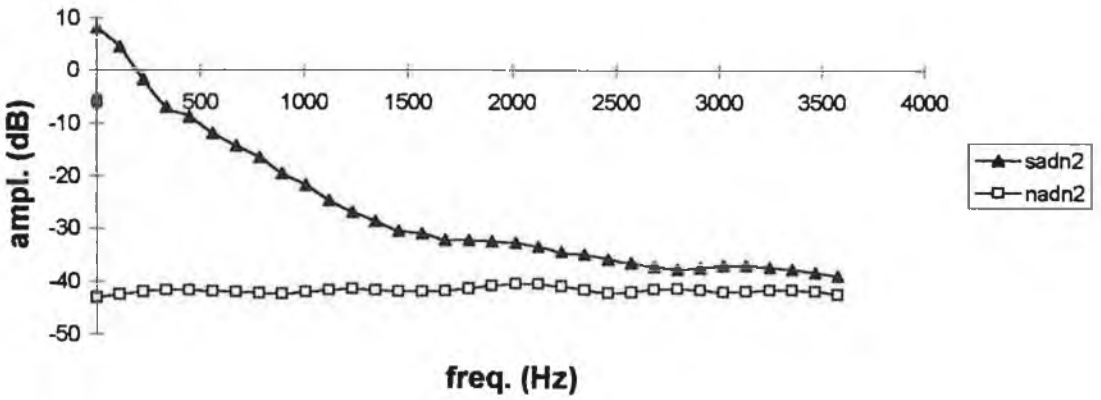


fig.5.12 *Two cycle Fourier Series averaged energy spectrum for random additive noise of std. dev. 2 %.*

So, considering the four cases of cyclic and random jitter, shimmer and additive noise we see that there are clear spectral differences. With this in mind it may therefore be possible to develop quantitative spectral measures that differentiate the four types. For example, a constant value of H/N(ω) for all ω would indicate shimmer. A first examination of the random jitter and additive noise graphs indicates that they are somewhat similar with early harmonic structure still present and higher harmonics completely missing.

ACOUSTIC INDEX /PERTURBATION TYPE	HARMONIC LEVEL (H)	NOISE LEVEL (N)	HARMONIC TO NOISE RATIO (H/N)
Shimmer	Constant	Increases in direct proportion to signal characteristics at that frequency.	Constant for all frequencies
Random Jitter	Reduced	Increases-signal dependent. Amplitude of noise is greater at lower frequencies.	Reduced. Decreases with increasing frequency.
Cyclic Jitter	Variable - dependent on individual pitch periods	Increased- a subharmonic regime is introduced.	Reduced-dependent on pitch period relationship
Random Additive Noise	More variable	Increased independent of signal	Reduced- decreases with increasing frequency

Table5.1 *Summary of Spectral Characteristics of Shimmer, Random Jitter, Cyclic Jitter and Additive Noise*

However, closer examination reveals that the noise level in the additive noise case is considerably higher and therefore a measurement of H/N ratio from 1 to 4 kHz indicates a perturbation of either additive noise or jitter but a subsequent measurement of the noise level will reveal which of the two is actually present. The subharmonic regime of the cyclic jitter spectra suggest a method for quantifying this type of perturbation. A summary of the spectral characteristics of the four perturbation

measures are given in Table5.1. To surmise, we have hypothesised what the spectral characteristics of jitter, shimmer and additive noise are and then produced spectra from the synthesized data in order to examine our hypotheses, which were found to be in good agreement. Therefore, what is now required is some definite spectral measurements to objectively prove the above assumptions in a quantitative manner. Section 5.3 provides a detailed description of the quantitative analysis techniques that have been developed in order to make these spectral calculations.

5.2.3 Harmonic to Noise Ratio of the Glottal Source and it's Relation to the Harmonic to Noise Ratio of the Output Radiated Speech Waveform

In acoustic analysis of tape recorded speech, the acoustic speech waveform is a complex signal consisting of the source excitation convolved with the vocal tract filter function, followed by radiation at the lips. The objective is to take measurements from this output signal and make inferences regarding the source signal and hence the underlying vibratory mechanism.

From the source/filter model of speech production we have that

$$s(t) = e(t) * v(t) \qquad \text{eqtn.5.14}$$

$$s(t) = (e(t) + n(t)) * v(t) \qquad \text{eqtn.5.15}$$

$$S(\omega) = \left[E(\omega) + N(\omega)\right] \times V(\omega) = E(\omega) \times V(\omega) + N(\omega) \times V(\omega) \qquad \text{eqtn.5.16}$$

$s(t)$, $S(\omega)$ = output waveform and Fourier transform

$e(t)$, $E(\omega)$ = source signal and Fourier transform

$n(t)$, $N(\omega)$ = additive noise and Fourier transform

$v(t)$, $V(\omega)$ = impulse and frequency response of the vocal tract

where * indicates convolution.

As discussed in section 5.1 the harmonic to noise ratio (H/N) taken from the output speech waveform or spectrum is a commonly used measure in assessing vocal pathology. It is pertinent to ask, in what sense is the harmonic to noise ratio of the speech waveform indicative of the harmonic to noise ratio of the source signal. Rearranging 5.7 gives

$$\frac{H}{N}(\omega) = \frac{E(\omega) \times V(\omega)}{N(\omega) \times V(\omega)} \qquad \text{eqtn.5.17}$$

which is the H/N ratio at frequency 'ω'. Within the limits of the source filter model, the H/N ratio at a given frequency location is the same for source and filtered waveform. Therefore we can write

$$\frac{H}{N}(\omega)(\text{waveform}) = \frac{H}{N}(\omega)(\text{source}) \qquad \text{eqtn.5.18}$$

However, the harmonic to noise ratio is usually determined over the complete frequency range or over a band limited region but not at discrete frequencies. Therefore eqtn.5.15 is summed in order to obtain the overall harmonic to noise ratio. The ratio is generally expressed in dBs.

$$\frac{H}{N}(\text{waveform}) = 10 \times \log 10 \left[\frac{\sum_{\omega} E(\omega)V(\omega)}{\sum_{\omega} N(\omega)V(\omega)} \right] \qquad \text{eqtn.5.19}$$

A consideration of this case shows

$$\frac{H}{N}(\text{waveform}) \neq \frac{H}{N}(\text{source}) \qquad \text{eqtn.5.20}$$

eqtn.5.19 is a generic formulation of the harmonic to noise ratio as calculated by various investigators. Variations include inverting the ratio, giving the noise to signal

ratio or giving the signal to noise ratio by including the noise and harmonics as signal. Bandlimiting the range over which the ratio has been calculated has also been investigated. Despite the fact that several variations of eqtn.5.17 that have been implemented, no study has attempted to relate the S/N (out) to the S/N (source). A simple numerical example will help illustrate the problem of simply using eqtn.5.17 to make inferences regarding source characteristics.

Example

As a simple illustration of eqtn.5.17 we take two frequency values, one low (ω_L) and one high (ω_H) and consider some numerical values.

$E(\omega_L) = 1000,\ E(\omega_H) = 10$

$N(\omega_L) = N(\omega_H) = 1$

$H(\omega_L) = 10$

$H(\omega_H) = 100$

$$\frac{H}{N}(\text{source}) = \frac{1000+10}{1+1} = \frac{1010}{2} = 505 = 27\text{dB}$$

$$\frac{H}{N}(\text{output}) = \frac{1000\times10+10\times100}{10+100} = \frac{10100}{110} = 20\text{dB}$$

Therefore, H/N (source) ≠ H/N(output) and the high frequency component has gone from having very little effect to dominating the ratio.

Taking an alternative approach allows us to recover the H/N ratio of the source.

$$\frac{H}{N}(\text{source}) = \frac{1}{M}\sum_{w=0}^{M}\frac{E(w)H(w)}{N(w)H(w)} \qquad \text{eqtn.5.21}$$

This ratio, which we indicate by H/N_S, reflects the H/N ratio of the source in a true sense only when the noise is equal for all frequencies. This is so for truly mean zero,

Gaussian noise. But the ratio can be of use even when this is not the case. As stated in the introduction, a prime objective is to make measurements that will correlate with either the physical underlying processes of vocal fold vibration or with perceptual based measures. Other researchers have made similar remarks:

"We also need to consider the purpose of our objective analyses. One analytic goal may be to determine the conditions of the vocal folds in order to study the pathophysiological mechanisms of voice disorders. In this case acoustic characteristics irrelevant to perception may be important. If the goal of the analysis is to predict the perception of the voice quality, then we have to consider features of our hearing mechanisms such as masking, frequency sensitivity, and so on." (Gauffin et al[15]).

It can be seen that 5.18 attempts to match physical attributes, in order to obtain information regarding the glottal source. In addressing perceptual correlations, a ratio which we have termed the geometric dB mean (eqtn.5.19) has been developed.

$$\frac{H}{N}(\text{geometric}) = \frac{1}{M}\sum_{\omega=0}^{M} 10 \times \log_{10}\left(\frac{H(\omega)E(\omega)}{H(\omega)N(\omega)}\right) \qquad \text{eqtn.5.22}$$

The term geometric mean comes form the fact that the additions involve logarithms and is therefore somewhat equivalent to taking the product of the linear values and taking the M^{th} root i.e. the geometric mean. This effectively gives greater weight to the higher frequency components in a crude manner at matching the frequency analysis processes of the ear. Extracting the ratio from the dB signal probably gives too much weight to the higher frequency components[16] and perceptually based frequency weighted ratios[17], having similar form to the above ratio, have been developed for vocal quality assessment. To date, these ratios have not been applied to the investigation of pathological voice types.

5.2.4 Harmonic Intensity Level

The preceding section examined the harmonic to noise ratio and suggested variations that may be of use in studying both source and perceptual characteristics. Previous to this, section 5.2 outlined possible causes of 'noise' found in pathological voice types.

Therefore, we have examined the ratio itself and the denominator but have somewhat neglected the numerator i.e. the energy at harmonic locations. The harmonic energy is perhaps the most important of the three. Recent developments have tried to develop a frequency domain parameter set relating to the LF model of glottal flow (fig.5.13) developed by Fant and co-workers[18]. The level of the harmonics is of prime importance in these developments. Often, in spectral analysis, the level of the harmonics is often overlooked, possibly because the spectrum is usually given in dB and relative values are often of more immediate interest. In it's most basic form, the level of harmonics provides an additional parameter for investigating the waveform. However, taking particular ratios between specific harmonics provides a means of assessing different flow characteristics and hence information can be obtained regarding the vocal fold vibrations.

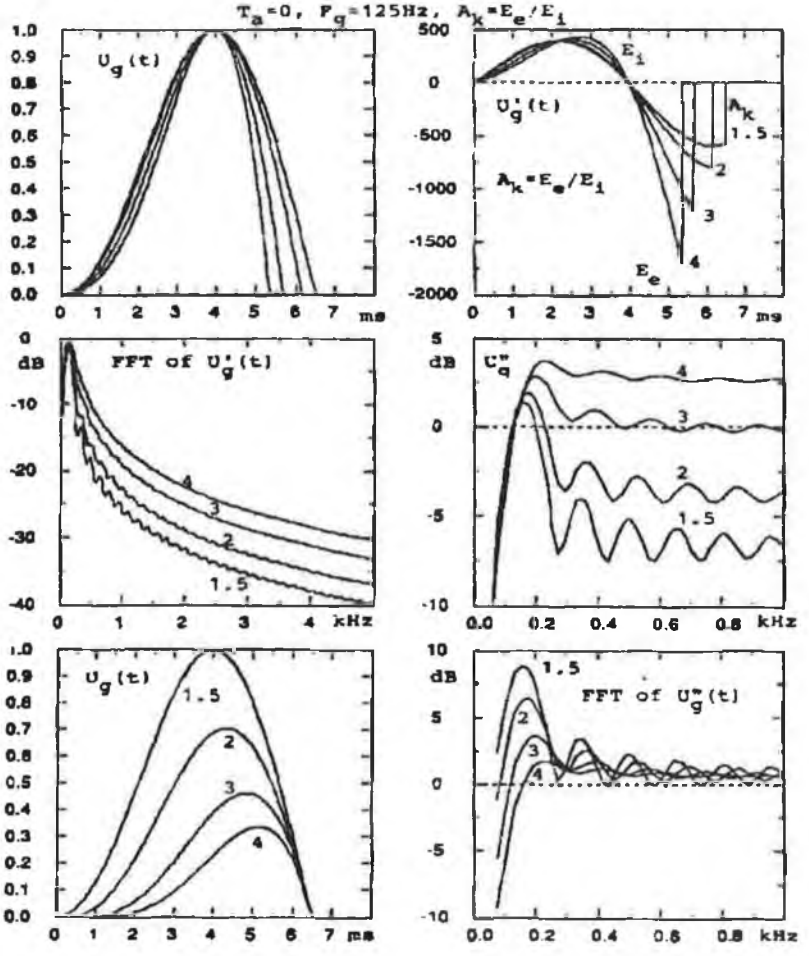


fig.5.13 *LF model of glottal flow. The top two rows: LF-flow $U_g(t)$, flow derivative $U_g'(t)$, and spectrum of U_g'' (t) at varying E_e/E_i, constant U_0, and T_a = 0. In the bottom row: LF flow and spectrum $U_g''(t)$ when maintaining constant E_e (T_a=0).*

The four parameters of the model are R_k, R_g, R_a and E_e, which are related to the three basic time events: (1) the location of the flow peak T_p; (2) the discontinuity point T_e at glottal closure and (3) the duration of the return phase T_a. Varying the model parameters produces changes in the frequency spectra as shown in fig.5.13 and an equivalent frequency domain parameter set has been derived. One important advantage of a frequency domain parameter set of glottal flow is that it can be used in conjunction with tape recorded speech (the time domain model requires accurate determination of low frequency phase). The use of such a frequency domain based glottal flow parameter set should be of considerable use for investigating vocal pathologies. However, we have seen in the previous section how the perturbation measures of shimmer, jitter and additive noise disrupt the harmonic structure in the spectrum and therefore destroying the possibility of applying the above mentioned frequency domain parameter set. In the next section techniques are developed in order to overcome these problems.

5.3 Analysis Techniques:

Following, is an account of nine techniques (Table5.2) which have been developed in order to provide some form of harmonic to noise ratio for investigating pathological voices. In presenting the techniques, any deviations from their original design are clearly stated and the reasons for the changes are detailed explicitly. The order is given, insofar as possible, to facilitate a continuity of ideas as opposed to listing the methods in chronological order. In this way we can begin to appreciate the difficulties that are encountered in obtaining the harmonic to noise ratio. The development leads to three novel approaches, one of which addresses the problems mentioned in 5.2. regarding the separability of jitter, shimmer and additive noise. All of the programs were coded in the Matlab programming language. This section deals expressly with the methodology and a full description of the results is given in the section 5.4.

ANALYSIS TECHNIQUE (PROGRAM NAME)	DESCRIPTION	ANALYSIS LENGTH (SHORTEST UNIT)
kitnos3.m	noise reducing filter	205ms
harmony4.m	Spectrum Analysis	205ms
harmper2.m*	Periodogram Analysis	205ms
noise6.m	Spectral Analysis	seven periods
harm4.m	Spectral Analysis	four periods
kojnos3.m	Fourier Analysis	three periods
psha2.m*	Fourier Analysis	two periods
harmyum.m	Time Domain	one period
psha1.m*	Fourier Analysis	one period

Table.5.2 *List of nine analysis techniques implemented in the present study where * indicates a novel procedure. Source code is given in appendix A.*

5.3.1 Noise Reducing Filter - Kitajima[19]

As stated by Kitajima "the basic idea is to pass the voice through a noise reducing filter at first, and then compare the pre- and post- filtered voices in their effective values". A stable portion of the vowel a/ is taken, sampled at 5 kHz (cut off frequency 2.5 kHz) for 205 ms and padded up 4096 samples. The approach is to obtain the power spectrum of the signal, apply a moving average filter to this harmonic spectrum and count as signal the components above the moving average estimate, and as noise those components below the moving average filter. This situation is depicted in fig.5.14 and given in equation form as

$$|E(f)|^2 = |Z(f)|^2 > |Z(f)|^2 - S(f) \qquad \text{if } |Z(f)|^2 > S(f)$$

$$|E(f)|^2 = 0 \qquad \text{if } |Z(f)|^2 \leq S(f)$$

eqtn.5.19

$$|E(f)|^2 = |Z(f)|^2 - [S(f)+C(f)] \quad \text{if } |Z(f)|^2 > S(f) + C(f)$$

eqtn.5.20

where is Z(f) is the spectrum of the original signal
S(f)=moving averaged spectrum
E(f)= noise reduced spectrum
C(f) = standard deviation of S(f)

A noise reducing filter is then developed based on 5.19 and 5.20. The filter is given as

$$H(f) = E(f)/Z(f)$$

eqtn.5.21

Z(f) is already known and E(f) is obtained as above and therefore H(f) could be calculated for each subject. A voice signal is then passed through the filter H(f) and the filtered voice was designated as Y(f). This is shown schematically in fig.5.15. The

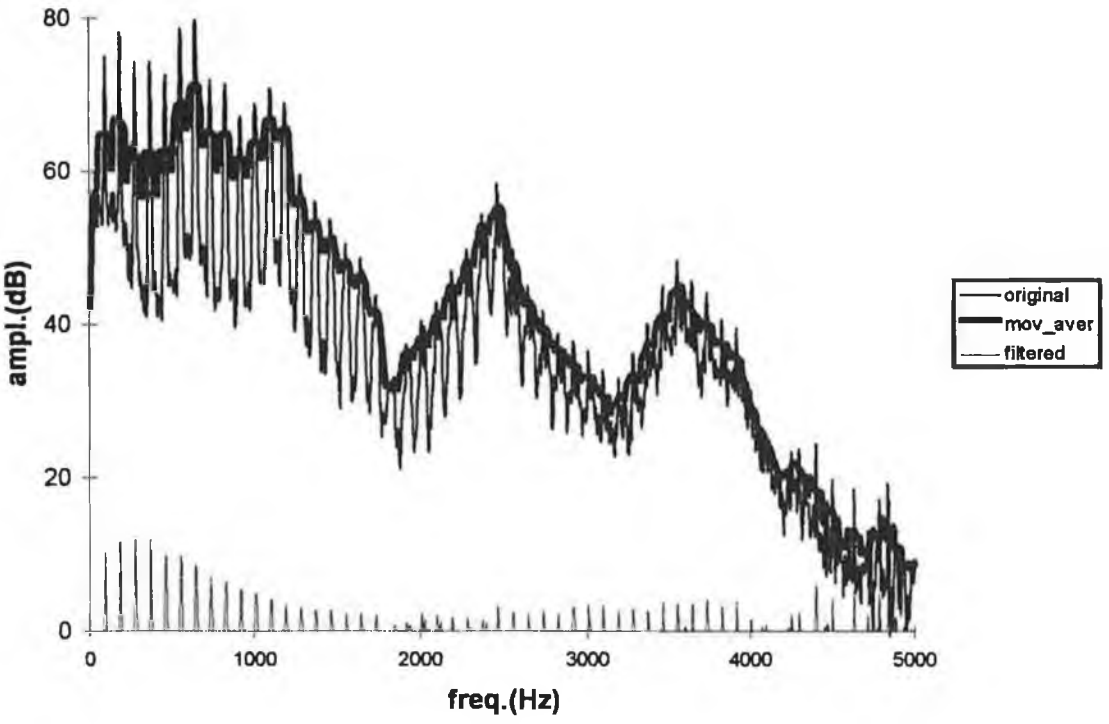


fig.5.14 *Moving average filter S(f), applied to speech spectrum Z(f), resulting in the 'filtered' spectrum E(f) (lower spectrum).*

residue signal of the filtering is then given as N(f) = |Z(f)-Y(f)|. The noise ratio is then calculated as

Noise ratio = rms of Y(f)/rms of N(f) eqtn.5.22

Some simple adjustments in technique help provide a more efficient harmonic to noise ratio estimate. Problems regarding the above implementation were a result, in part of the hardware limitations, which only allowed for 0.2 s of speech in the analysis frame. This was further reduced by a misguided assessment of the spectral analysis ..."taking the side lobes of the FFT into consideration, the beginning and end of the signal were not used. The mid-section , that is, one fourth of 205 ms of Z(f) and N(f), was applied in the formula." A quarter of the signal had been needlessly disregarded. The approach can be used to provide a convenient estimate of the harmonic to noise levels but the 'filter' is best not thought of as a noise reducing filter. Once the estimates have

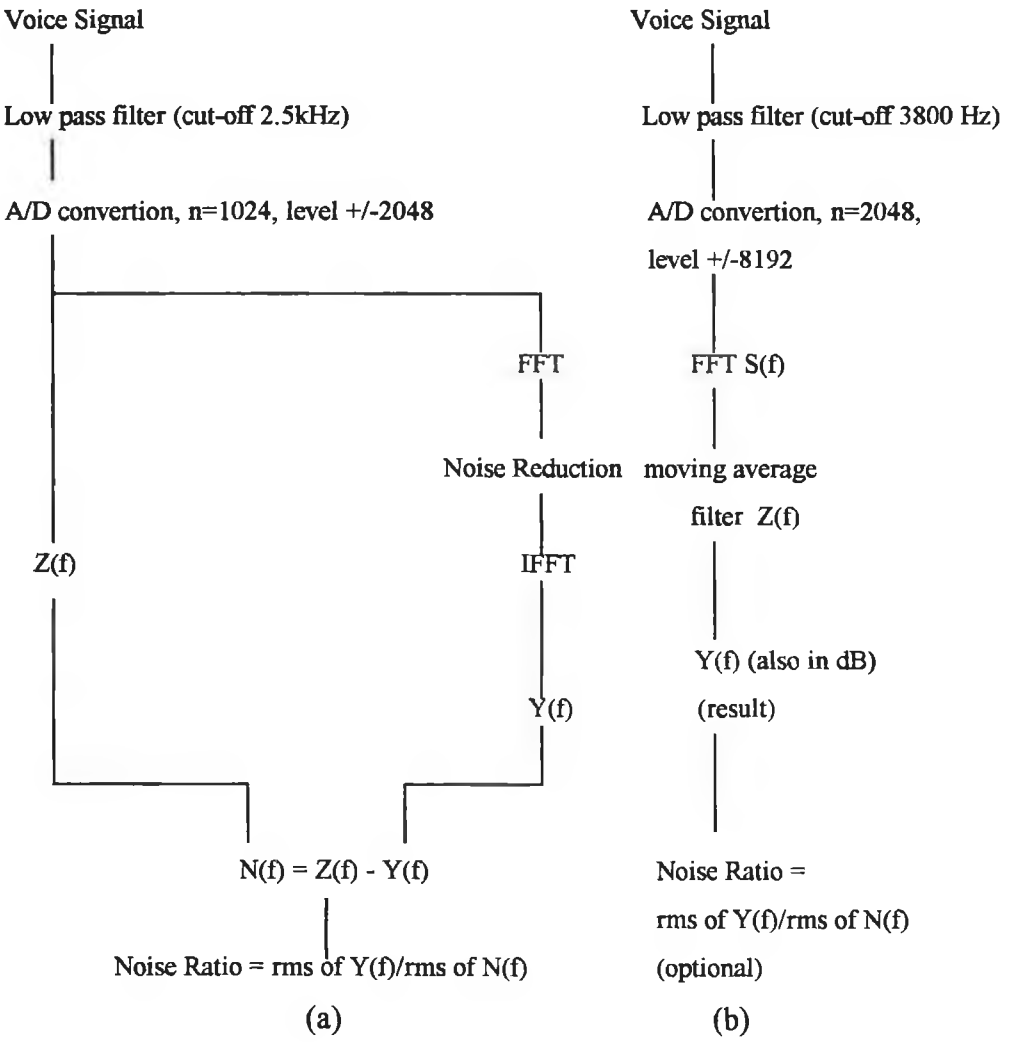


fig.5.15 (a) *Schematic diagram of analysis steps involved for the noise reducing filter as implemented by Kitajima and* (b) *as implemented in the present study*

been taken from the spectrum (fig.5.14), the harmonic to noise ratio can be directly inferred. Therefore, subsequently developing a filter based on these measurements is probably not necessary. The filter gives no extra information regarding the signal or noise being obtained. Another simple, but important modification is to average 'n' spectral samples for the moving average filter, where 'n' is dependent on the pitch period and not static (41 points).

Referring to the schematic diagram of the method as implemented by Kitajima, the many extra steps that were necessary for the filtering, involving a forward and inverse FFT can be seen. The filtering simply provides a noise estimate via a multiplicative

process based on the noise estimate that was originally obtained through an additive process. The method implemented here (fig.5.15(b)), simply applied a moving average (dependent on the pitch period) filter to the spectrum. Then the resulting signal plus it's standard deviation were subtracted from the original spectrum to obtain the noise reduced signal. Two separate estimates were taken, one based on the linear spectrum (kitnos3.m) and the other was obtained from calculations taken directly from the dB spectrum (kitnosdB.m). The process of calculation is best illustrated by referring to fig.5.14. In the case of the dB spectrum the average is calculated from the spectrum in dBs, giving what might be termed a geometric mean of the amplitudes. The linear ratio returned form 'kitnos3.m' is calculated as shown in eqtn.5.22 and then converted to dB which is the more usual form and hence more useful for comparison purposes. Also, 10kHz sampling was used in this study, and frequencies up to 3.8 kHz were analysed.

5.3.3 Relative Harmonic Intensity - Hiraoka et al[20]

The relative harmonic intensity (Hr) is a direct measurement taken from the Fourier magnitude spectrum in linear scale and is given by

$$H_r = \left(\frac{\sum_{i \geq 2} p_i}{P} \right) \times 100(\%) \qquad \text{eqtn.5.23}$$

The vowel a/ was used for analysis, sampled at 20 kHz. A 4096 point FFT was taken, corresponding to approximately a 0.2 second segment of speech at this sampling frequency. In order to match this condition with 10 kHz sampling, 0.2 second of the speech sample was extracted for analysis and padded out to 4096 points for Fourier transform analysis. The method given[20] for locating f0 was originally used, though it was found preferable to use any accurate f0 extraction method to locate the fundamental peak and then find the subsequent harmonic locations by searching for peaks in the region of n×f0±ml, where 'ml' represents the main lobe width of the analysis window. Surprisingly, the method used for calculating the amplitude of the i^{th}

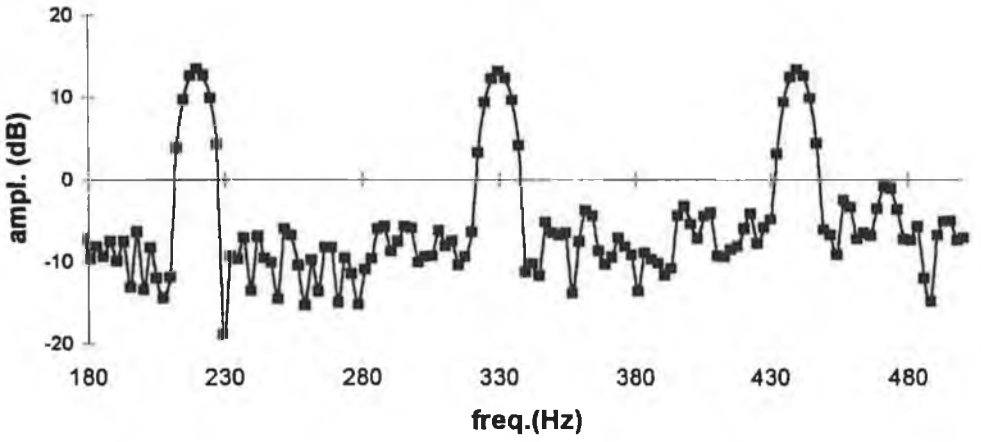


fig.5.16 *Section of 4096 point FFT power spectrum of 2.048ms of speech giving a mainlobe width of 8×pad/2/M=8×4096/2/2048=8 points.*

harmonic is not given although it is implicit that the peak amplitude at the harmonic location was taken from the linear spectrum. In this implementation, the total energy of the main lobe was taken as the energy for the i^{th} harmonic and taking the sum for all harmonics gave total harmonic intensity. The main lobe width[21] (fig.5.16) for the Hamming window used is given by '8×pad/2/M' where pad is the FFT length and M is the sample length taken. Several other noise indices were taken from the spectrum to give a total of ten different measures reflecting S/N, H/N, Hr, Sr and HNgeo and different regions of the spectrum were investigated.

5.3.3 Periodogram Averaged Harmonic Analysis (PAHA)

In considering the estimation of noise levels in pathological voices we have generally referred to the noise as being an additive random component. In that sense we are making an a priori assumption that there exists an underlying deterministic process which has been obscured or contaminated through the addition of random noise. However, we can also view the system under investigation to be the result of a stochastic process giving rise to a stationary random signal and based on this viewpoint make inferences about the underlying structure, if any, through the

application of statistical analysis tools. Such an analysis tool is the periodogram estimate or power spectral density function. It has been shown, however, that the periodogram does not provide a consistent estimate as the window length increases and that the variance of the estimate is of the same size as the power spectrum estimate under investigation[22] (Consider for comparison, the mean of a stationary random process which approaches the true mean as the window length increases). However, the variance can be reduced if we make several consecutive estimates of the signal i.e. N estimates reduces the variance by 1/N. Welsh[23] has shown that overlapping by 2:1 and hence increasing the number of estimates by a factor of two reduces the variance further, by almost a factor of two also. In terms of the power spectrum, the expected value of the average periodogram estimate is the convolution of the true power spectrum with the Fourier transform of the autocorrelation of the window function[21]. The spectral consequence of the autocorrelation is to double the mainlobe width[22] (fig.5.16). For a rectangular window this gives

$$c_{ww} = \begin{cases} L-|m|, & |m| \le (L-1), \\ 0, & \text{otherwise} \end{cases} \qquad \text{eqtn.5.24}$$

$$C_{ww} = \left(\frac{\sin(\omega L/2)}{\sin(\omega/2)} \right)^2 \qquad \text{eqtn.5.25}$$

$$P_{xx} = \frac{1}{2\pi LU} \int_{-\pi}^{\pi} F_{xx}(\phi) C_{ww}(e^{j(\omega-\phi)}) d\phi \qquad \text{eqtn.5.26}$$

where c_{ww} is the autocorrelation of the rectangular window

C_{ww} is the Fourier transform of c_{ww}

P_{xx} is the power spectral density or periodogram estimate

A Hamming window was used in this analysis for which the above spectral broadening occurs in the same manner. Therefore, a side effect of the averaging and hence

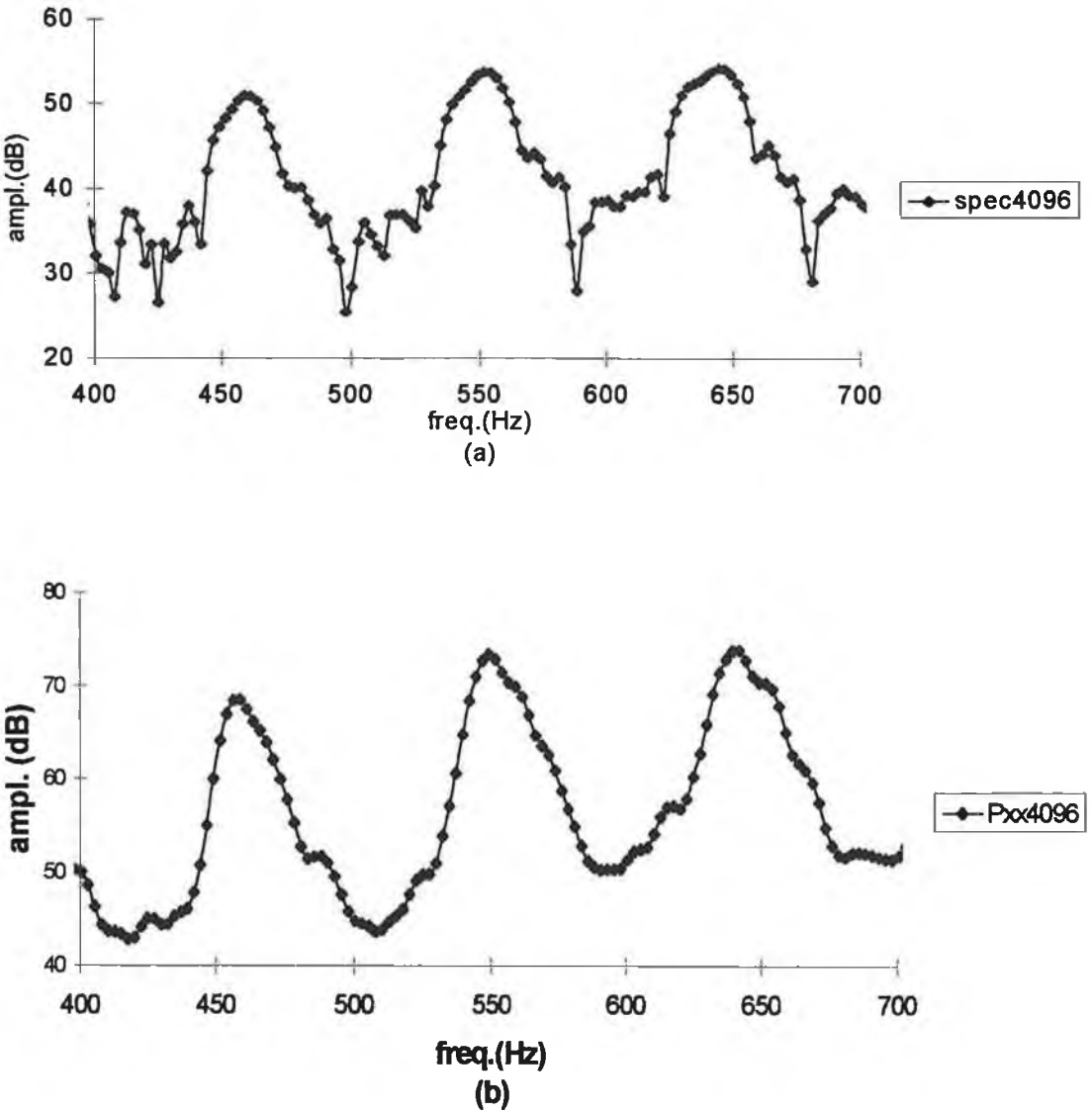


fig.5.17 *Spectral peak broadening and reduced variance of (b) periodogram estimate with respect to Fourier transform values (a) for spectral section of vowel a/ for one of the 'normal' participants in the present study.*

reduced variance of the estimate has been a reduction in spectral resolution. In order to maintain good frequency resolution and to facilitate direct comparison with the Hiraoka method we have chosen a window length of 2048, padded up to 4096 and

hopped by 1024 points providing 8 independent spectral estimates for about 1.2 seconds of speech. The harmonic estimates were obtained as per the Hiraoka method

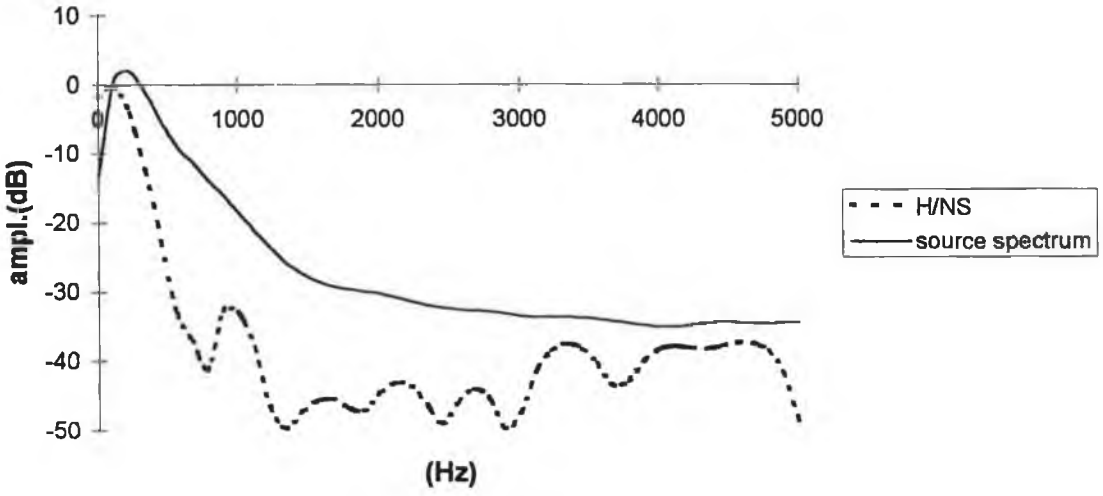


fig.5.18 *Comparison between original source spectrum (g110ar4) for 110 Hz file with std. dev. 4 % additive noise and the $H/N_s(\omega)$ derived source spectral ratio calculated from the output radiated waveform.*

but with double the mainlobe width and all of the above mentioned ratios were also calculated. In addition to these ten ratios, the H/N_S ratio (section 5.2.) was also estimated, using various harmonic numbers which provided a further six indices. The spectra obtained as a result of the $H/N_S(\omega)$ procedure are shown in figure 5.17. These spectra bare a close resemblance to the source spectra as shown in fig. 5.18. It is interesting to consider this derivation in terms of the frequency response of our linear time invariant system to a white noise input.

$$S_{yy}(e^{j\omega}) = C(e^{j\omega})S_{xx}(e^{j\omega}) \qquad \text{eqtn.5.27}$$

$$S_{yy}(e^{j\omega}) = C(e^{j\omega})\sigma_x^2 \qquad \text{eqtn.5.28}$$

where S_{xx} is the power spectrum of white noise and $C(e^{j\omega})$ is the system function. σ_x^2 is the variance of the noise signal and S_{yy} is the output of the system.

5.3.4 Normalised Noise Energy - Kasuya et al[23].

The normalised noise energy (NNE) is the only method to include phase in the noise calculation, or rather to have taken phase into consideration. In the vast majority of speech processing applications phase is not considered and can be quite difficult to calculate and ironically is very sensitive to noise. Secondly, the ear is not responsive to phase information from the speech signal. In any case when considering random processes, the random signal is by definition considered to have random phase.
The speech signal in the m^{th} frame is given by

$$x_m(n) = s_m(n) + w_m(n), n = 0,1,\ldots,M-1 \qquad \text{eqtn.5.29}$$

taking Fourier transformation

$$X_m(k) = S_m(k) + W_m(k), k = 0,1,\ldots,M-1 \qquad \text{eqtn.5.30}$$

Then the NNE is defined as

$$NNE = 10\log\left(\frac{\left[\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{m=1}^{L}|W(k)|^2\right]}{\left[\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{m=1}^{L}|X_m(k)|^2\right]}\right) \qquad \text{eqtn.5.31}$$

where $x_m(n)$ = m^{th} frame of vowel phonation
with periodic component $s_m(n)$ and additive noise
component $w_m(n)$
M is the number of samples within the frame
$N_L = [Nf_LT]$, $N_H = [Nf_HT]$
where L is the number of frames and f_L and f_H determine the highest and lowest frequencies

The denominator in equation 5.31 is calculated directly from the DFT estimate and therefore the problem is to devise a method of estimating the numerator or noise. Representing the squared magnitude of the DFT of the signal plus noise gives

$$|X_m(k)| = |S_m(k)|^2 + |W_m(k)|^2 + 2|S_m(k)||W_m(k)|\cos[\theta(k) - \phi(k)], k = 0,1,\ldots,M-1 \qquad \text{eqtn.5.32}$$

Since the signal becomes small in the harmonic dip region (fig.5.19) the noise can be estimated directly from the spectrum as

$$|W_m(k)|^2 = \left\{ \sum_{r \in D_i} |X_m(r)|^2 (N_i)^{-1} + \sum_{r \in D_{i+1}} |X_m(r)|^2 (N_{i-1})^{-1} \right\}, k \in P_i$$

eqtn.5.33

An interpolation of estimates from adjacent dip regions is used in the estimation of noise in the peak region. Therefore the phase has been considered, yet not calculated. From the development of the spectral consequences of various perturbations outlined in section 5.2, it has been observed that the harmonic energy remains quite stable in some instances (shimmer) and reduces in others (jitter). Taking this into consideration, perhaps it is not appropriate to estimate the noise in the peak regions in this manner. We have seen that the noise floor moves up rather than that harmonic energy reduces in the perturbation of shimmer and additive noise. The main problem with this estimate of noise in the peak region is that here the magnitude will always add, therefore implying that the noise actually adds to the signal strength. If the phase had been included the signal would on average contribute zero energy to the peak region, simply adding more variability to the estimate which is consistent with our development above. Therefore in the absent of phase information it is not applicable to estimate the noise in the harmonic region in this manner and this type of approach is more suited to noise considerations where the noise is a competing sinusoid or other well determined signal. The final result in the analysis is simply a doubling of the noise component of

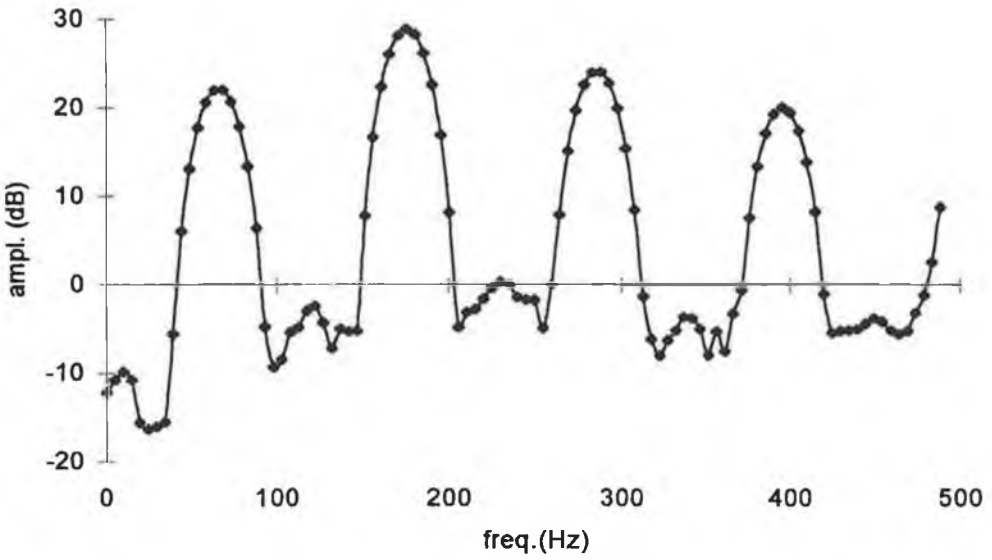


fig.5.19 *Spectral section taken from synthesised output file (s110ar4) with std. dev. 4% noise. The energy at harmonic locations is given as the sum of the energy within* $\pm 2 \times pad/2/(7 \times T) = \pm 2 \times 2048/(91 \times 7) = \pm 6$ *points as shown. Energy within the dip regions (noise) is calculated by summing energy contributions between successive harmonic mainlobe regions.*

the signal and therefore does not interfere with the overall dB value to a great extent (+3dB).

The speech signal was bandlimited at 5 kHz and sampled at 10 kHz with an accuracy of 12 bits per sample. Seven periods were extracted for each frame of the analysis and the process was repeated every 20 ms until the end of the sample length (not given) for the vowel e/. A correction procedure was also programmed for cases in which the energy in the dip region could not be calculated. This involved giving the estimate based on the estimate from the previous dip region. However, as we have stated the broadening of bandwidths can occur as a consequence of jitter and therefore this procedure may falsely compensate against this spectral manifestation. The 'bw' should properly be called the mainlobe width not to be confused with the 1/2 power or 1/4 power 'bw' which are different.

5.3.5 Pitch Synchronous (Four Period) Analysis-Muta et al[24]

The Japanese vowel /u/ was extracted from running speech for analysis. Four pitch periods were extracted for analysis, and the analysis was carried out every 6.4 ms up to 200 ms. Both synthesis and patient and normal data were used in the analysis. In taking four periods of an harmonic signal, three points appear in the mainlobe and the fourth point appears in the valleys (fig.5.19). This hence provides a convenient arrangement for calculating the H/N ratio.

$$P_N(k) = \min_{i=-1,0,1,2} P(4h \leq +i) = P_{Nh}$$
$$(4h - 1 \leq k \leq 4h + 2) \qquad \text{eqtn.5.34}$$

$$R_{NS} = 10\log\left(\frac{\sum_{k=3}^{4L+2} P_N(k)}{\sum_{k=3}^{4L+2} P(k)}\right) \qquad \text{eqtn.5.35}$$

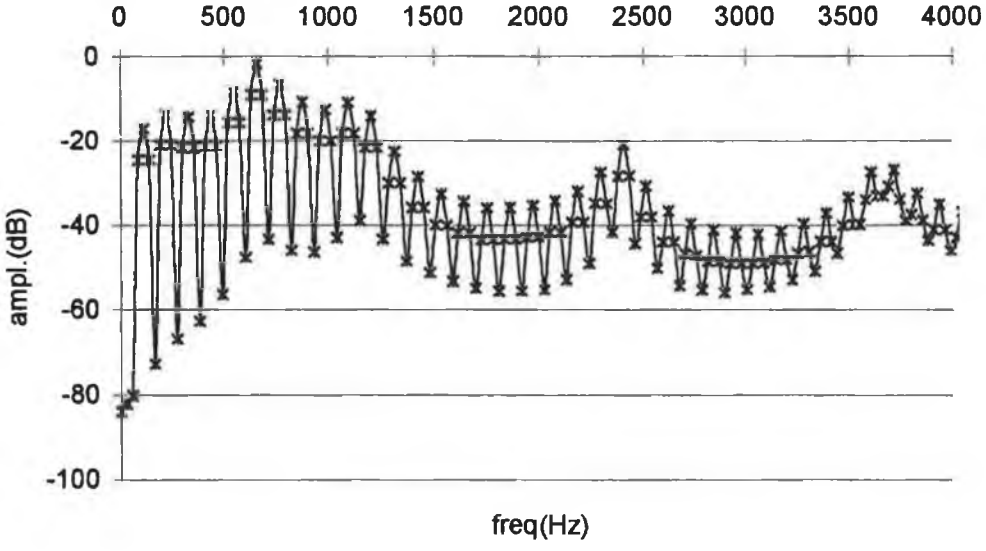


fig.5.20 *Hamming window applied to four periods of 's110ar1', synthesised vowel a/, and Fourier transform taken providing a basic coding scheme for harmonic to noise ratio calculation.*

A major difference from all other approaches is that only the first sixteen harmonics were chosen for the analysis. The justification for limiting the frequency range was that the vowel used was the Japanese vowel /u/ which has the lowest first three formants of all the vowel sounds. Despite this however Muta et al still state that "generally the harmonic structure in the voice signal shows greater distortion in higher harmonics than in lower harmonics". Simply taking the first sixteen harmonics is therefore not recommended. However, we shall see in the results section that there may be considerable advantage in considering voice samples by harmonic number rather than over a given frequency range.

5.3.6 Partial Sum of the Fourier Series - Kojima et al[25]

We have discussed in some detail the basis behind this method in section 5.2.1 when explaining the spectral consequences of jitter, shimmer and additive noise. The Kojima implementation took three periods for the analysis interval. However, it should be pointed out that the inference that for the same length of data, the Fourier series offers better frequency resolution than the Fourier transform is incorrect. In the Kojima et al paper, two spectra are shown, similar to the spectra in fig.5.20, where a) is derived from the Fourier series and b) is calculated from the Fourier transform. A examination of these spectra shows that a) is simply the interpolated version of b). In fact, computationally, obtaining the Fourier series coefficients and the values for the transform at discrete frequencies for a finite length of data is exactly the same. It is simply the theoretical interpretation that is different. Klingholtz et al[14] are also somewhat in error when stating that repeating the waveforms endlessly "eliminated random variations within the speech wave by this procedure. Jitter and shimmer of the three periods were transformed into periodic variations ". We have seen in some detail that when taking two periods (T_1 and T_2) of the speech waveform as a period (T) for Fourier series consideration, produces noise components at 1/T, due to any form of variation from period to period be it due to jitter , shimmer or additive noise. There is an element of truth in the above quote, however, as some of the noise is 'counted into'

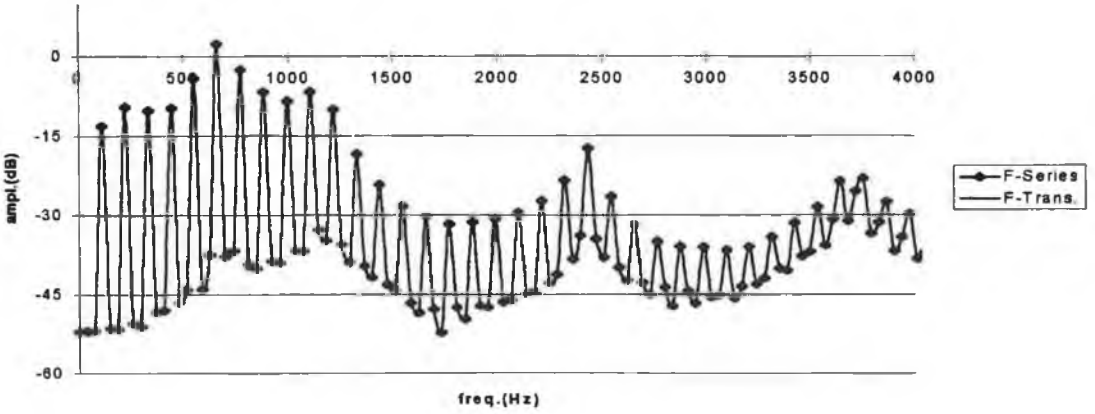


fig.5.21 *Equivalence of Fourier Series and Fourier Transform 'Computation', also illustrating the convenient scheme for H/N ratio calculation with every third component counted as 'harmonic'.*

the harmonics. The overall result of this is to make the harmonics more variable on a period to period basis due to the noise. However, these are added together in the numerator when calculating the ratio and we therefore would expect the variability to average out. The technique developed in section 5.3.9 overcomes this problem. However, both methods (or interpretations) are consistent and the resolution is therefore not different.

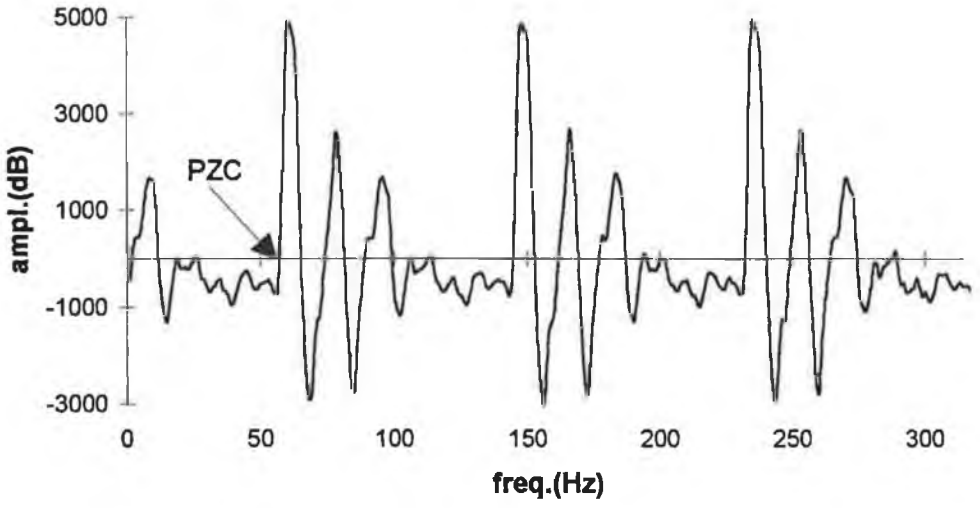


fig.5.21 *Illustrating the point of period extraction (PZC-positive zero crossing before major waveform peak) for calculation of the Fourier series (vowel a/-normal of present study).*

The pitch extraction was carried out by a time domain technique (zpitch3.m) as illustrated in fig.5.21. The beginning of each segment for analysis was located at the positive zero crossing occurring before the major waveform peak. Choosing the zero crossings as the starting point limits the possibility of discontinuities in the waveform and it's concomitant spectral leakage. The analysis length chosen was 325 ms as per Kojima et al with 10kHz sampling frequency as usual. With three periods in the analysis window every third component is counted as an harmonic component therefore providing an easy coding scheme for the analysis. The H/N ratio is given by

$$R_{av} = 10 \times \log_{10}\left(\frac{S_1 + S_2 + S_3 + \ldots + S_n}{N_1 + N_2 + N_3 + \ldots + N_m}\right) \qquad \text{eqtn.5.36}$$

where S_i = harmonic energy of i^{th} estimate

and N_i = noise energy of i^{th} estimate

5.3.7 Partial Sum of the Fourier Series - Two Cycle Analysis

Two periods were taken for the analysis window (as in the development in section 5.2). Each estimate was taken by simply moving the analysis forward one period, therefore each period was compared with the period before and after. Every other frequency component was therefore counted as noise. A clear advantage of this method is that it is more nearly pitch synchronous and the ease with which the noise can subsequently be extracted on inverse Fourier transform of every second series coefficient. Or, equivalently, the signal can similarly be extracted.

Also, the H/N_S ratio, which was thought to more directly represent the source, as outlined in section 5.2, was calculated and hence a source related spectrum was extracted every pitch period. Further examination of these spectra may reveal more specifically the vibratory pattern in the pathological voice. In the light of the negative results reported by Muta et al[23], where inverse filtering of pathological voices using linear prediction proved difficult due to the noise present in the signal, and considering

that we have taken advantage of the fact that random noise characterises the system in the HN_S analysis, it seems to provide an attractive alternative in approach.

5.3.8 Time Domain Averaging - Yumoto et al[26]

A straight forward time domain measure adopted from classical S/N ratio analysis of 'noisy' signals was implemented. The mean value of fifty periods was taken and subtracted from each successive period in order to obtain a noise estimate. The H/N ratio is therefore

$$\frac{H}{N} = 10 \times \log_{10} \left| \frac{\sum_{i=1}^{n} \int_{0}^{T_i} f_i(\tau)\, d\tau}{\sum_{i=1}^{n} \int_{0}^{T_i} [f_i(\tau) - f_A(\tau)]^2\, d\tau} \right| \qquad \text{eqtn.5.37}$$

T_i = i^{th} period

$f_i(\tau)$ = i^{th} waveform and the average waveform is given by

$$f_A(\tau) = \sum_{i=1}^{n} \frac{f_i(\tau)}{n} \qquad \text{eqtn.5.38}$$

In eqtn.5.38 the average energy within a period was calculated by considering the largest period (T) and setting $f_i = 0$, for $T_i < \tau < T$. Therefore, jitter is included in the noise estimate. In order to reduce the jitter we chose the median period (after investigating a number of alternatives) for analysis. (Hillenbrand[1] investigated using the minimum period). The pitch period was extracted by the method reported in 5.3.6. Also, a frequency domain analysis of the method was conducted by taking the FFT of the time domain average and the FFT of each individual period. This therefore satisfied Yumoto's call for a frequency domain analysis of the noise signal (denominator in eqtn.5.37) since,

$$F_i(\omega)-F_A(\omega) = N_i(\omega) \qquad \text{eqtn.5.38}$$

F_i = i^{th} pitch synchronous spectrum

F_A = spectrum of average time domain waveform

N_i = i^{th} noise spectrum

He also states that further study is necessary to scrutinise the relationship between the harmonic to noise ratio and the psychophysical measurement of the degree of hoarseness and we have in a sense considered this with our geometric dB mean ratio. The sampling rate for the analysis was 10 kHz and the vowel used was a/. Four ratios were taken in all, including two dB derived measures.

5.3.9 Pitch Synchronous Harmonic Analysis (PSHA)

With a view to obtaining spectral measures on a period by period basis[27] and in an attempt to overcome the influence of jitter and shimmer on the harmonic to noise ratio, the following novel technique was implemented.

A single period of voiced speech is taken and it is assumed that this period repeats itself in an identical fashion throughout the waveform. This is consistent with our digital model of voiced speech[28] which assumes that a short segment of voiced speech is taken from

$$s(n) = \sum_{m=-\infty}^{\infty} h(n + mN_P) \qquad \text{eqtn.5.39}$$

where as usual

$$h^*(n) = r(n)^*v(n)^*g(n) \qquad \text{eqtn.5.40}$$

where * indicates convolution and

v(n) = vocal tract filter function

r(n) = radiation at the lips

g(n) = glottal waveform

The Fourier Series (eqtn.5.1-eqtn.5.3) can now be applied to the extracted period to compute the Fourier series coefficients a_n and b_n, from which the harmonic energy is determined via eqtn.5.41

$$h_n = (a_n^2 + b_n^2)^{1/2} \qquad \text{eqtn.5.41}$$

This approach is in agreement with the more commonly used Fourier transform implementation of spectral estimation, whose frequency resolution increases with increasing window length. The increased window length in this instance, provides more identical waveforms in the analysis frame. The spectral consequence of this to provide more spectral estimates i.e. increased frequency resolution, but because the waveform is repetitive, the extra spectral estimates are simply zero (fig.5.22). The mainlobe width of the convolving window function decreases and in the limit as the repetitive waveform approaches infinity, the convolved spectral harmonic estimates approach the Fourier series coefficients. This development is useful for showing the equivalence of each approach but it should also be mentioned that the Fourier transform cannot be evaluated for a waveform of infinite extent.

The analysis is initally developed through examination of some simple test signals, in order to introduce the method in a maximally simple fashion. Two sawtooth waveforms and representations of their Fourier spectra are shown in fig.5.23. The waveform in (a) represents the normal waveform of period 'T = 10 ms' (taking the sampling frequency to be 10 kHz). Part (b) of the figure shows a waveform that is identical in every respect to the waveform in (a) except for the pitch period i.e. it represents a scaled version of the waveform (jitter) in (a) with new period T'=T/2 (200 Hz) chosen for simplicity. The equation for the sawtooth waveform is given as

$$f(t) = t \qquad 0<t<T/2, \quad = 0, \; t \geq T/2 \qquad \text{eqtn.5.42}$$

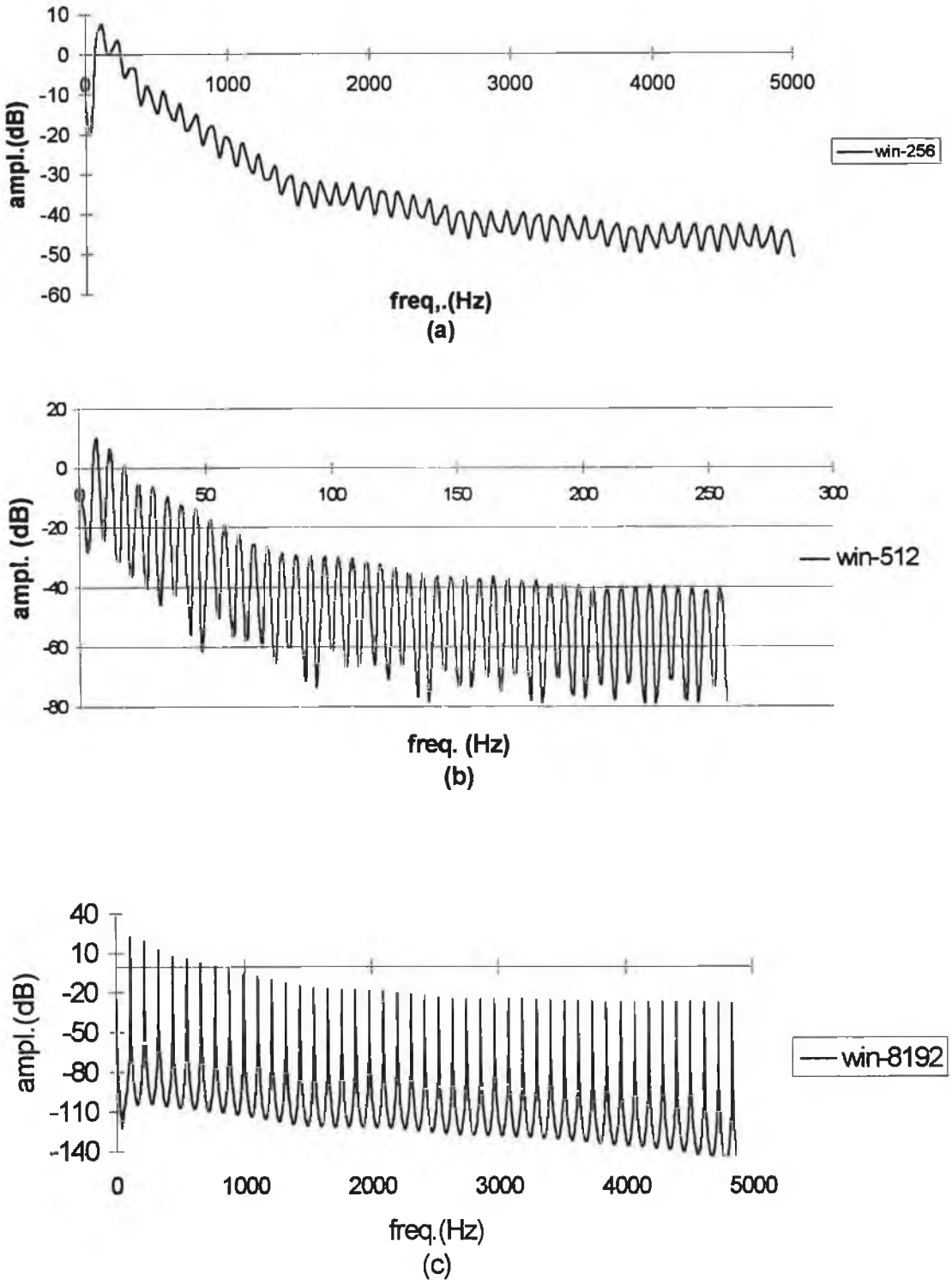


fig.5.22 *Increasing window length for Fourier Transformation of a perfectly periodic waveform (A Rosenberg glottal pulse-110Hz was used). Bandwidths of harmonics approach impulse functions i.e. the transform approaches the series coefficients.*

To compare the waveforms directly, independent of period length, a scaling parameter, which is given as the ratio between the periods, 50/100=1/2 in this example, is used i.e. f(t) is compared with f(k×t).

$$f(k\times t) = SF\times f(t) \qquad \text{eqtn.5.43}$$

where SF is the scale factor, which is dependent on f(t).

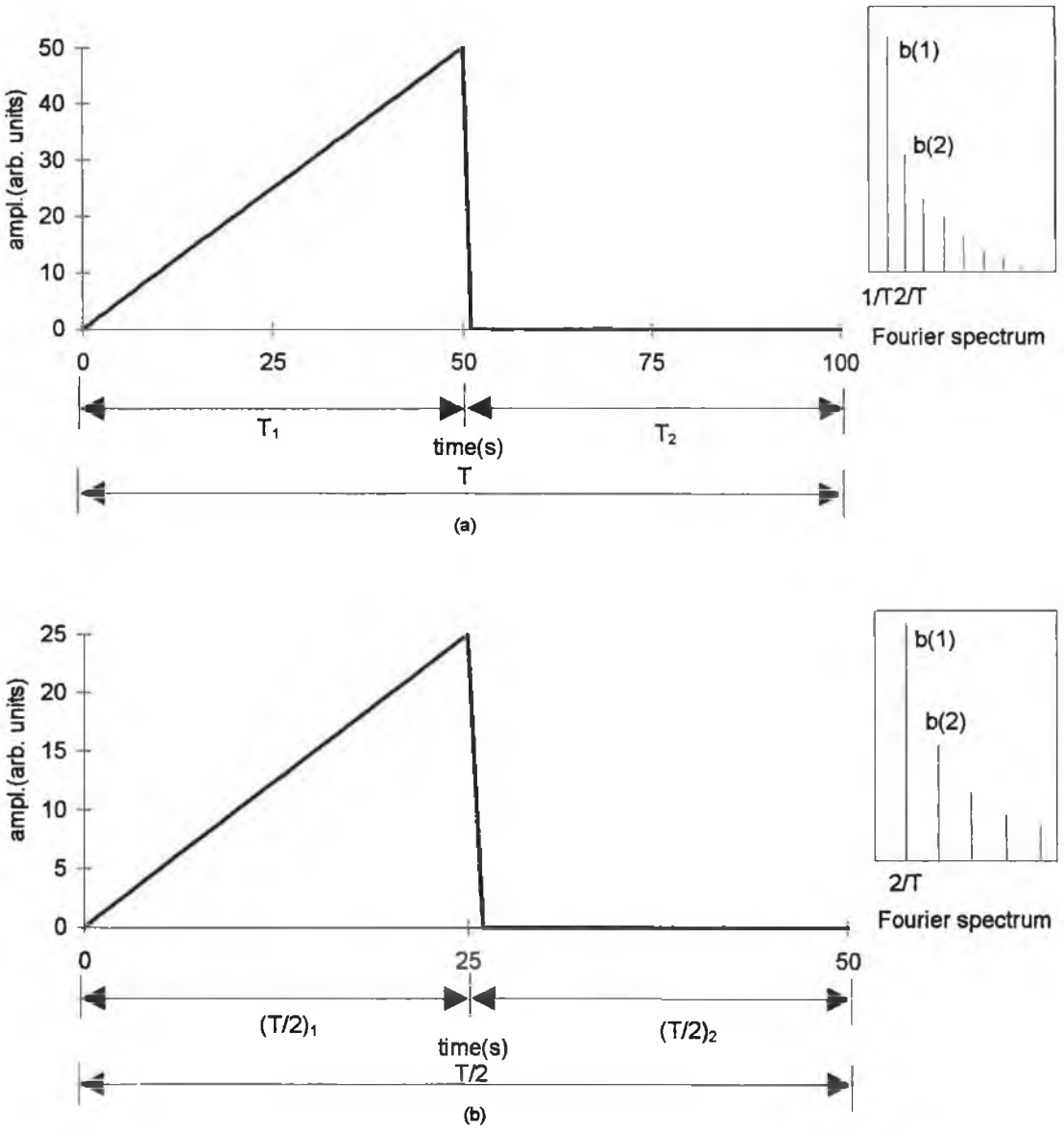


fig.5.23 *Sawtooth waveforms with periods of (a) 10 ms and (b) 5ms with corresponding Fourier coefficients (absolute values) shown in captions.*

For the particular case of the sawtooth waveform, chosen to have a positive slope of one, the scale parameter and scale factor are equal and division by the scale factor, 'k' according to eqtn.5.44 is required.

$$f(k\times t) = k\times f(t)=k\times(t) \qquad \text{eqtn.5.44}$$

Therefore, if the waveforms are identical in every respect except period length, scaling the waveforms with respect to each other eliminates the jitter component (eqtn.5.45).

$$f(t)-f(k\times t)/k = f(t)-k\times f(t)/k=0 \qquad \text{eqtn.5.45}$$

This is easily illustrated with reference to fig.5.23 and eqtn.5.45 and comparing the waveforms at, for example, point t=50 from (a) (T) which gives f(50)-(f(1/2×50))/(1/2)=0. For more typical pitch perturbation values we consider a 2% jitter signal, the perturbed signal having a period of 10.2ms (~98Hz). Subtracting the re-scaled signal as above gives f(t)-f(102/100×t)/(102/100) = 0, where 't' is evaluated at it's usual discrete sample value points. Therefore, for a direct comparison between periods, the re-scaled signal must be evaluated at non integer locations and hence an interpolation algorithm is required.

$$f_{\text{inter}}(i) = f(i) + (f(i+1) - f(i)) \times k \qquad \text{eqtn.5.46}$$

And substituting into eqtn.5.45 gives $f(t)-f_{inter}(k\times t)/k=f(t)-f(t)=0$ as before.

The method as outlined above, in conjunction with the Yumoto et al technique (eqtn.5.37) forms the basis of a jitter free harmonic to noise ratio measurement. However, the need for pitch dependent, pitch synchronous interpolation is avoided by viewing the signals in the Fourier domain.

In the above development for the sawtooth waveform, a scale factor (SF), which simply turned out to be the equal to the scale parameter (k), was required in order to compare two identical waveforms which differed only in period. In general the SF is dependent on the function being scaled. It is of interest to consider the scaling of

cosines, firstly, because the analytical expression for the glottal pulse model consists of cosine terms and secondly, and more importantly, it provides motivation for an alternative frequency domain scaling scheme. Re-scaling cosines gives

$$\cos(\omega' kt) = \cos(\omega t) \qquad \text{eqtn.5.47}$$

where $\omega = 2\Pi/100$ and $\omega' = 2\Pi/50$, $k = 50/100$
and t is the discrete time unit or sample point

From eqtn.5.47 it is observed that amplitude normalisation is no longer required. This is similarly true for our glottal pulse model whose analytic expression consists of two cosine terms. In applying the Fourier Series, any periodic waveshape can be represented as the sum of sine and cosine terms (eqtn.5.1) and utilising eqtn.5.47 to alter the harmonic frequencies it can be seen that any two waveforms that are identical in every respect excepting period length will have Fourier coefficients that bear the same relationship to each other, spaced at an integral number times the inverse of their period. The scale factor is simply the period length. Therefore implementing the scaling in the frequency domain removes the need for a priori knowledge of complicated scaling factors.
Expansion in the time domain results in frequency domain compression and an increase in amplitude, due to the length of the period. This "time compression-frequency expansion" property (eqtn.5.48) of the Fourier series is illustrated by the captions showing Fourier spectra in the top right hand corner of fig.5.23 (a) and (b).

$$s(kt) = \frac{1}{|k|} F\left(\frac{f}{k}\right) \qquad \text{eqtn.5.48}$$

f - frequency, k - scaling parameter
F- Fourier series coefficient, t - discrete time

From eqtn.5.48, if the waveshapes are identical in all aspects except for period, T, then the Fourier coefficients (i.e. the harmonic amplitudes for each period), are in direct

relation to one another (fig.5.23 (a) and (b) (captions))). For our 100 Hz and 200 Hz cosine waves eqtn.5.48 has harmonics spaced at 1/T and 1/T′ as expected. For more general functions, if we compare the normalised (i.e. divided by the window length) harmonics for T with those of T′, they should be identical i.e.

$$h(n\times 1/T)-h(n\times 1/T') = 0 \qquad \text{eqtn.5.49}$$

where $h(n) = n^{th}$ harmonic

In this manner we are comparing the waveforms based by harmonic number as opposed to the more usual comparison between 'same frequency' location. For the sawtooth waveform the signal f(t) can be time shifted for 'oddness' about x=0 i.e. f(t) = -f(t) and the waveform can then be written in terms of the Fourier series coefficients as

$$f(t) = \frac{2}{\Pi}(\sin \omega t - \frac{1}{2}\sin 2\omega t + \frac{1}{3}\sin 3\omega t - \frac{1}{4}\sin 4\omega t + \ldots)$$

eqtn.5.50

with cosine terms equal to zero. A scaled version is then simply obtained by substituting $\omega' = k\times\omega$ for ω. The amplitude coefficients for the sawtooth waveform have a '1/x' characteristic and the energy (eqtn.5.41) therefore follows a '$1/x^2$' curve i.e. the spectrum drops off at 6dB per octave beginning at the fundamental. The magnitude of the Fourier coefficients as opposed to the energy is shown in the captions in fig.5.23 and fig.5.24 for ease of illustration. The second advantage of this approach is that the need for interpolation has also been removed. Adding the spectra according to harmonic number is also of benefit for considering the average glottal flow characteristics.

Shimmer is also conveniently removed using the pitch synchronous approach. If the waveform is normalised pitch synchronously (in either domain) then the problem is immediately removed. Note that here we have considered shimmer to mean that the waveform is the same in every respect except amplitude at every point in the cycle. Therefore, this approach enables us to directly compare scaled periods and hence forms

the basis of a measurement technique that can perform an harmonic intensity analysis which is independent of jitter and shimmer. In order to obtain the harmonic to noise ratio, the pitch synchronous harmonics are averaged according to harmonic number (not frequency location) to form an average harmonic spectrum. The average harmonic spectrum is then subtracted from the individual pitch synchronous spectra in order to obtain the spectral noise estimate and hence the H/N ratio

$$\frac{H}{N} = \left[\frac{\sum_{i}^{L} \sum_{j}^{M} \left(h_i(T_j)\right)^2}{\sum_{i}^{L} \sum_{j}^{M} \left(h_i(T_j) - h_{i(AV)}\right)^2} \right] \qquad \text{eqtn.5.51}$$

$h_i(T_j)$ = i^{th} harmonic of j^{th} spectrum and

$h_{i(AV)}$ = average of i^{th} harmonic

M = total number of spectra

L = total number of harmonics

T = time

Based on our development, if $h_i(T_j)$-$h_i(av) \neq 0$, then either a wave shape change has occurred or additive noise is present in the signal i.e. eqtn.5.51 provides jitter and shimmer free harmonic to noise ratios, where the jitter component has resulted from waveforms that are identical in every respect except for different periods. We will term this scaled jitter.

In order to introduce the method in it's most basic form we have carried out the analysis on simple waveshapes, including the glottal waveform. Now we consider the important deviation that is introduced when analysing the output speech waveform. This occurs as a result of adding by harmonic number as opposed to exact frequency location. The problem arises as a result of the harmonic-formant interaction[1,29]. If we had developed the technique using the output waveform and considered the waveforms to be the same in every respect except period we would have been dealing with waveforms that are impossible to realise in practice. The fundamental frequency, f0 = 1/T, governs the harmonic source spectrum frequency locations. The amplitudes

of these harmonics are modified on passing through the vocal tract filter. For the output waveform therefore, two cycles cannot be the same in every respect except for a period difference as they receive different resonant contributions. We are assuming, of course, that the vocal tract resonance configuration is exactly the same in each case. In the frequency domain we see that the slight offset in harmonic structure leading to a slightly different resonance contribution to the harmonic output spectrum. Therefore the jitter has effectively been turned into harmonic shimmer. It is interesting to note that jitter cannot exist independently of shimmer for the output radiated speech waveform. Through comparison of typical jitter values for normal (0-1%)[30] (and pathological) voices with the resonant bandwidths for the first five formants (Table.5.3) we can try to estimate the magnitude of the effect. The relationship between the location of the fundamental and formant locations must also be taken into consideration. A correction scheme could then be developed based on a correlation between jitter, for a given f0 and the resultant 'harmonic shimmer'.

FORMANT	FREQUENCY	BANDWIDTH
1st	650.3	94.1
2nd	1075.7	91.4
3rd	2463.1	107.4
4th	3558.3	198.7
5th	4631.3	89.8

Table.5.3 *Formant data for Russian vowel a/ from Fant*[16]

Another, ultimately more useful, approach is to use pitch synchronous inverse filtering. Rosenberg[31] has obtained excellent results for inverse filtering based on pitch synchronous analysis. The method as introduced by Mathews et al[32] who state "the contributions from the vocal tract can be uniquely separated and examined" is to first estimate the model parameters defined by $H(\omega_j) = R(\omega_j)V(\omega_j)G(\omega_j)$, where R, V and G indicate frequency representations of radiation at the lips, the vocal tract and glottal waveform respectively. Values are computed for $H(\omega_j)$ and matched to the spectrum

of the waveform under investigation. The parameters are then adjusted so as to minimise the error. The spectrum of the glottal pulse waveform is then computed according to

$$G(k) = \frac{X_n(e^{j\frac{2\pi}{Np}k})}{R(e^{j\frac{2\pi}{Np}k})V(e^{j\frac{2\pi}{Np}k})}, 0 \le k \le N_p - 1 \qquad \text{eqtn.5.52}$$

where N_p is the number of samples in the p^{th} period.

Therefore the pitch synchronous spectral approach not only allows a convenient means for eliminating jitter and shimmer artifacts from the signal but can also be used as a method for inverse filtering. The result is important for our development in that it we can combine the two, first obtaining the glottal frequency spectra according to eqtn.5.51 and therefore nullifying the harmonic-formant interaction effects and then applying the H/N ratio (eqtn.5.50). It is a convenient arrangement in that we are using the same analysis techniques. Also, of course, the inverse filtering not only eliminates the effects of the formant frequencies on the H/N ratio but also supplies the glottal spectrum (and pulse, if required).

An objection may be made to calling the approach pitch synchronous in that (from eqtn5.50) an average of several cycles is required in order to obtain our estimate. A slight modification of the equation however gives us the harmonic to noise ratio based on the difference in consecutive spectra. A combination of such approaches may provide the optimum approach. Obtaining the spectrum pitch synchronously allows us to take a lot of measurements on the signal and as stated in section 5.3.5, opens up the possibility of making spectral measurements specifically related to vibratory events.

In our development, the jitter artifact was considered to result form a waveform identical in every respect with it's neighbouring waveform except for period differences. We have call this scaled jitter. Several possibilities exist for changing the pitch period other than simply scaling the periods. Fig.5.24 illustrates some examples. It can be imagined that the vibratory mechanism is functioning correctly but that due to some abnormality in tissue properties of the folds that consecutive closure events occur

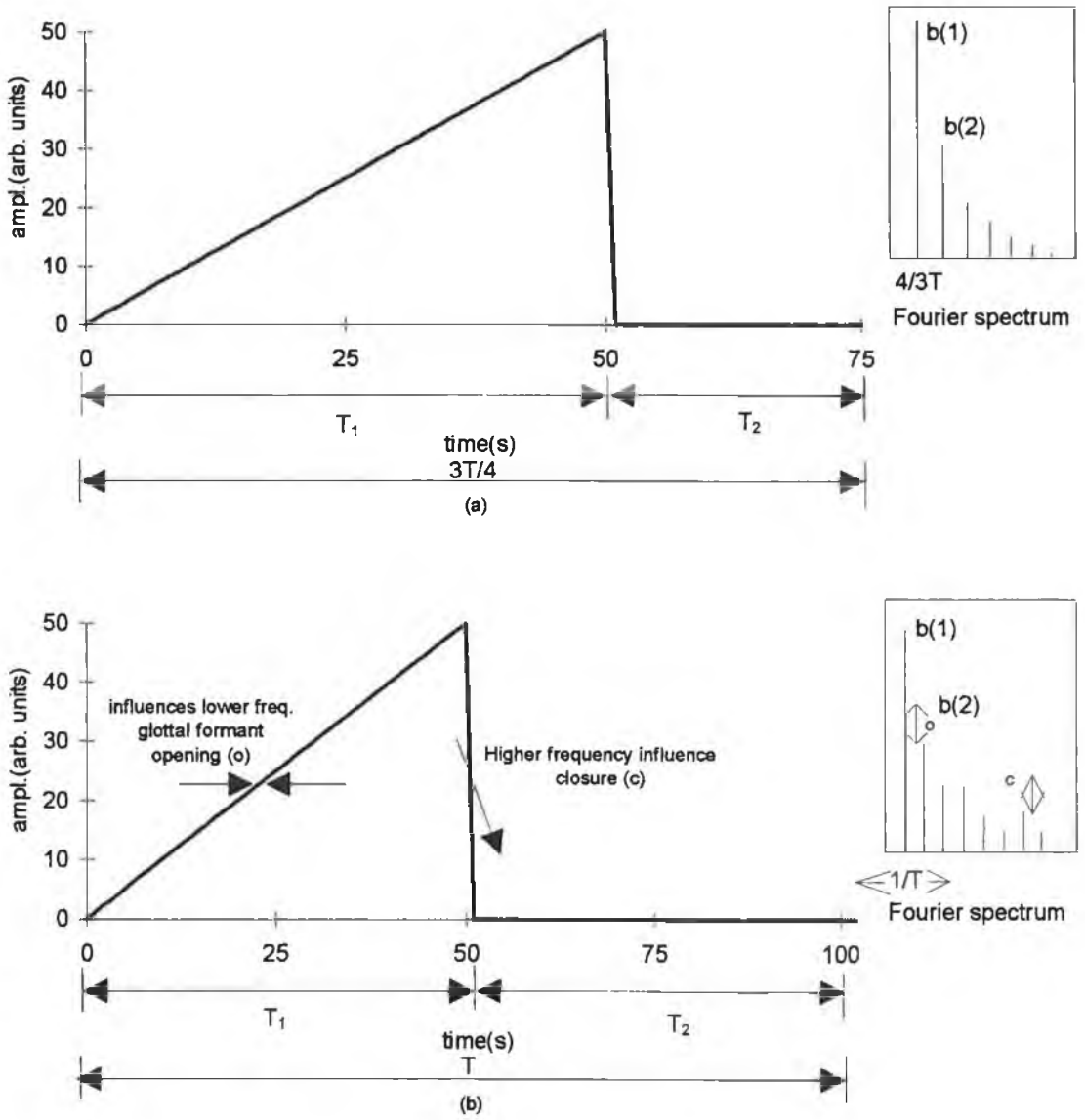


fig.5.24 (a) *Period change due to shortening of the closed phase. Spectral envelope is maintained and for the sawtooth waveform since the Fourier coefficients are 1/x the relative heights of harmonics are also maintained.* (b) *Other glottal events that may be the cause of period changes with the rising part of the waveform corresponding to glottal abduction which primarily influences the low frequency end of the spectrum. The falling edge corresponds to adduction which mainly affects the upper partials.*

in quite a random fashion. This case is illustrated in fig.5.24(a) where the open phase remains exactly the same but the closed phase is either elongated or truncated with respect to normal. Rothenberg[33] has considered a somewhat similar example of

aperiodicity where "the vocal fold vibrations are periodic, but an irregular mucous bridge is making the onset of the airflow aperiodic". Ladefoged[34], in attempting to provide a jitter free index representing the random noise components associated with breathiness proposed using "only part of a cycle and compare(d) it with the corresponding part of the next cycle". This technique, seems applicable for use on glottal waveforms with differing closed phases. Although, for use on the output radiated speech waveform the harmonic-formant interaction will also be present. The spectral consequence of changing the closed phase is to change the relative height of the harmonics. However, the spectral envelope remains the same and therefore "zero-padding"[21] the periods until they are of equal length will regain the relative 'harmonic' strengths. Possible mechanisms for the aforementioned scaled jitter might include differences in tension of the thyroarythenoid or cricothyroid muscles or differences in the mass of the folds taking part in the vibration from cycle to cycle. The period may also change due to a change in adductory or abductory function (fig.5.24(b)) which may result from changes in cricoarythenoid activity. Other possible mechanisms are found in cases involving vocal pathology where the presence of vocal fold nodules or mass lesions give rise to aperiodicities and turbulent flow. It is of considerable interest to spectrally characterise these conditions.

List 1-7 provides the basis for a possible algorithm for investigating glottal characteristics.

1. If $|h_i(T)-h_{iAV}|$ in eqtn.5.51 gives a value of zero then there is no additive noise present in the signal, no change in open quotient (OQ - open time to closed time in one period of oscillation) and no change in waveshape.
2. 'noise' present indicates either (a) additive noise, (b)OQ has changed or (c) waveshape change.
3. Check for (b) spectral envelope may be the same, therefore 'zero pad' to make periods equal and calculate $|h_{izc}-h_{avzc}|$, where h_{izc} is the i^{th} harmonic for the spectrum derived form the 'zero padded' waveform.
4. If noise $\neq 0$ then

1. a) or c)
2. a) $|h_i-h_{av}|$ difference is constant for all h_i.
3. Check for (c) abduction(look for lower f changes), adduction(tilt - look for higher f changes)

For more advanced procedures matching the spectral changes with the LF model is the required approach[28]. In the actual implementation the pitch period was extracted according to the method indicated in fig.5.21. The number of points taken for calculating the partial sum of the Fourier series is given by f_cut/fsam×2×median(period), where f_cut=3800 Hz (cut off frequency of low pass filter) and fsam=10 kHz (sampling frequency).

5.4 Results:

In this section an examination and interpretation of the main results obtained from the present implementations of the methods detailed in the last section is given and compared to the results obtained from the original analysis. The order of presentation is the same as in the previous section. The analysis programs were run on all synthesis files and on the patient and normal data. A presentation of all the results is not possible due to the number of ratios returned from all programs. In our systematic manner of evaluating the H/N ratio techniques we firstly ran the program on the synthesis files with three levels of additive noise of std. dev. 4%, 8% and 16% for six values of fundamental frequency (f0) in equal steps from 80 to 350 Hz. If the results from this analysis were encouraging the response of the ratio with respect to the jitter and shimmer files was examined. The potential diagnostic strength of the ratio was then evaluated by examining the ratio with respect to the patient/normal data set. In each section that follows the H/N ratio refers to the H/N ratio calculated by that method e.g. the harmonic to noise ratio for the Kojima technique is given by eqtn.5.36, section5.3.6 and the results are given in section 5.4.6.

5.4.1 Noise Reducing Filter

The results obtained by Kitajima for the S/N show moderate correlation to results obtained from spectrographic ratings. The improved version of the method, implemented in the present study, which included an f0 dependent averaging, a broader frequency range, two less FFT operations and a more accurate assessment of the filtering operation was not very accurate at showing the variation of the harmonic to noise ratio with f0 (fig.5.25(a)). However a much improved harmonic to noise ratio pattern is obtained when a dB-derived mean is used. (fig.5.25(b)).

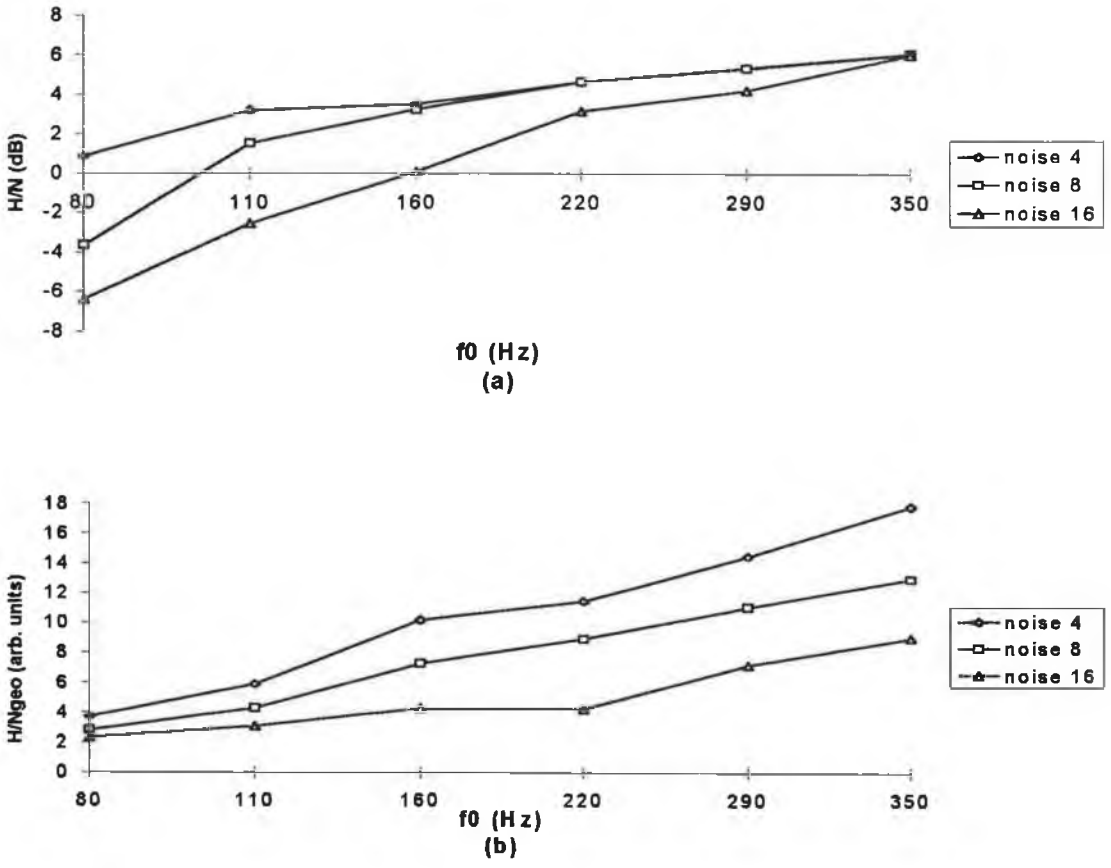


fig.5.25 *Implementation of Kitajima's Noise Reducing Filter (a) H/N vs f0 and (b) dB derived 'geometric' ratio H/Ngeo vs f0 for three levels of additive source noise having std. dev. 4%,8% and 16%. The trend of (b) with respect to f0 is taken as 'normal', showing equal increments for each level of noise at a given f0 location.*

The trend in part (b) of fig.5.25 showing H/Ngeo vs f0 is explained in the next section and for now it is simply considered to represent normal. It can be seen that equal increments in H/Ngeo occur for increases in additive noise levels at a given frequency location. The improvement in representing the noise increases is similarly shown in figures 5.26(a) and (b).

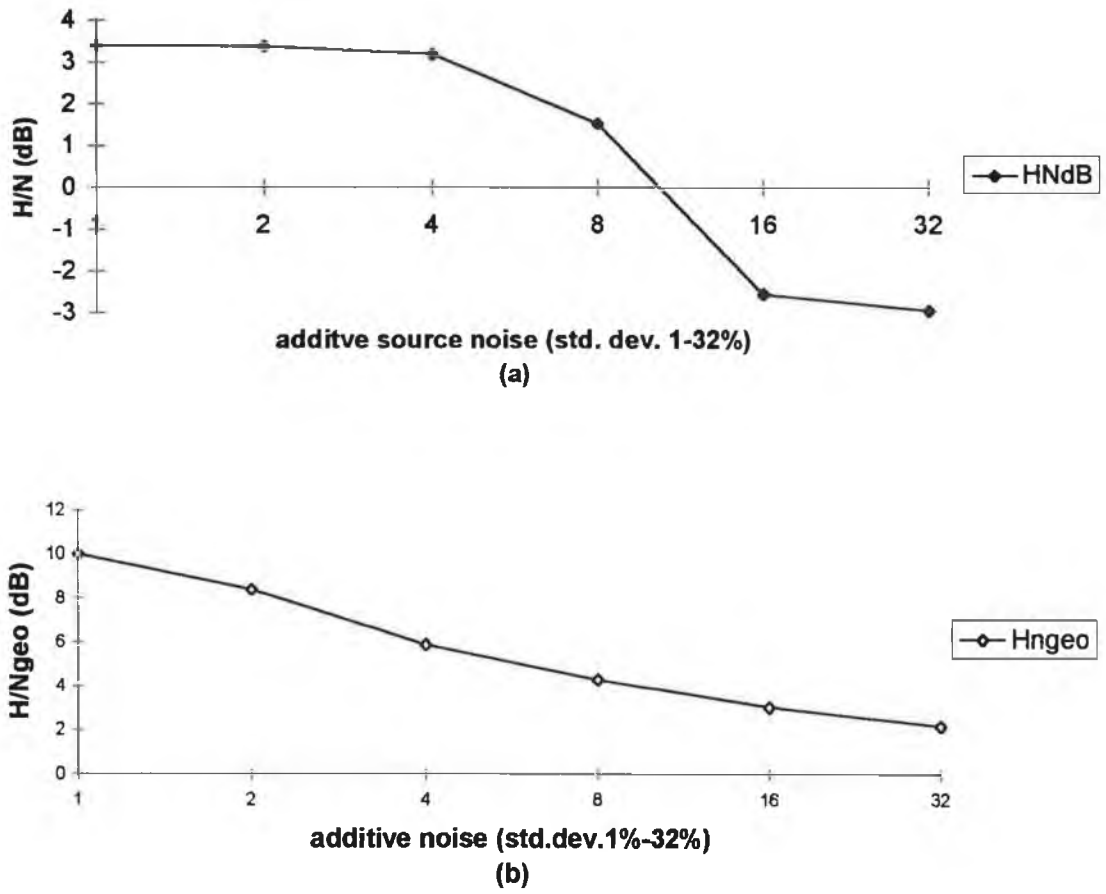


fig.5.26 *Implementation of Kitajima's Noise Reducing Filter (a) H/N vs additive noise and (b) dB derived 'geometric' ratio H/Ngeo vs additive noise. The H/Ngeo displays a more regular response with respect to equal increments of additive source noise.*

In (b), for five doubling in noise levels there is a corresponding decrease of about 2 dB per doubling. The improvement here is due to the fact that a dB spectrum was used at the outset before averaging. A similar result would have been obtained if we had left the original spectrum and summed all values greater than the moving average and then taken the dB values (not with respect to the noise). It is interesting to note that

accurate estimate of the noise levels are not required in order to obtain a reasonable estimate of the harmonic to noise ratio trend with f0. All that is required is that a level with respect to which to take as noise is taken in a consistent manner. The response of the method to all perturbation measures is shown grouped together fig. 5.27.

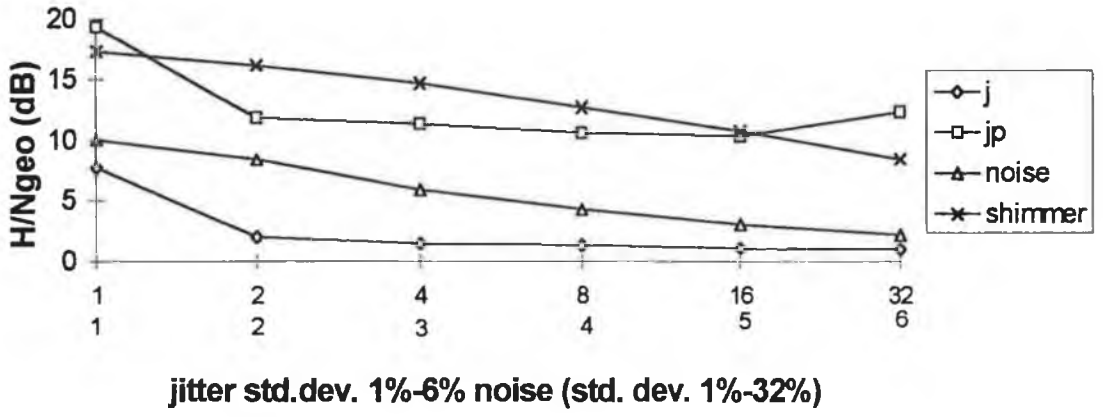


fig.5.27 *The response of the H/Ngeo ratio to all four perturbation measures. The ratio is linearly responsive to additive noise levels as required and somewhat insensitive to cyclic jitter and shimmer. However, the method is most sensitive to random jitter.*

In consideration of the basis for the method, using a moving average filter applied to he speech spectrum and recalling the spectral characteristics of the four perturbation measures (section 5.2) fig.5.27 is the expected result.

Figure 5.28 show how H/N and H/Ngeo performed as potential indicators of vocal pathology. As expected, perhaps, H/N shows no separability and although there is also considerable overlap between patient and normal data for the H/Ngeo ratio, seven normals show distinctly higher values. The result is significant at the 5% level using a one tailed, two sample, equal variance t-test, therefore showing some potential differentiability due to the modified approach.

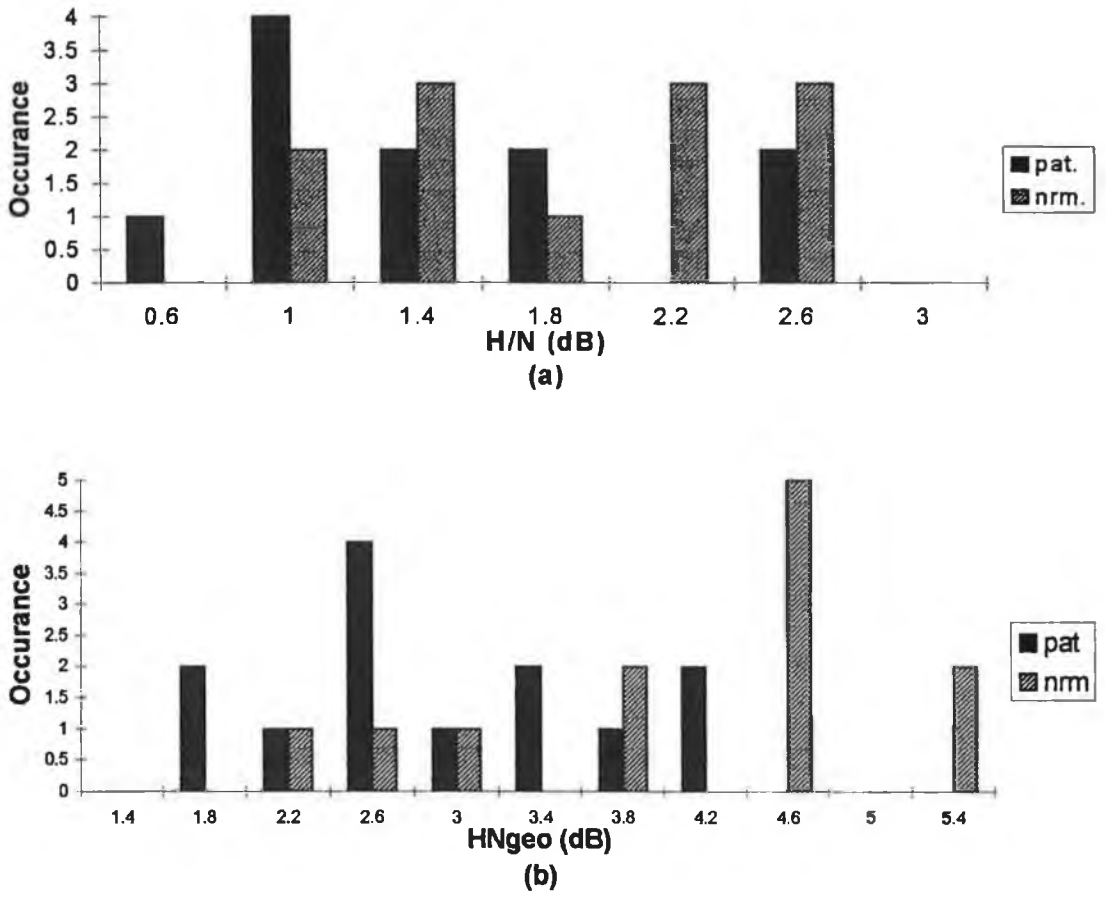


fig.5.28 *Performance of* (a) *H/N ratio and* (b) *H/Ngeo as potential indicators of vocal pathology for the data set -13 patients (varying pathologies) and twelve 'normals'.*

5.4.2 Relative Harmonic Intensity (Hr)

Hiraoka's implementation stresses the fact that a high relative f0 amplitude which is known to be a good indicator of breathiness may be missed by conventional harmonic to noise ratio estimates. This echos what we have said in section 5.2 regarding the fact that the waveshape could be the same from period to period and therefore have a high harmonic to noise ratio and yet could still show considerable pathology due to the unusual, although consistent waveshape. The Hr ratio as defined in eqtn.5.23 guards against missing this anomaly. Our implementation sticks closely to the Hiraoka method although they have failed to state how the harmonic energy estimate was calculated. Their method was tested on a group of 36 normals and 30 patients. Improved separability with Hr as opposed to Sr was reported. Figure 5.29 shows the H_r index

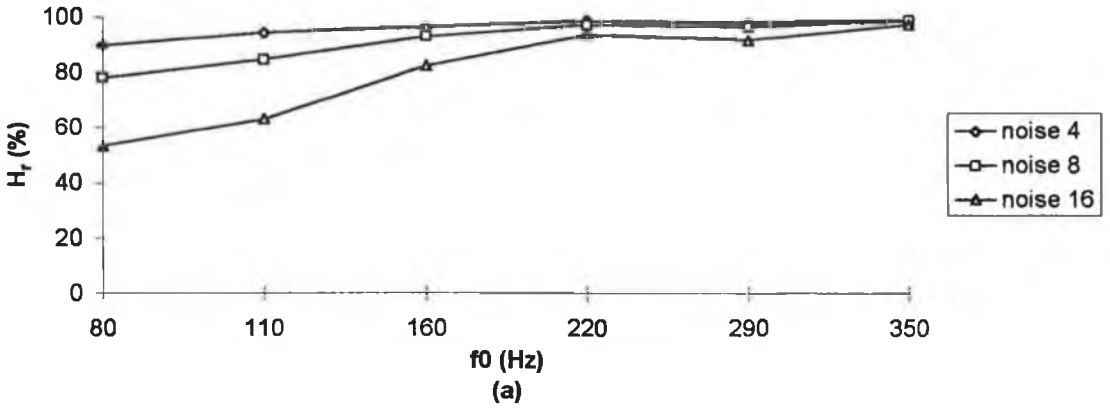


fig.5.29 *Variation of H_r (%) index with f0 for three levels of additive noise*

plotted against f0 for three levels of additive noise. The trend for the Sr ratio is very similar (not shown). As stated previously, the trend of increased harmonic to noise ratio with increasing f0 is explained in detail in section 5.5. The Hr ratio for the patient/normal data are shown in fig. 5.30. The Hr value showed greater separability than the Sr index (not shown) but considerable overlap still remains with the result not being significant at the 5% level using the one-tailed, equal variance, two sample, student's t-test.

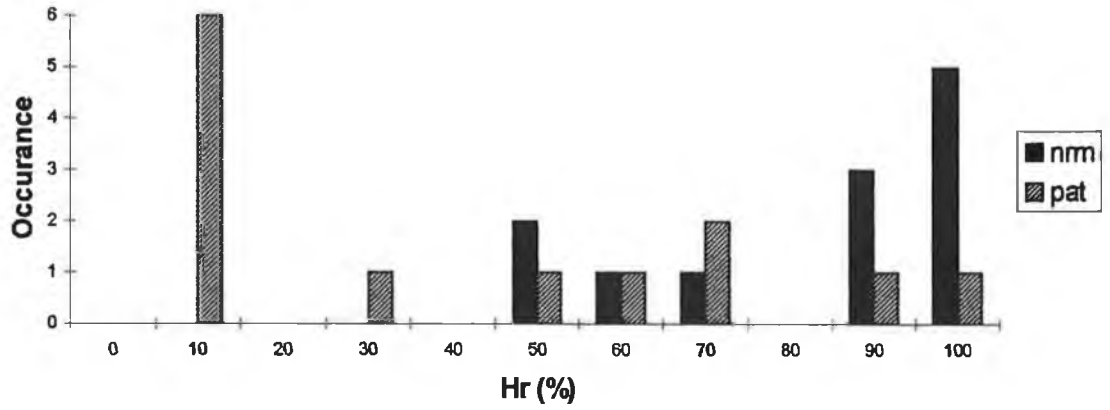


fig.5.30 *Differentiability performance of Hiraoka's Hr index for the patient(13) / normal(12) data set. The result is not significant at the 5% level (one-tailed student's t-test)*

5.4.3 Periodogram Averaged Harmonic Analysis (PAHA)

The PAHA technique is a new approach for determining the H/N ratio for voiced speech. As stated in the analysis section, for a random signal, averaging (n) successive power spectral densities reduces the variance of the resultant spectral estimates by a factor of 1/n. It seems applicable to use this method of periodogram averaging for investigating voice pathologies as we have often modelled the noise as additive random noise. Direct comparison of the results obtained from this approach with results obtained from the Hiraoka approach is possible if we apply the H/N ratio to the single spectrum (i.e. Hiraoka's method). Figures 5.31 (a), (b) show the H/N values for the PAHA and Hiraoka analyses.

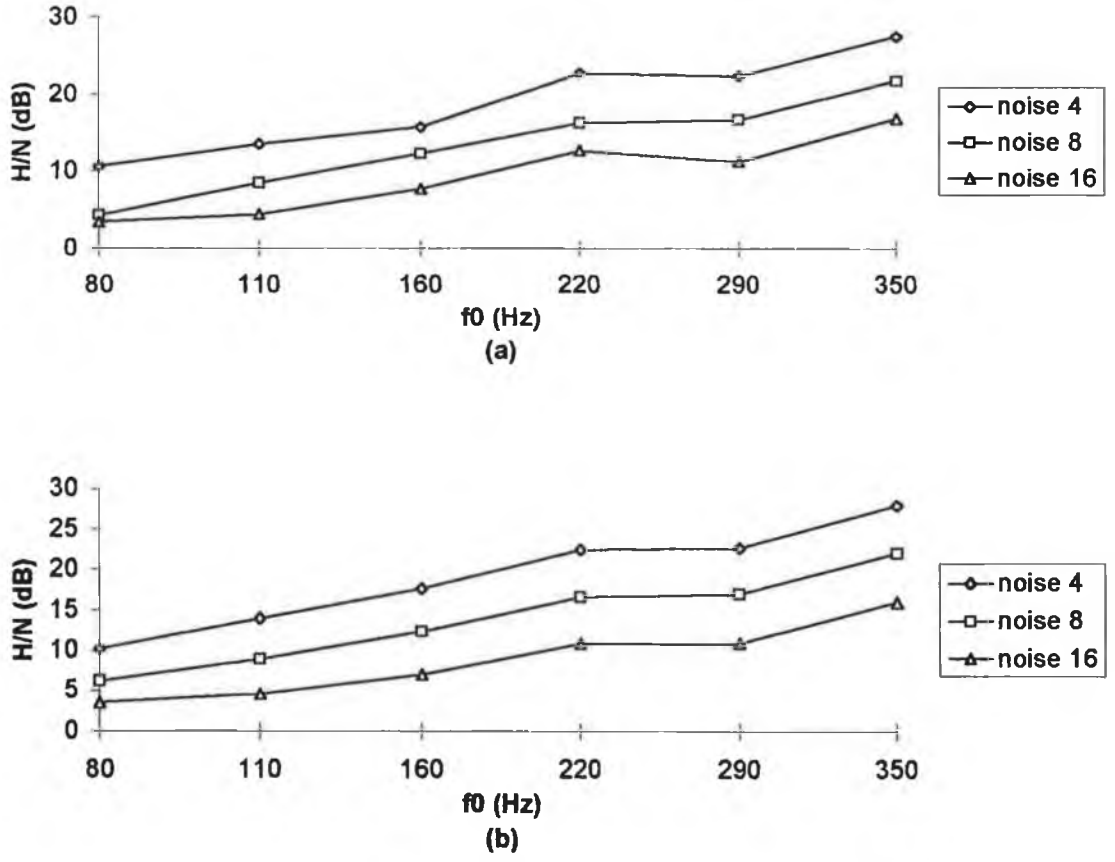


fig.5.31 *H/N vs f0 for (a) Hiraoka method and (b) periodogram averaged method (PAHA). The reduced variance of the spectral estimates is very evident.*

The benefit of overlap and averaging is immediately apparent in the PAHA case. The results reflect more accurately the noise levels present in the voice signal due to more consistent spectral estimates. Figure 5.32 shows the H/N ratio plotted with respect to the perturbation measures.

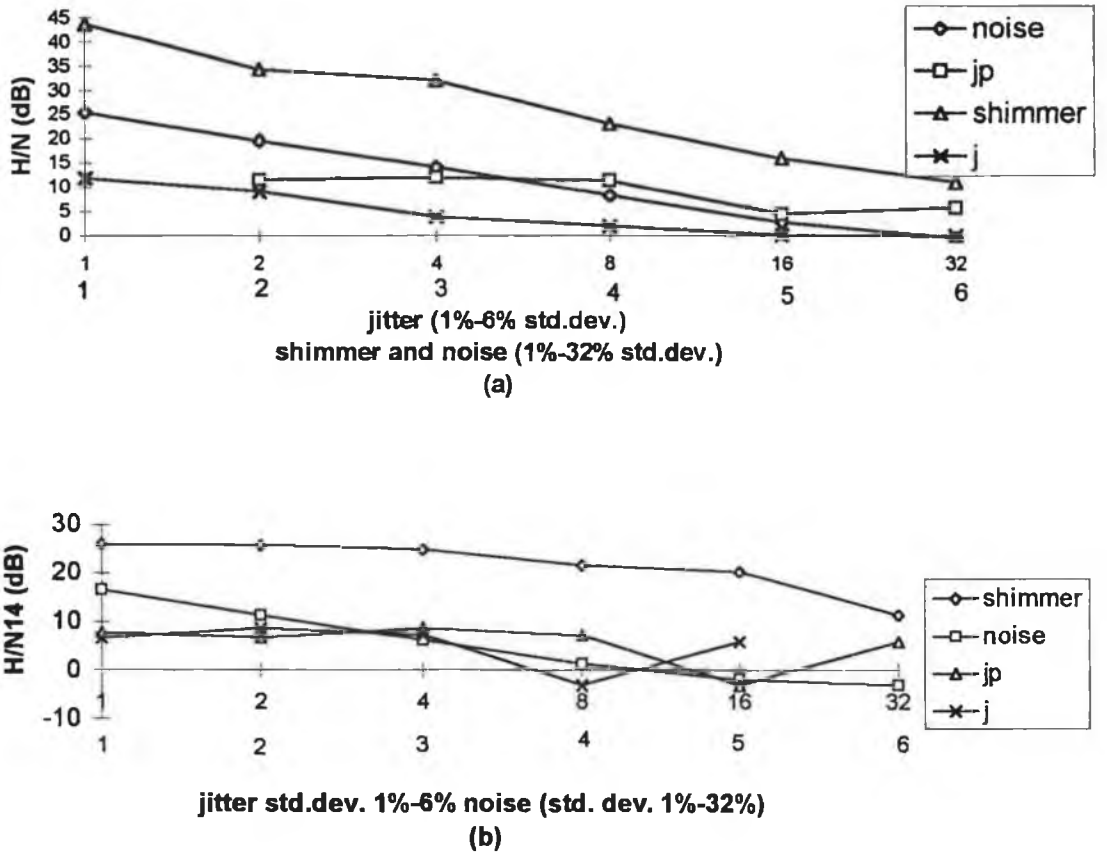


fig.5.32 (a) *Periodogram averaged harmonic analysis (PAHA) with H/N ratio for the four perturbation measures of additive noise, cyclic jitter (jp), random jitter (j) and shimmer* (b) *with H/N14 ratio - limiting the frequency range from 1-4 kHz*

The result (fig.5.32 (a)) is similar to the modified Kitajima approach (fig.5.27). The method is very sensitive to random jitter variations, even at 1% std. dev. random jitter. The noise levels are reflected well and the ratio is somewhat insensitive to shimmer. Again, these results are in agreement with and can be explained by the spectral characterisation development in section 5.2.

In section 5.2 the motivation for different ratios, reflective of perceptual and physical characteristics was developed. Three new ratio types were introduced with the PAHA technique, corresponding to limiting the frequency range (H/N14-harmonic to noise ratio for frequencies between 1 to 4 kHz - fig.5.32 (b)), perceptually based ratios (H/Ngeo and H/Ngeo14-geometric means - fig.5.33) and source correlated ratios (HN_S - fig.5.34).

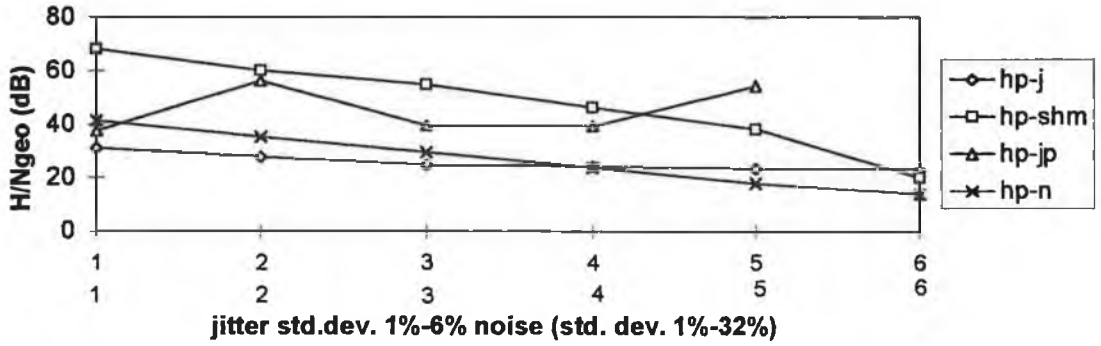


fig.5.33 *Geometric dB mean ratio vs the four perturbation measures of jitter (j), shimmer (shm), cyclic jitter (jp) and additive noise (n).*

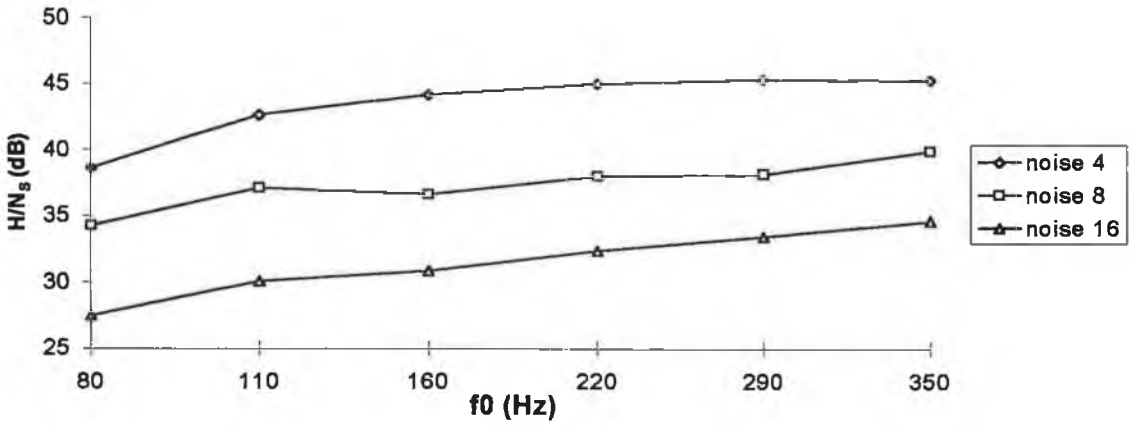
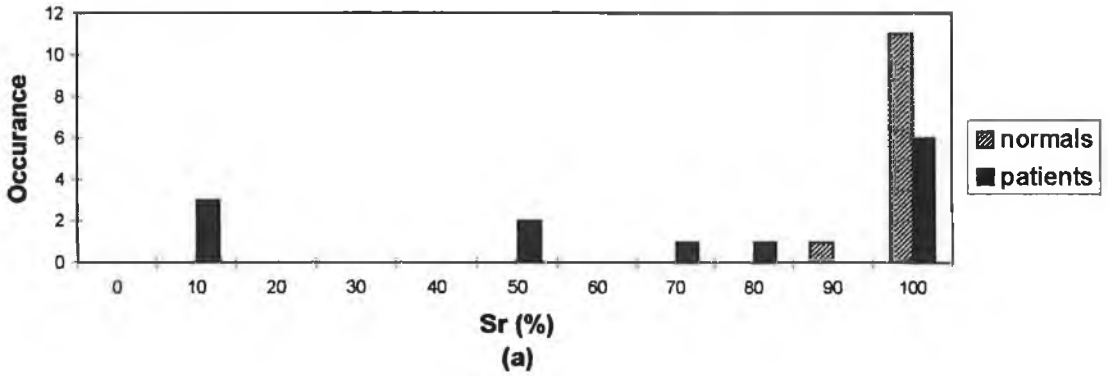
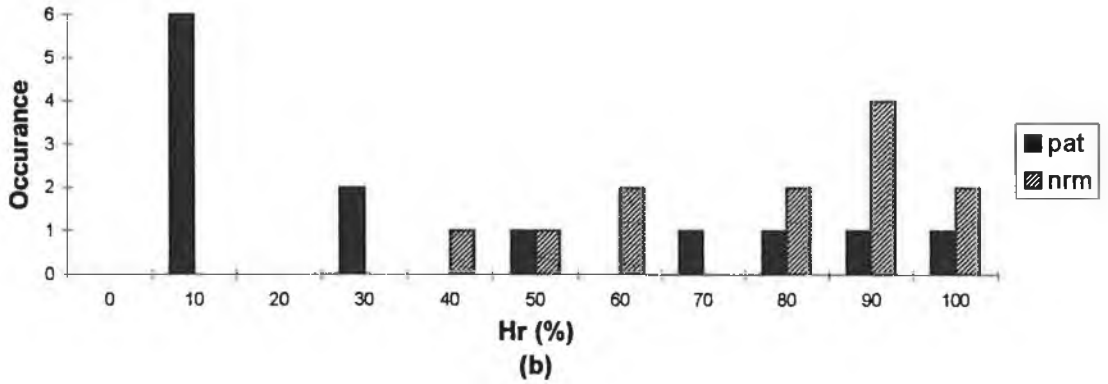


fig.5.34 *Variation of H/N_S source ratio with f0 for three levels of additive noise. Note the variation with f0 is considerably reduced.*

The ability of all the above mentioned ratios at separating the patient and normal data is shown in figures 5.35(a),(b) and 5.36 (a), (b) and (c).





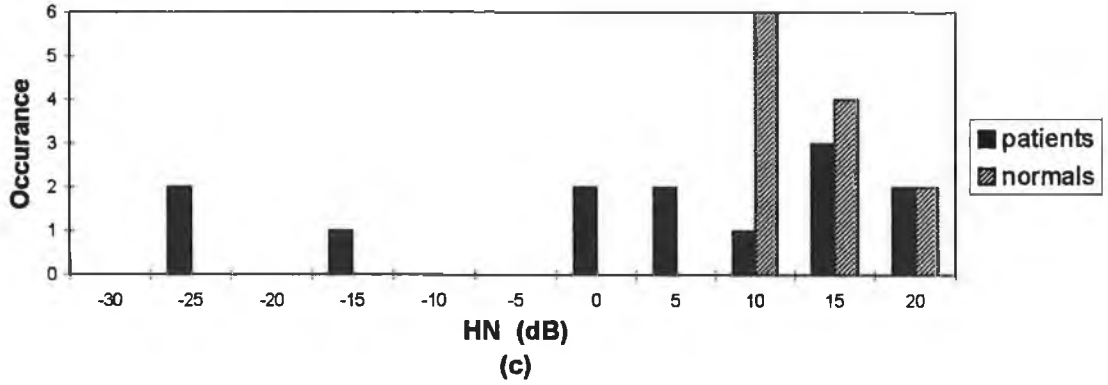


fig.5.35 (a) *Total percent of harmonic energy to signal energy, Sr (%), no significant differentiability,* (b) *total percent of harmonic energy (excluding f0) to signal energy Hr (%) showing some improvement but considerable overlap still exists,* (c) *harmonic to noise ratio, H/N (f0 not included in H calculation). A similar graph results when including f0 in H.*

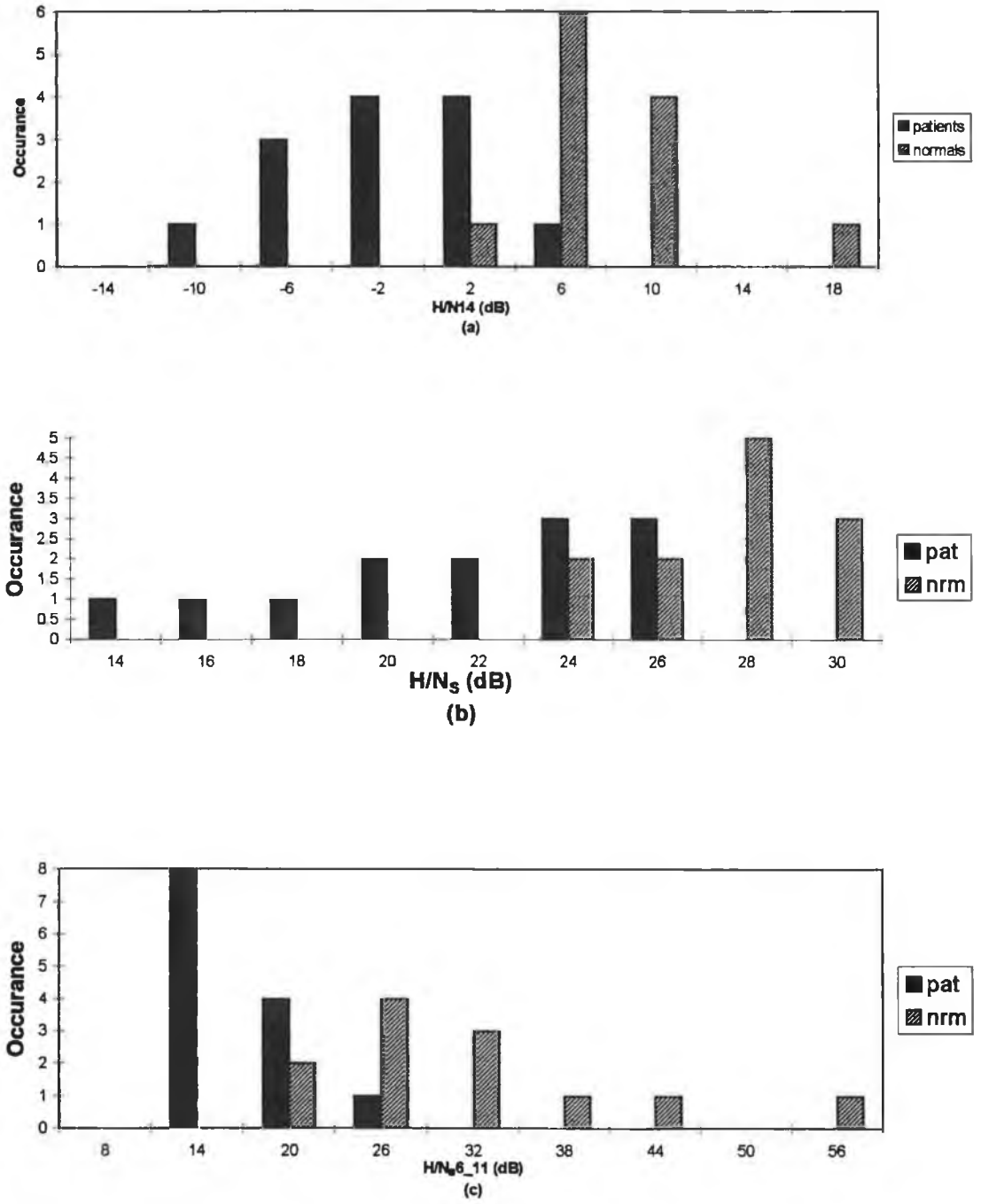


fig.5.36 *Considerable improvement in separability is achieved with* (a) *bandlimiting the H/N ratio and applying source related ratios* (b) H/N_s *and* (c) $H/N_{s6\text{-}11}$ *(bandlimited (according to harmonic number) source ratio). All results aresignificant at the 5% level (one-tailed, equal variance student t-test).*

5.4.4 Normalised Noise Energy (NNE)

In determining the NNE as defined in eqtn.5.31, Kasuya et al examined five frequency regions for investigating vocal pathology. They found 1-5 kHz to provide the highest degree of separability. We have chosen 1-4 kHz (NNE_{14}) based on this finding as the cut off frequency of the low pass filter was 4 kHz. Their ratio was tested on an extensive set of pathological voices (186) of varying etiologies and 64 normals. The error rate reported for normals was 9.4 % and 24.2 % in the case of pathology. Nonetheless, they found their method to be superior to Hr (Hiraoka) and H/N (Yumoto) at separating glottic cancer patients from normals with NNE giving approximately half the number of errors as the other two methods. NNE is shown for the synthesis data with increasing f0 in fig.5.37.

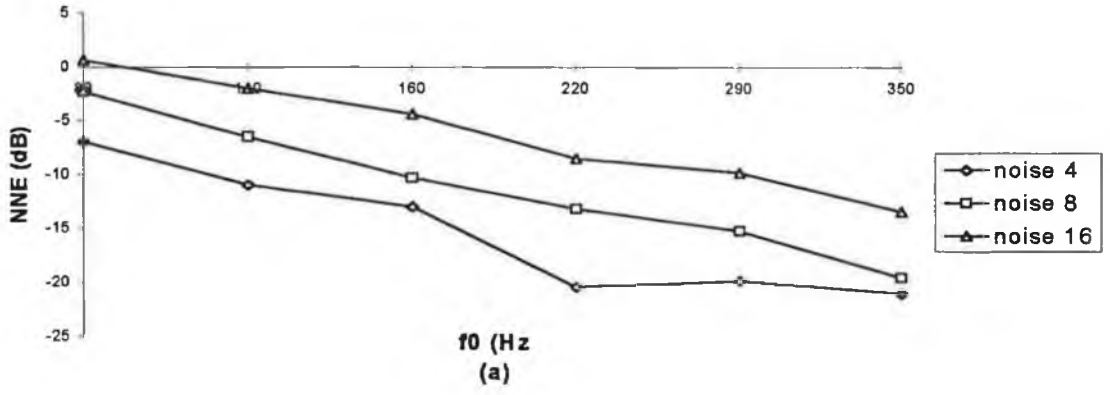


fig.5.37 *NNE vs f0 for three levels of additive noise. Approximately equal intervals for each increase in noise level at a given f0 location.*

The response of NNE to the four perturbation measures (not shown) is very similar to the PAHA response, showing a linear response to noise but sensitive to jitter. In it's ability to separate our patient/normal set, NNE proved to be a rather poor indicator whereas NNE_{14} showed good separability (fig.5.38).

One possible point to query about this study is that samples were taken at stable pitch and increased loudness. Samples providing the smallest NNE were taken as representative for that person. But of course the increased loudness causes an overall change in the spectral composition with a decrease in the first formant bandwidth,

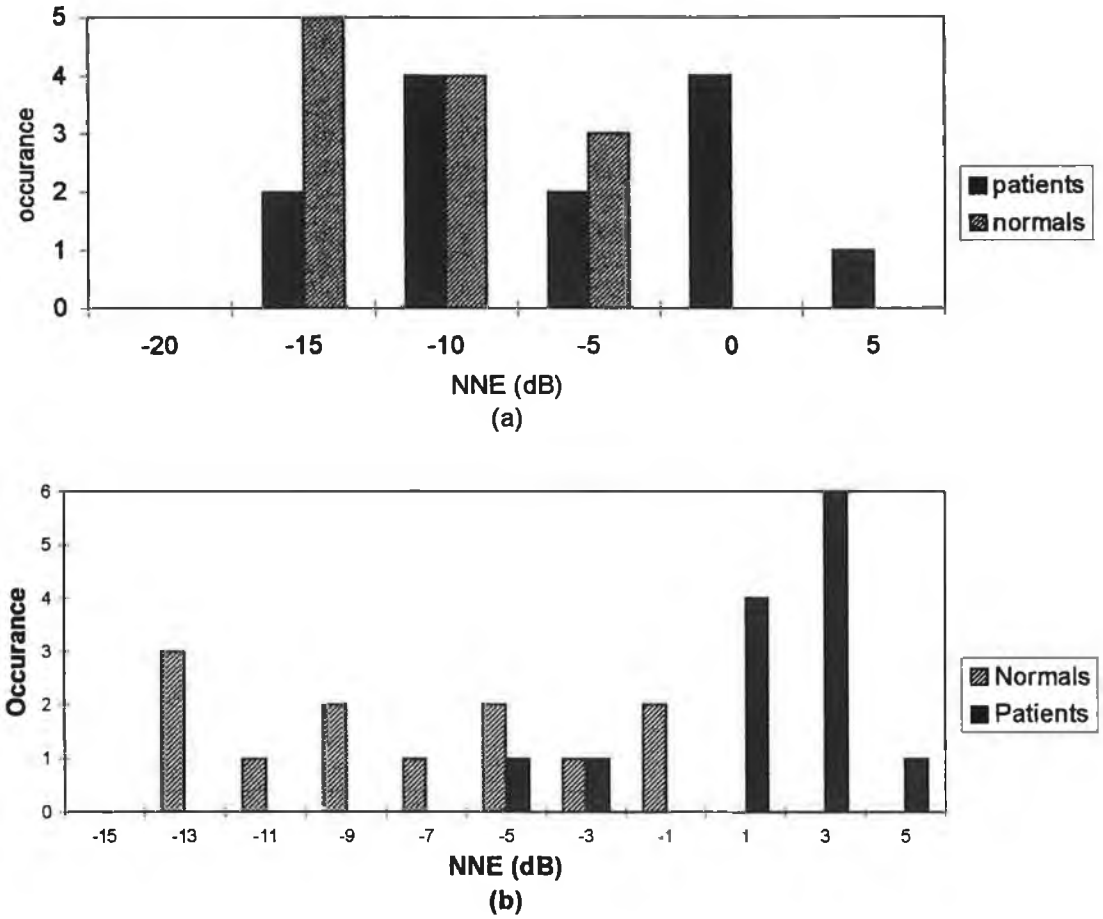


fig.5.38 (a) *NNE and* (b) NNE_{14} *(bandlimited from between 1 an 4 kHz). Only the latter is significant at the 5% level (student's t-test).*

evidenced in the time domain pulse by a less rapid decay of the frequency of the first formant. However, it does raise the question of what the best criterion should be for obtaining samples. Whatever is chosen, the format must necessarily be of a simple nature, be accurately rated perceptually and cause minimum discomfort to the patient.

5.4.5 Pitch Synchronous (Four Period) Analysis

The Muta et al technique differs from all others in that only the first 16 harmonics are taken to represent the speech waveform. The justification for this was explained in section 5.40. As a result of this approach, 1600 Hz is covered for a 100 Hz signal and 3200 Hz is covered for a 200 Hz signal. It is interesting to note that this study uses the

N/S ratio to compare pre- and post- op samples and no comparison is made therefore between patient and normal data. All patients (only six participants) shows a decrease in the N/S ratio. However, if we take post op data to represent normal and the pre op data to represent the true patient data then the ratio does show overlap. The implication is therefore that the method may be useful for intra patient analysis as might have been expected. However a further complicating factor to this assumption is that patients often show a considerable change in fundamental frequency before and after surgery[35]. Surprisingly then, there seems to be some merit in this approach of representation by harmonic number as opposed to complete frequency range. This isssue is discussed in the next section (5.5). As shown in fig.5.39, the N/S ratio doesn't reflect the noise to signal ratio changes with f0 very well. Not surprisingly, the N/S ratio shows little ability to separate the patient/normal data (fig.5.40).

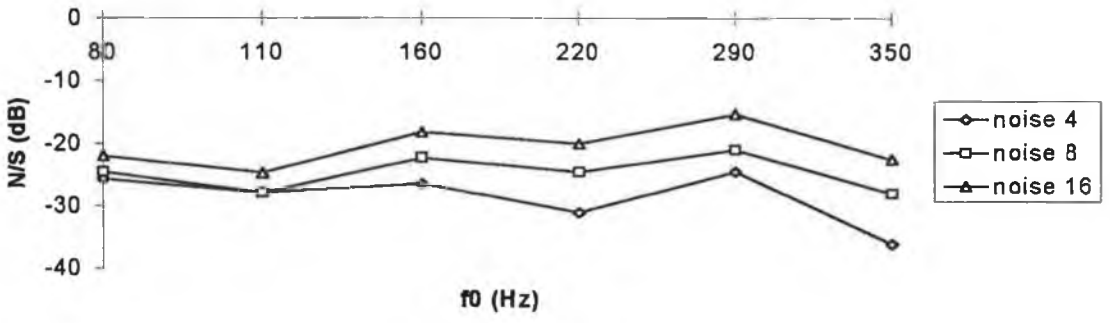


fig.5.39 *N/S vs f0 for three levels of additive random source noise std. dev. 4%, 8% and 16 %.*

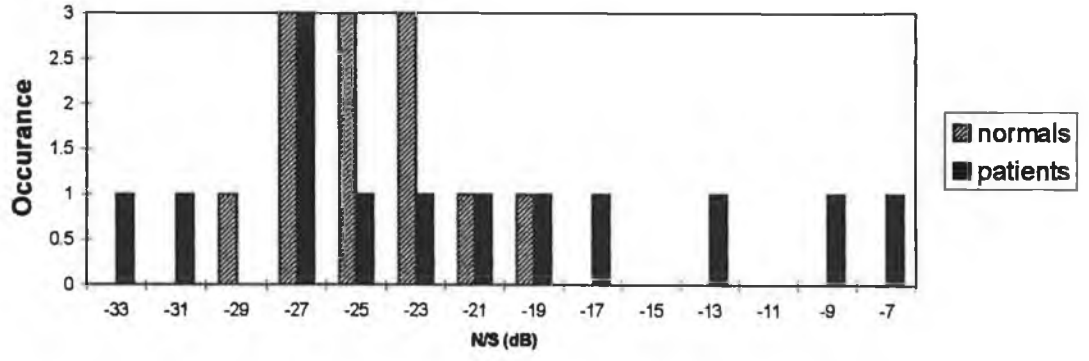


fig.5.40 *N/S for the patient and 'normal' data set.*

5.4.6 Partial Sum of the Fourier Series - Kojima et al

In Kojima's study, a set of fourteen males and fourteen females were used as normals and a set of twenty males and ten females comprised the patient data. The results were compared with spectrographic and auditory impressions. Some overlap was evident for the data set with normals ranging from 15 to 23 dB and patients ranging from -1.5 to 20.3 dB. The method is attractive due to it's simplicity in coding, with two points noise, one point harmonic etc. . However, the method is easily offset due to jitter. Figures 5.41 (a) and (b) show the measure with respect to f0 (for three levels of additive noise) and the four perturbation measures. The ability of the index at differentiating between the patient/normal data set is shown in fig.5.43.

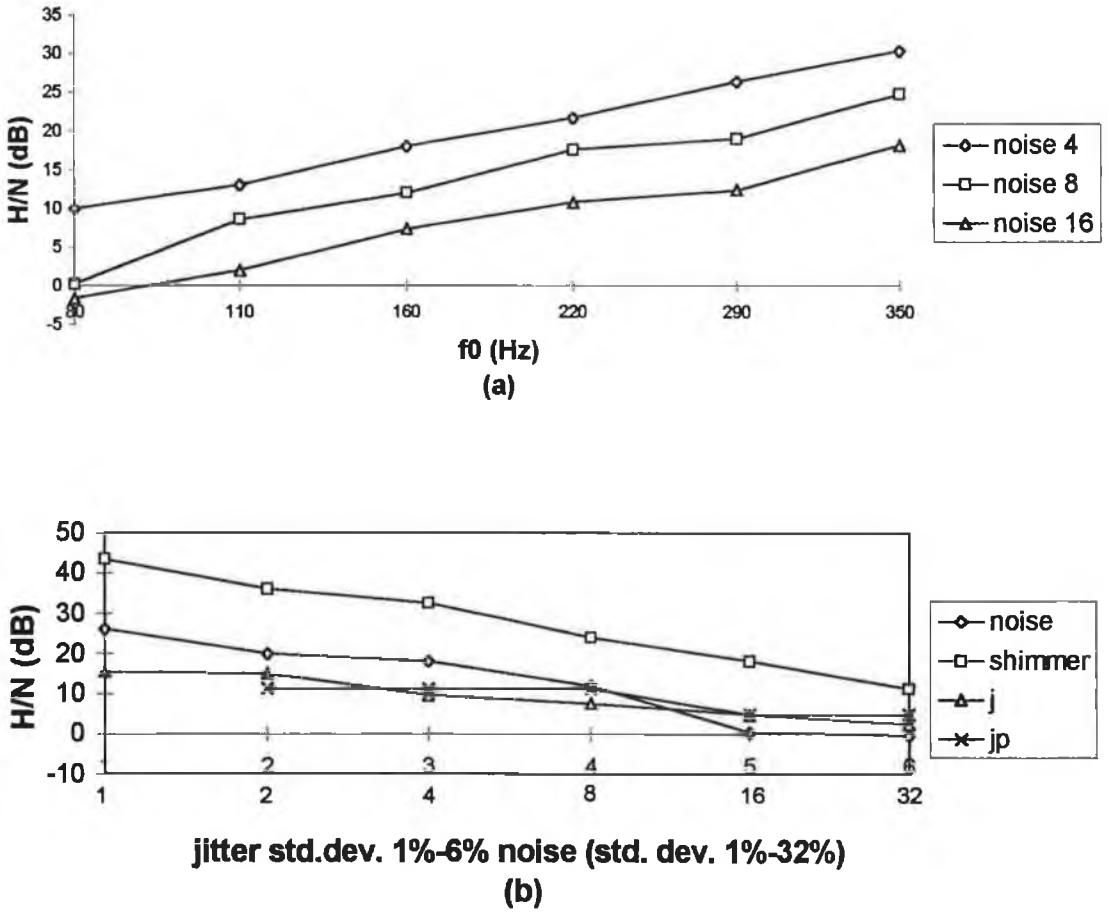


fig.5.41 *Response of the Kojima H/N ratio to (a) f0 for three levels of additive noise and (b) the four perturbation parameters where 'j' indicates randon jitter and 'jp' cyclic jitter.*

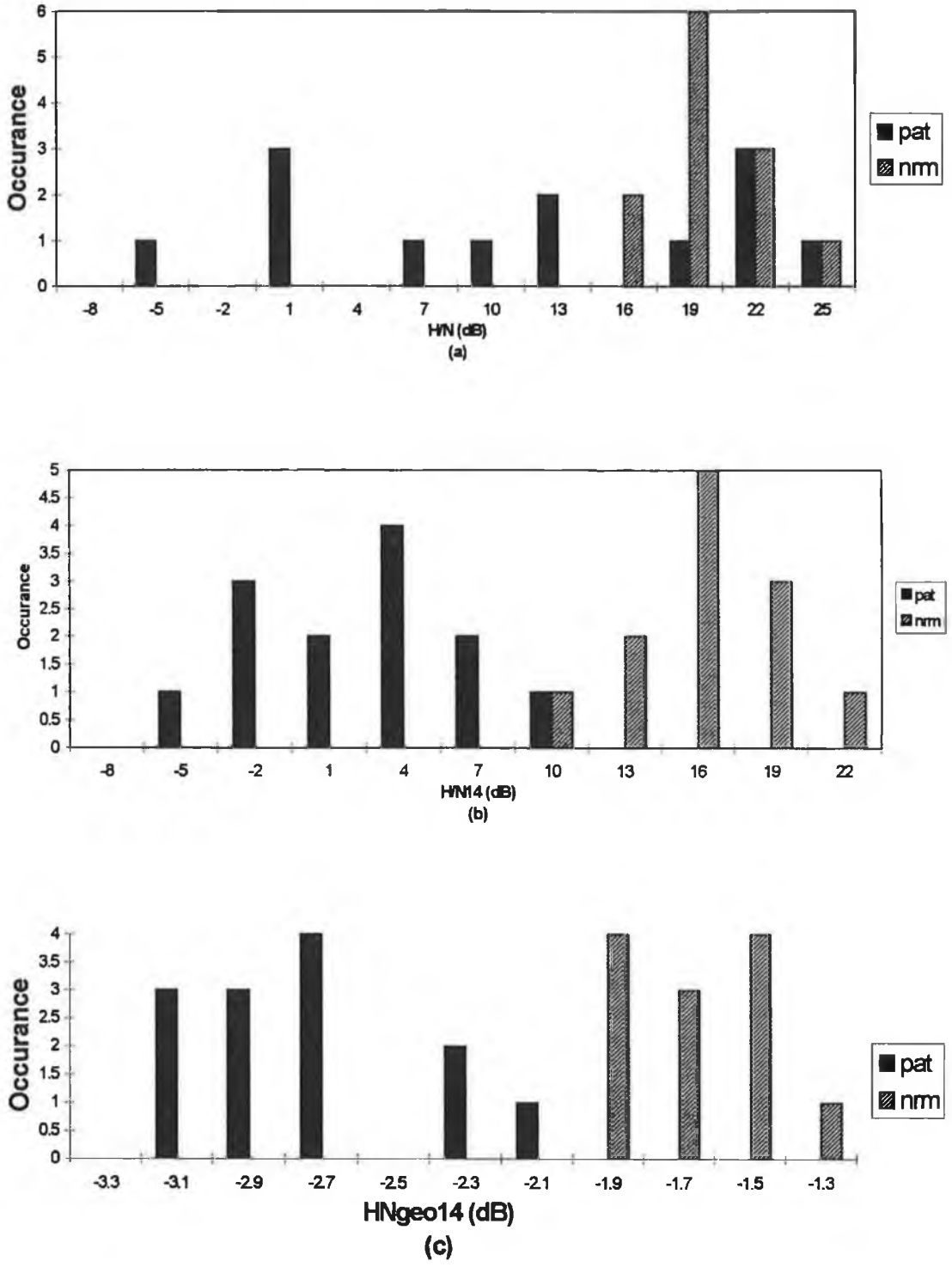


fig.5.42 *Improved separability of the patient/normal data set with respect to the* (a) *H/N ratio, though use of* (b) *H/N14 and* (c) *H/Ngeo14, both of which give results that are highly significant at the 5% level. (HNgeo produced a similar result to HNgeo14).*

5.4.7 Partial Sum of the Fourier Series - Two Cycle Analysis

Two cycles of the waveform were taken and moved on one cycle at a time therefore providing an H/N ratio pitch synchronously. In this manner jitter and shimmer still contribute to the noise estimates. Again, the method is very appealing by nature of it's simplicity. A further advantage is that we can simply go back to the time domain with either the noise or noiseless signals, simply by disregarding every second Fourier coefficient in taking the inverse. The response of the H/N ratio with respect to f0 and for additive noise and perturbation measures was essentially the same as for the Kojima method (fig.5.41). Many new ratios were also investigated, including two dB derived measures and six source or H/Ns based measures including bandlimited versions. The performance of a selection of these measures with respect to the patient/normal data set are shown in fig.5.43 to 5.44. Improved separability is obtained through use of the source based, perceptually based and bandlimited approaches (H/N not shown-poor separability).

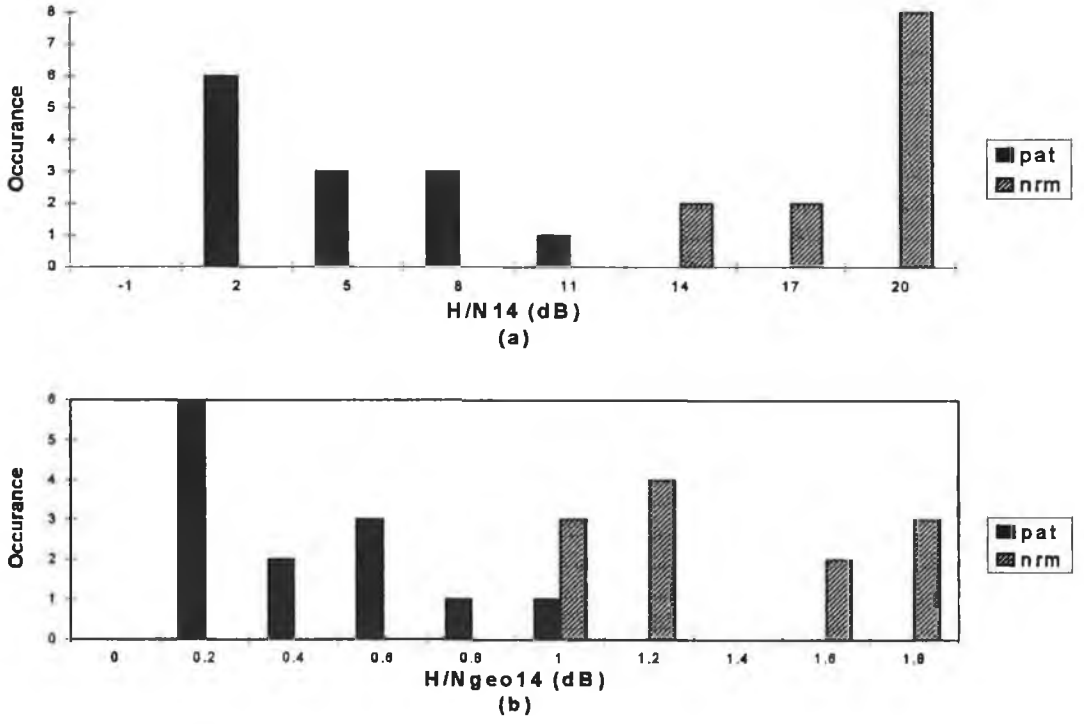


fig.5.43 (a) *H/N14-bandlimited ratio and* (b) *Hngeo14-bandlimited perceptually based ratio for the patient/normal data sets. Both are highly significant at the 5 % level.*

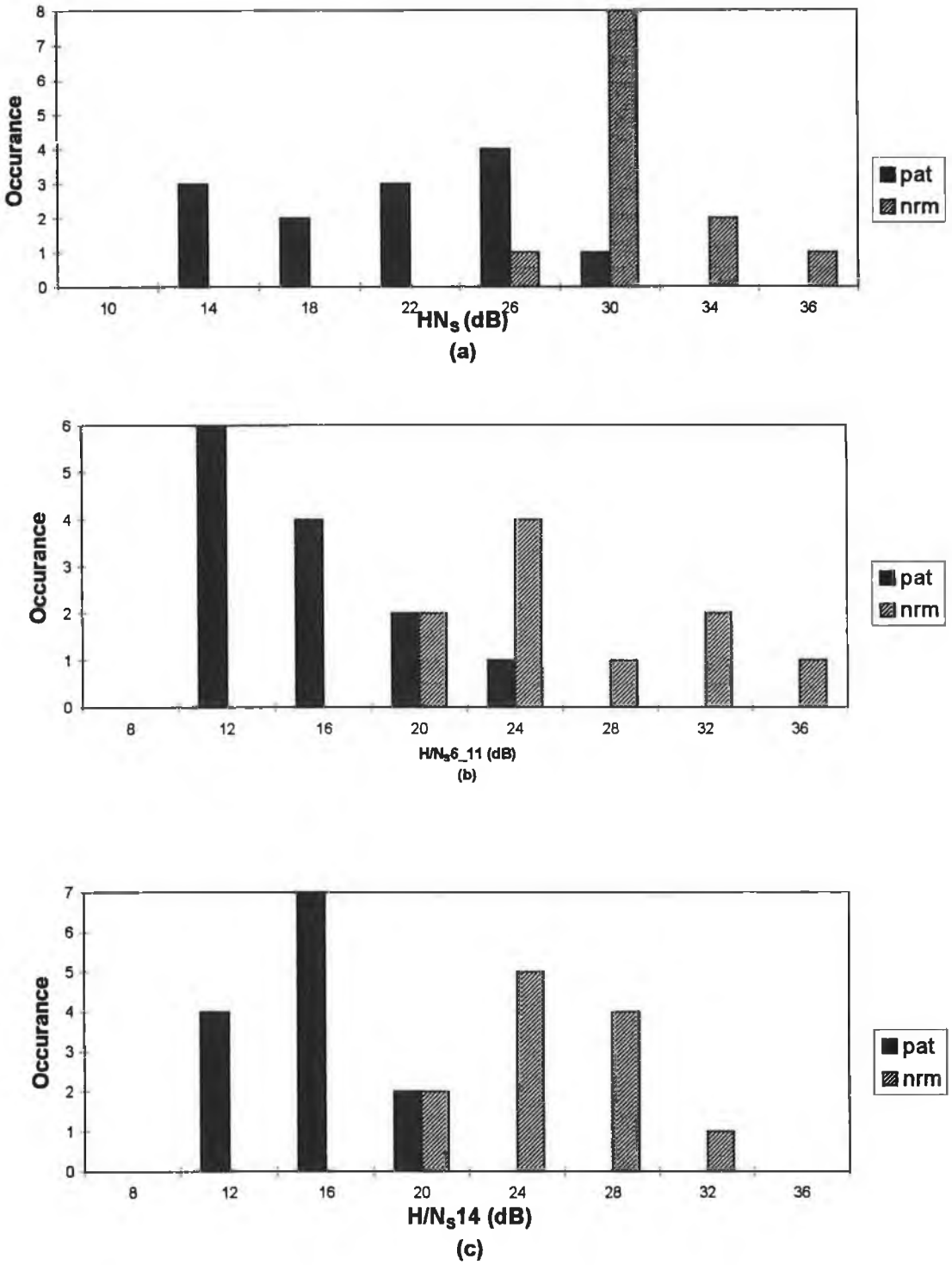


fig.5.44 *Source related H/N ratios* (a) H/N_S, (b) $H/N_{S6\text{-}11}$, *bandlimited according to harmonic number and* (c) H/N_{S14}, *1-4 kHz bandlimit. All measures are significant at the 5 % level (student's t-test).*

5.4.8 Time Domain Averaging - Yumoto et al

The Yumoto paper reported good separability of their patient (12 males/eight females)/normal (22 males/20 females) data and the post surgery improvements also shows good agreement. The H/N ratio for males (average 12.2 dB) did not vary significantly from the H/N ratio for females (average 11.5 dB) and therefore the sets were combined and compared with the patient data. Their values for normals ranged from 7 to 17 dB and pre-op patient data from -15.2 to 9.6 dB, with post-op patient data ranging from 5.9 to 17.6 dB. It is interesting to note that their patient pre-op values went as low as -15.2 dB (~30 times more noise than signal) despite the fact that they could "demarcate pitch periods even in the hoarse voices". This therefore suggests that despite having clearly defined pitch markers, the signal behaved very erratically between these markers i.e. had a very different waveshape from period to period. Although this type of behaviour may occur in some conditions such as spastic dysphonia, the more likely explanation is that tracking errors did in fact occur given the magnitude of the ratio. The response of the H/N index to increases in f0 and the perturbation measures is shown in fig.5.45. Our approach also included a frequency domain analysis from which the geometric dB mean (H/Ngeo) once again proved to be superior to the H/N ratio at separating the patient/normal data. In fact the patient/normal data was completely separated using this method (fig.5.46).

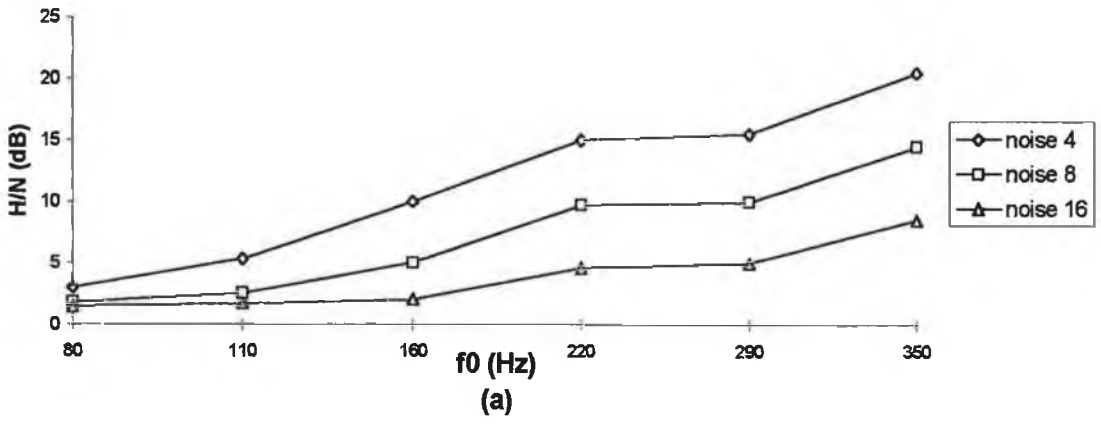


Fig.5.45 (a) *Response of H/N time domain ratio to changes in f0 with three levels of additive random noise with std. dev. 4 %, 8 % and 16 %.*

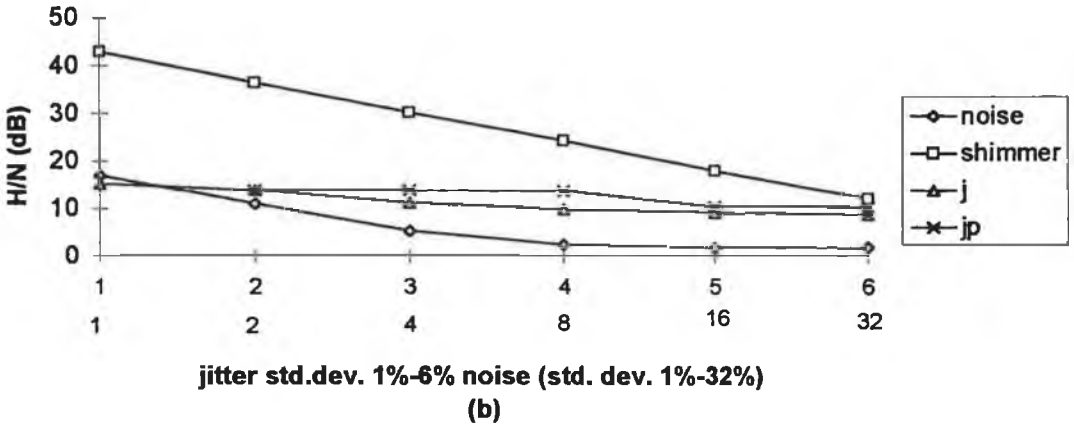


fig.5.45 (b) *Response of the H/N index to the four perturbation measures of random jitter-'j', cyclic jitter-'jp', additive noise and shimmer. Note the reduced sensitivity of the measure with respect to jitter a compared to all previous analyses.*

In part (b) of fig.5.45, the reduced sensitivity of the H/N index is a result of the fact that the method is pitch synchronous and also because the median period was used in estimating the average period (see eqtn.5.37 and eqt.5.38).

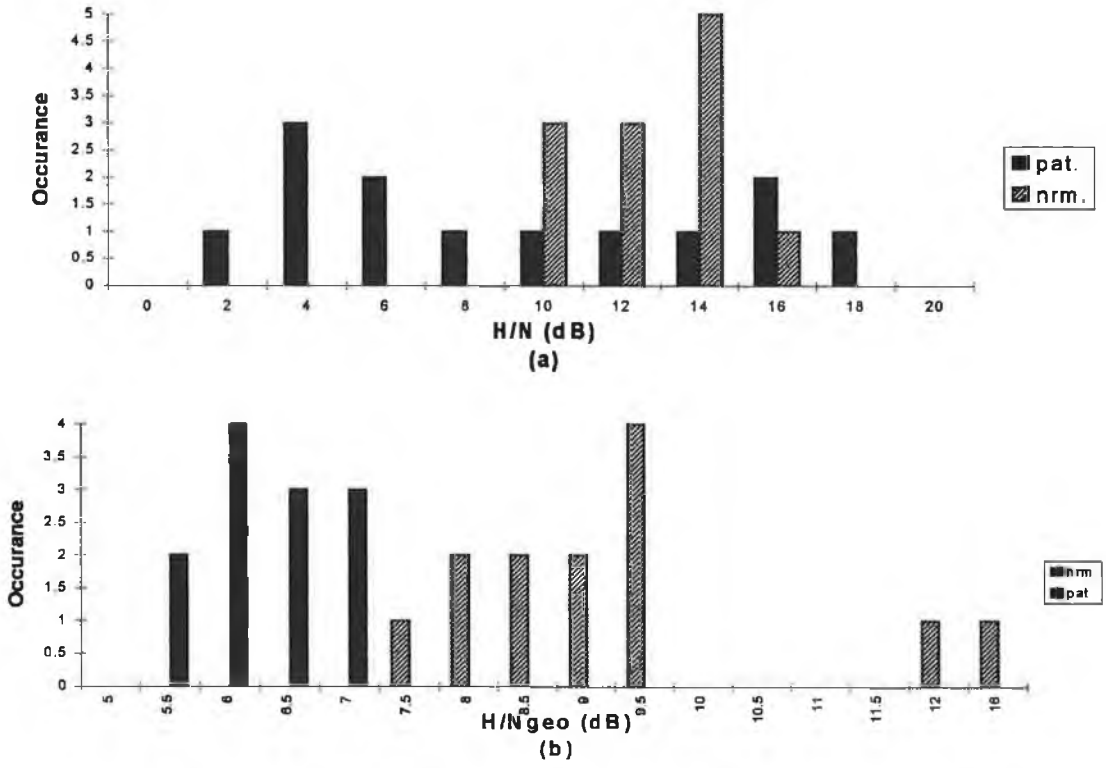


fig.5.46 *Improved separability of patient/normal data sets using* (b) *H/Ngeo (significant at the 5 % level) as opposed to* (a) *the traditional H/N ratio.*

5.4.9 Pitch Synchronous Harmonic Analysis (PSHA)

This novel procedure was designed to provide a measure that is indicative of noise levels in pathological vocal qualities, independent of jitter and shimmer perturbations. Obtaining a spectrum pitch synchronously provides us with several measures including various shimmer measures and distortion factors as well as the usual spectral measures. In addition to this it allows comparison with the frequency domain implementation of the four parameter glottal flow model. Figure 5.47(a) shows the harmonic to noise ratio plotted with respect to f0 for the three levels of additive noise.

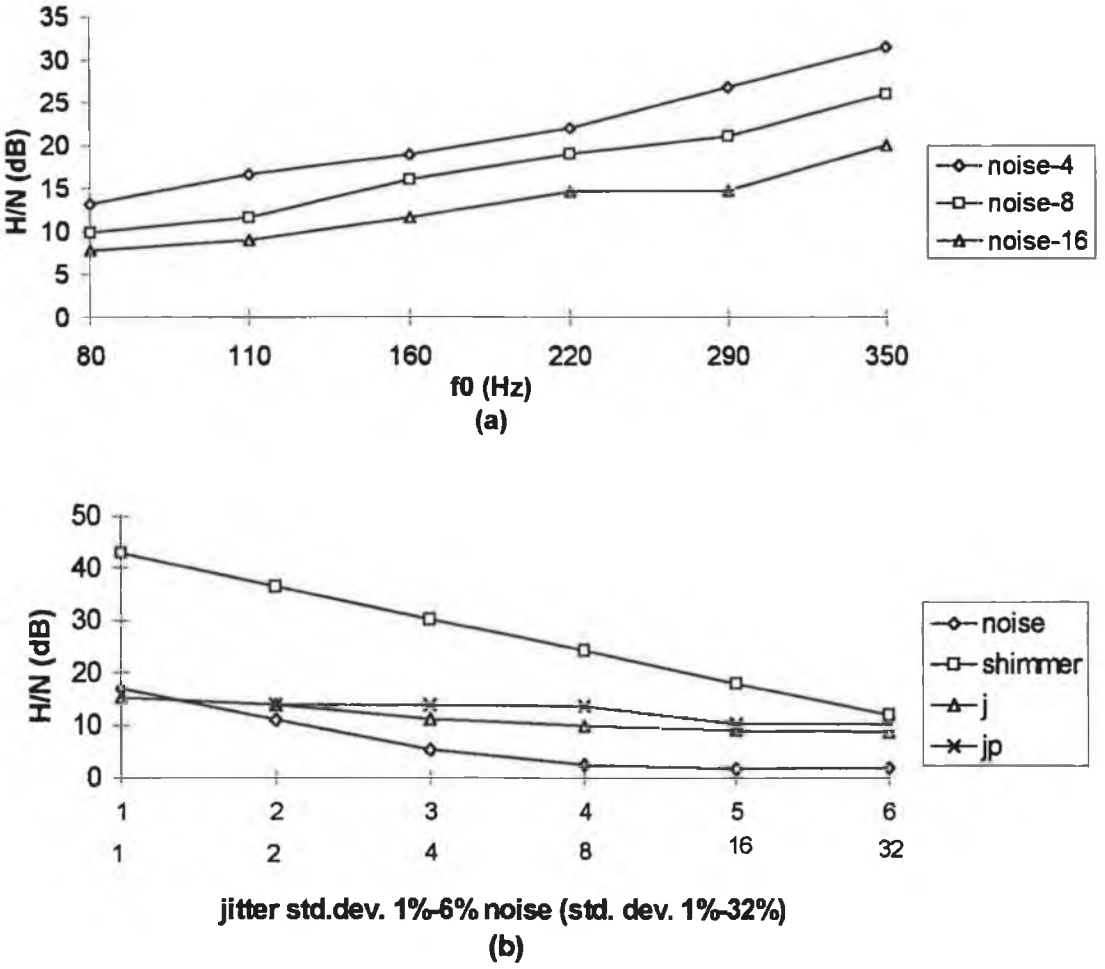


fig.5.47 (a) *The usual f0 trend with approximately equal decrements in the H/N ratio (at a given f0) for increases in noise level.* (b) *The response to the perturbation measures shows a marked improvement on other methods.*

Part (b) of fig.5.47 shows the variation with respect to the perturbation measures. For the worst cases of jitter and shimmer the H/N ratio is still above 10 dB whereas the ratio has reaches this index at ~2% std. dev. random additive noise. The effects of jitter and shimmer on the index has been totally eliminated due to the harmonic formant interaction process as mentioned in section 5.3.9. The index was also tested on the glottal source data in order to examine if the index was truly jitter and shimmer free as hypothesised.

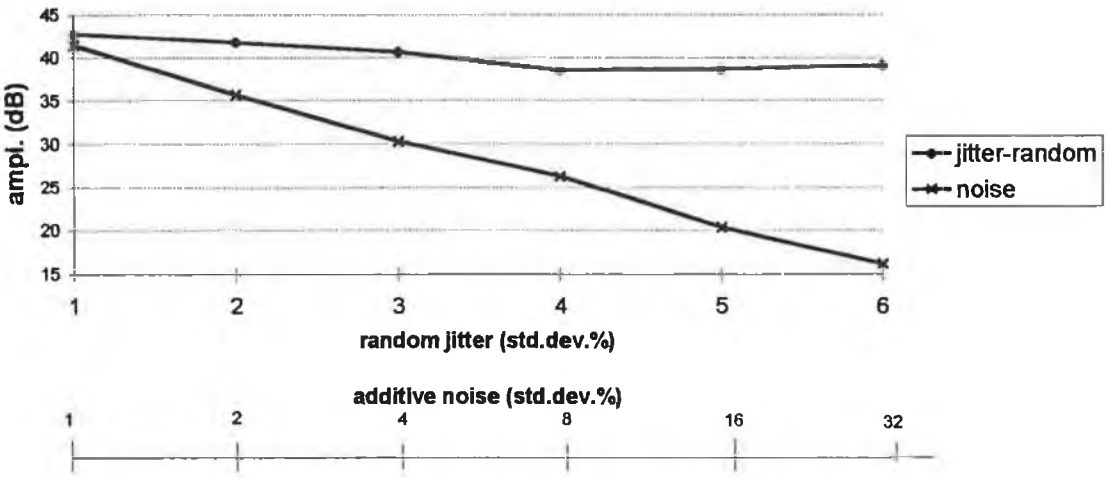


fig.5.48 *The performance of the pitch synchronous harmonic to noise ratio (eqtn.5.51) with respect to random jitter and random additive noise of the glottal source.*

The ratio shows a good linear response with respect to the additive noise levels and the harmonic to noise ratio maintains a level of about 40 dB (with slight variation) up to 6 % std. dev. random jitter. Interpolation of the time domain data, in order to locate the positive zero crossing before the major peak would reduce the slight variability of the index. Shimmer was completely eliminated by normalising the waveform prior to obtaining the spectral estimates and hence gave an infinite harmonic to noise ratio.

For the patient/normal data, the H/N ratio gave poor separability (not shown). The H/N_{14} ratio gave values ranging from 10 to 20 dB for the normal data and from 0 to

7.5 dB for the patient data, therefore completely separating the two data sets (fig.5.49 (a)). The geometric dB mean is shown in part (b) of the figure.

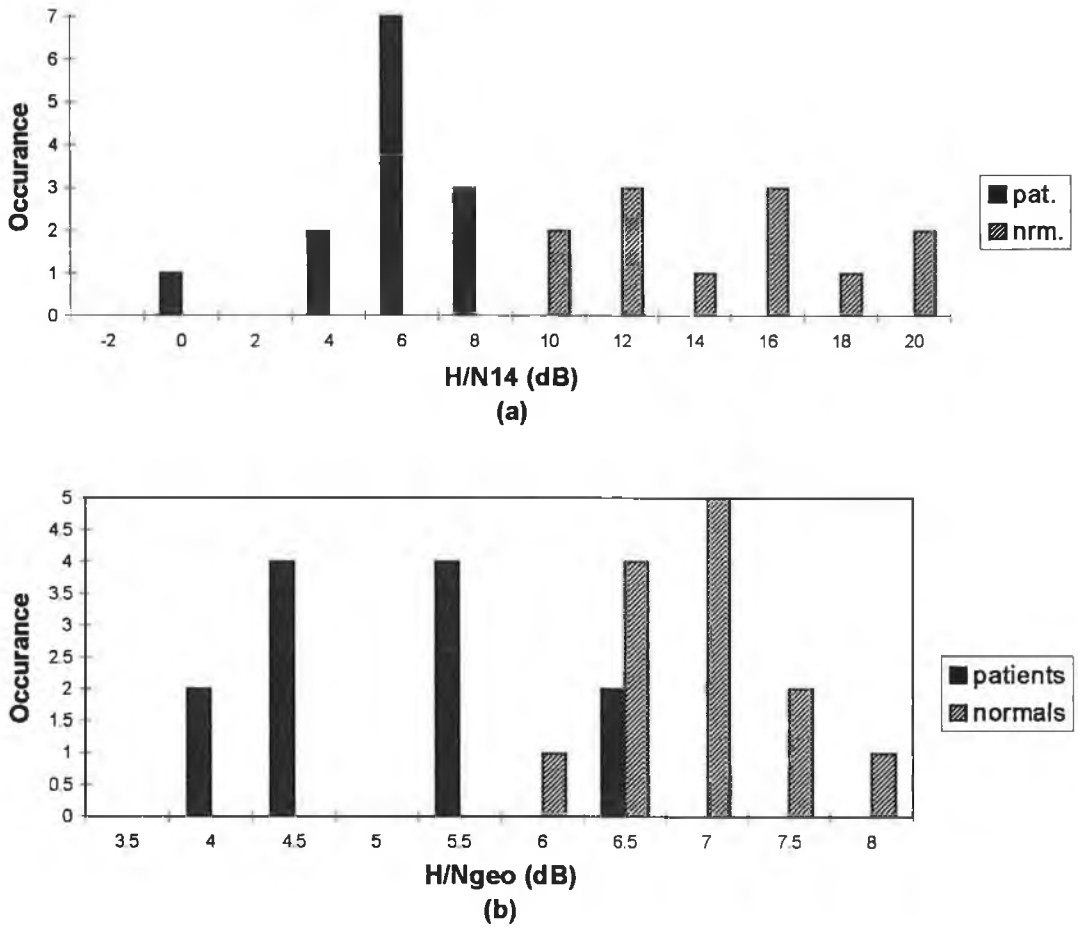


fig.5.49 (a) *Bandlimited and* (b) *geometric dB mean, pitch synchronous harmonic to noise ratios. Both results being highly significant at the 5 % level (two sample, equal variance, student t-test).*

The total harmonic, average percentage amplitude perturbation, THAPAP, defined as

$$\text{THAPAP} = \left[\frac{\sum_{i}^{L} \sum_{j}^{M-1} \left| h_i(T_{j+1}) - h_i(T_j) \right|}{h_{AV}} \right] \qquad \text{eqtn.5.53}$$

where h_{AV} is the mean harmonic value taken over all spectra.

THAPAP is shown for the patient and normal data in fig.5.47(a). Substituting dB values for h_i in the numerator of equation 5.53 and removing the demoninator gives the total harmonic shimmer index (fig.5.47(b)). Other indices such as APAP (average percentage amplitude perturbation-eqtn.5.53 for 1st harmonic) and distortion factor (amplitude of f0 divided by total signal amplitude) were also examined. However, they did not perform well at separating the patient/normal data set.

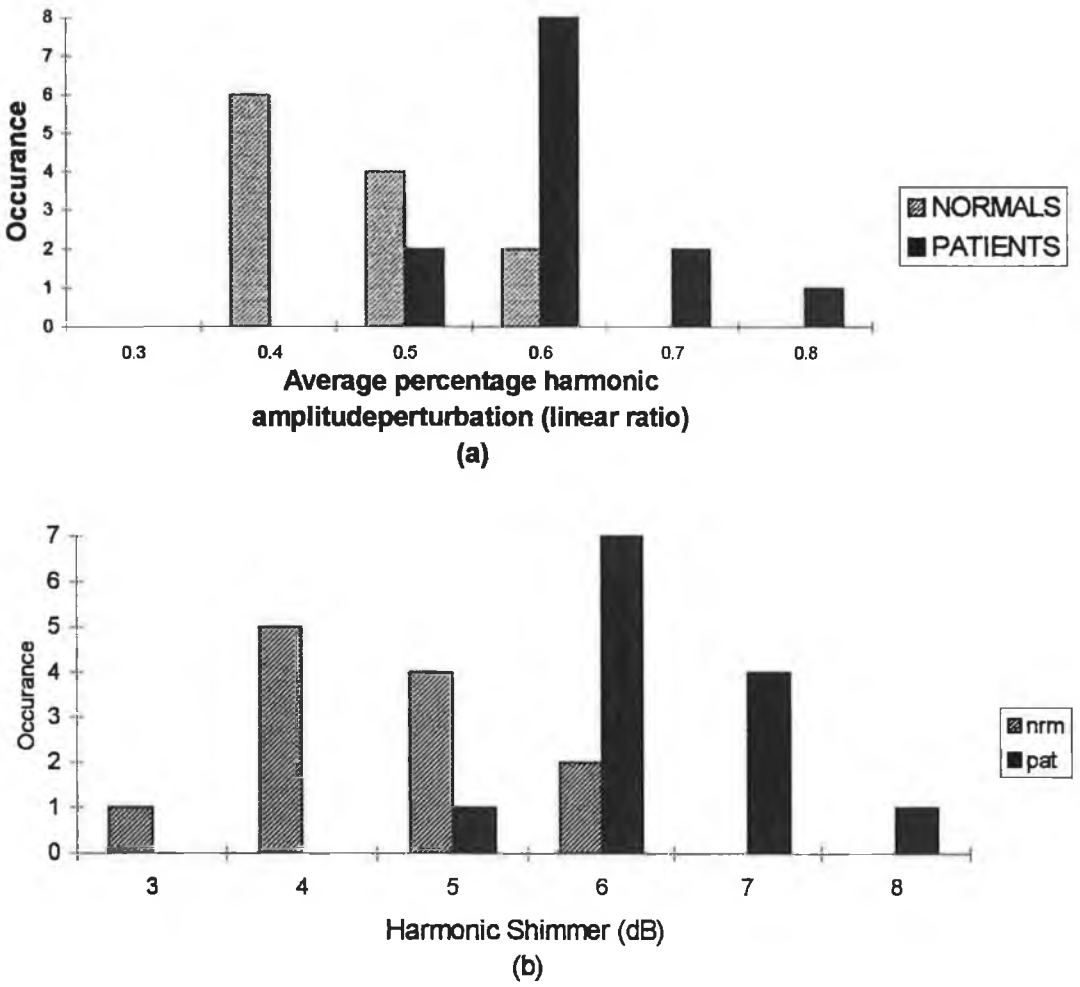


fig.5.50 (a) *(THAPAP) Total harmonic, average percentage amplitude perturbation and* (b) *(TSHM) harmonic shimmer showing good separability of the patient/normal data set, both of which were significant at the 5 % level of the one tailed, equal variance,two sample mean, students t-test.*

5.5 Discussion

5.5.1 Variation of Harmonic to Noise Ratio with Fundamental Frequency for the Synthesis Data : Analysis Considerations

One of the most striking features of the graphs in the results section is the variation of the harmonic to noise ratio with fundamental frequency (f0). All methods except one - the four period, 'pitch synchronous' approach by Muta et al, show this trend of increased H/N ratio with f0 (see fig.5.31 for example). In fact, with the synthesis files used by Muta et al, which came from Titze's SPEAK program[36], an f0 trend was also noticed, although it had a different characteristic to the variation shown here. This variation was simply attributed to the type of synthesis used. In a report which determined the harmonic to noise ratio using the cepstrum technique, de Krom[37] noticed a similar variation of H/N ratio as that encountered here i.e. H/N ratio increased as f0 increased.

In order to investigate possible causes of the f0 trend, the periodogram averaged harmonic analysis program (PAHA-section 5.3.3) was used. Of basic concern in spectral analysis is the resolution required for a certain measurement. Depending on resolution, different characteristics of a signal are revealed. Obvious examples of this in speech analysis are the narrowband and broadband spectrograms, the former resolving the harmonic frequencies and the latter showing more gross characteristics i.e. the formant tract. Due to the coherent addition of the discrete Fourier transform, it is the 6 dB bandwidths that determine spectral resolution, as opposed to the 3 dB criterion of classical signal analysis[12]. Two factors determine whether two signals spaced at given frequency locations will be resolved : (1) the difference in frequency between the signals and (2) the bandwidth of the Fourier estimates. Increased fundamental frequency therefore produces greater separation between the harmonic locations. The window length (and type) determines the bandwidth of the Fourier estimates. Therefore, for a given window length, we might expect different harmonic resolution with f0 variation for the synthesis files and consequently a different H/N

ratio. To test this hypothesis a scheme was developed whereby the ratio of the fundamental to the analysis window length was kept constant. Thus, if the bandwidth limit of the FFT is causing the f0 trend then this approach should produce a flat spectral response .i.e. H/N is equal for all f0s for a given noise level.

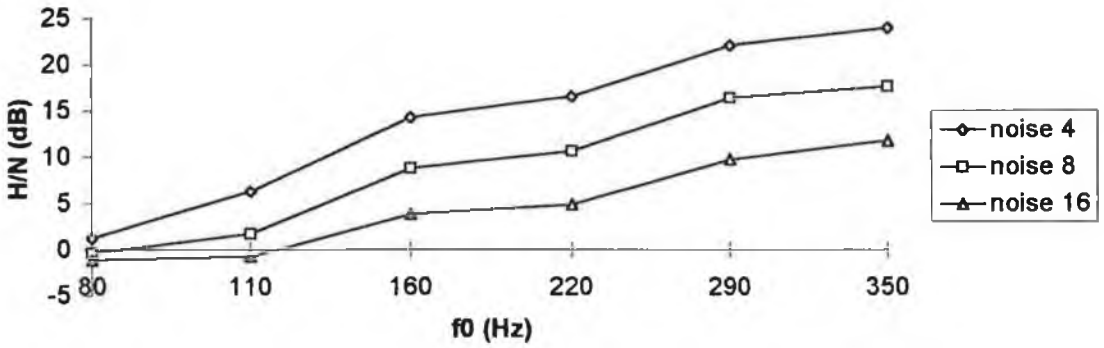


fig.5.51 *Variation of harmonic to noise ratio with f0 with increasing window lengths as f0 decreased. (paha) (see fig.5.31 for comparison).*

Figure 5.51 shows the variation of the H/N ratio with f0 for three levels of additive random noise. The characteristic trend of H/N with f0 is practically unchanged. This is perhaps not unexpected as Kasuya's NNE technique takes seven periods for analysis and hence varies the discrete Fourier transform (DFT) resolution (even though the window length is padded up to 1024 points for all analyses) and still obtains the characteristic f0 trend. Furthermore, taking a 4096 point DFT for a signal sampled at 10 kHz gives a mainlobe width of 8*1024/4096*10000/4096 = 10 Hz (approx.) and is therefore sufficient to resolve even the 80 Hz signal more than adequately. Another argument in favour of the f0 trend not being a bandwidth/resolution effect is that the Yumoto technique, which is based in the time domain gives the same characteristic curve (fig.5.45).

As a result of these findings it was postulated that the trend may in fact be due to a statistical artifact. When noise is added to the source signal, a certain amount of random Gaussian noise, given by a standard deviation of say, 'x' is added. Now, imagine the pitch period is doubled and the same noise of std. dev. 'x' is added to the signal. Looking at this over a single cycle, there is more noise, but correspondingly

more signal, and therefore the signal to noise ratio remains the same. To see why this may not be the case, a single point in a given cycle is considered (fig.5.52). We are considering a random, mean zero signal, therefore the mean of the additive noise component added to this part of the signal is also zero[21].

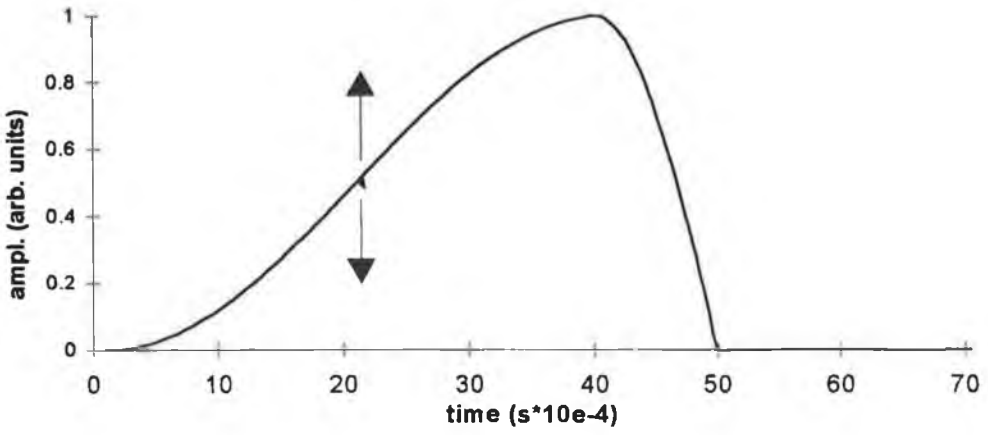


fig.5.52 *Illustration of statistical variation of a single point in the glottal waveform. The variance of a single point is equal to the variance of the signal.*

For a given window length, the average we obtain for the point shown in the figure is better for the higher frequency signals simply because more of them occur in the analysis frame. Two compensatory factors are required. Firstly, the window length is (imagine period length 2:1) doubled to compensate for the two to one ratio of number of periods per window and secondly the number of points per period must be compensated (double again). Therefore, in order to obtain equally accurate estimates of the mean for a signal whose periods differ by a 2:1 ratio we must use a four times longer analysis length for the longer period signal. Neither of these compensations are necessary, of course in the perfectly averaged signal. The hypothesis was checked using the Yumoto technique i.e. the analysis length was determined by the above mentioned statistical and f0 relationship e.g. 160 Hz with 1024 points gives 80 Hz with 4096 points. However, the characteristic f0 trend still remained (fig.5.53) and it was therefore concluded that the data did not in fact require any special statistical considerations i.e. the statistical variability is removed using standard analysis lengths with no special compensatory factors required.

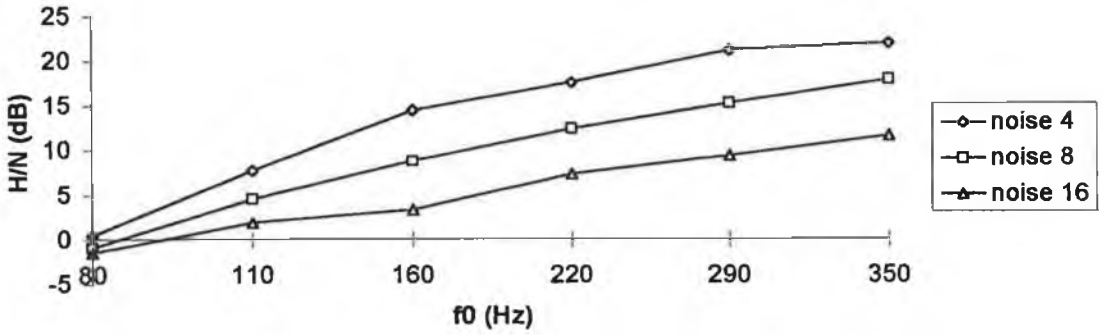


fig.5.53 *Variation of harmonic to noise ratio with f0 with increasing window lengths as f0 decreases in order to compensate for statistical artifacts. (see fig.5.31 for comparison). (Yumoto Technique).*

Further investigations, involving observation of the source spectra (fig.5.54), revealed the true nature of the f0 trend. The H/N ratio is plotted for the three levels of noise versus f0 in fig.5.55 for the glottal source files. It can be seen that the signal to noise ratio of the source data is in fact equal for all frequencies. It is interesting to note that fig.5.3.4, the source derived H/N ratio recaptures this linear characteristic. In one sense therefore it is seen that the H/N variation is due to the synthesis. However, it is simply due to the greater weight given to the higher frequencies as explained in section 5.2.3 as opposed to any peculiarities due to vocal tract filtering.

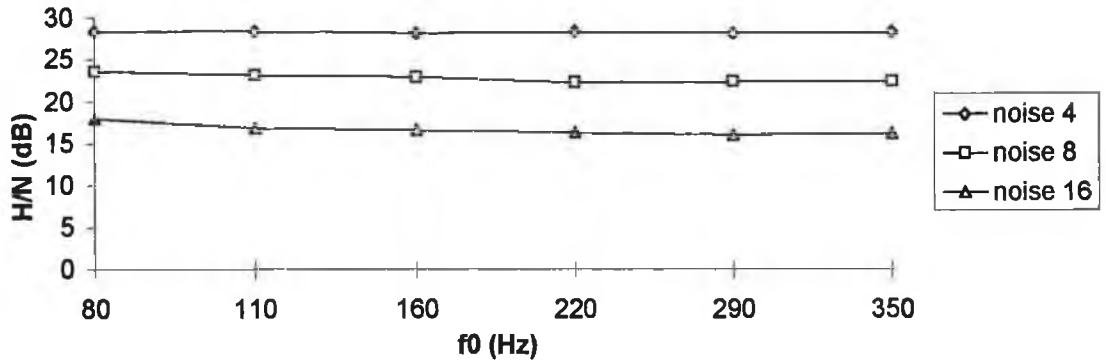


fig.5.54 *Periodogram harmonic to noise ratio vs f0 for the glottal source data. H/N reflects the amount of noise added to the signal at all fundamental frequencies accurately (i.e. ~6 dB reduction for each doubling of noise level).*

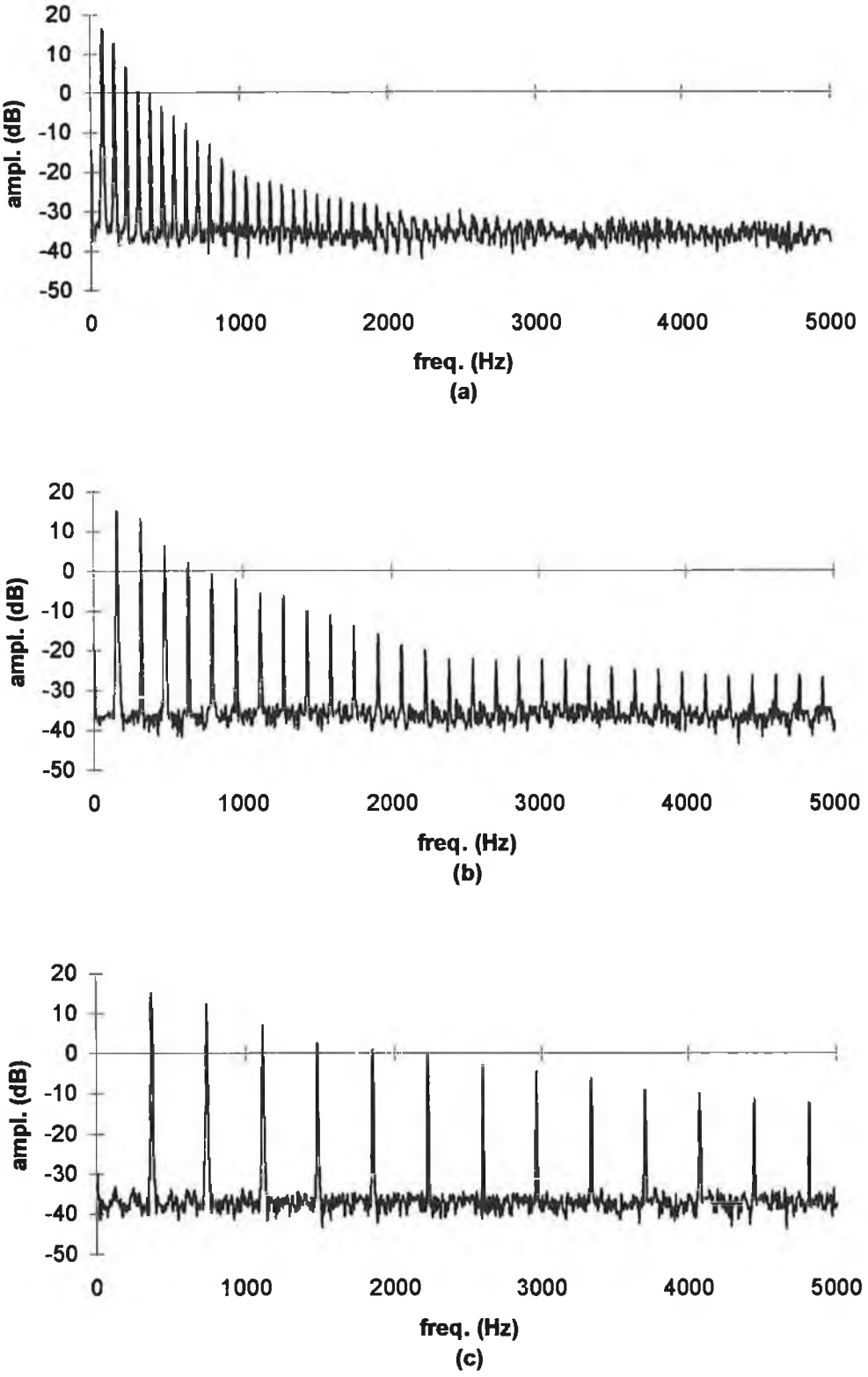


fig.5.55 *Periodogram analysis of glottal source data for* (a) *80 Hz,* (b) *160 Hz and* (c) *350 Hz signal with 4% std. dev. additive noise.*

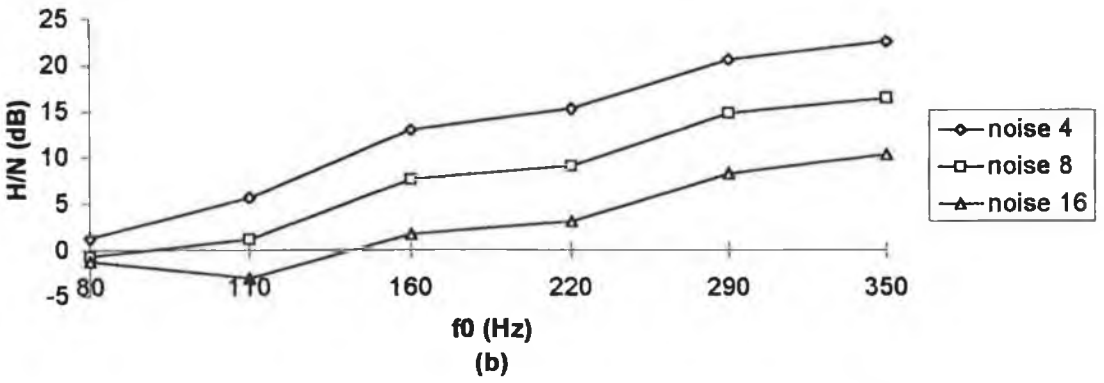


fig.5.56 *Periodogram H/N_{14} ratio for the glottal source data.*

This is clearly illustrated if the H/N_{14} ratio is taken for the glottal source data (fig.5.55). The characteristic f0 trend is regained. Examination of the source spectra for std. dev. 4 % additive noise for the 80 Hz, 160 Hz and 350 Hz source signals helps provide the answer for the observed f0 trend. For the 80 Hz file the lower partials dominate in the calculation of the H/N ratio. However, when the signal is bandlimited from between 1 and 4 kHz the resulting H/N ratio is greatly reduced. Considering the 350 Hz file, the harmonic frequencies are still very prominent in the 1-4 kHz range, only giving a slight reduction in the H/N ratio.

The basis behind this occurrence was examined in section 5.3.9 when 'scaled jitter' was investigated. The different source signals with different fundamental frequencies can be viewed as scaled versions of each other and therefore their relative harmonic strengths are equal. The first fourteen harmonics for the 80 Hz file have occurred by 1120 Hz whereas the first fourteen harmonics for the 350 Hz file span the complete frequency range up to 5000 Hz. The harmonic to noise ratio is compared according to harmonic number in fig.5.56 using pitch synchronous harmonic analysis (psha). The response with f0 is almost flat. Therefore, as was stated when discussing the Muta et al technique, there is considerable benefit in considering the signals according to harmonic number as opposed to frequency range. An obvious objection to this is that the formant frequency locations differ only by 25 % for male and female speakers whereas their pitches have an octave difference. Therefore for the output radiated speech waveform a set frequency range is probably more appropriate. However, for

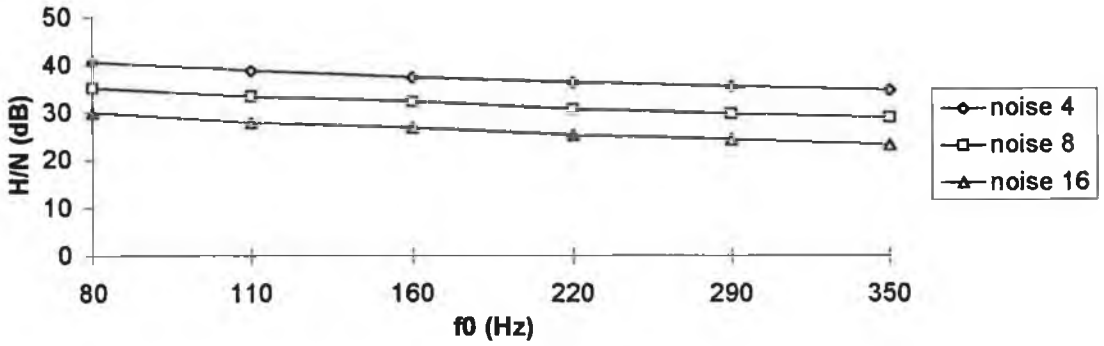


fig.5.57 *Periodogram H/N ratio taking the first 14 harmonics for the glottal source data. Approximately linear but the lower frequencies have slightly higher H/N due to higher sampling.*

inverse filtered or other source related data, analysis by harmonic number is the preferred approach. Note that in fig.5.55 a spectrum equivalent to the 80 Hz file could have been obtained for the 350 Hz file if it had been sampled at a frequency that maintained the same ratio between sampling rate and pitch period (as for the 80 Hz file). An appropriate sampling frequency is therefore another issue for consideration. Interpolation could also be used. Kasuya[38] has shown that a high sampling frequency (40 kHz) is required in order to capture the high frequency noise components accurately. Twenty kHz seems a reasonable compromise between excessive data and reasonably accurate determination of the signal. The benefit (H/N ratio unchanged) of analysing by harmonic number echos what was stated in section 5.3.9 regarding scaled signals. It can be seen therefore that the present consideration regarding H/N variation with f0, the problem of jitter and shimmer and the question of inverse filtering, all have a common solution in the form of pitch synchronous harmonic analysis.

5.5.2 Comparison of Analysis Techniques Based on Spectral Characterisation of Perturbation with Inferences for Future Development of Quantitative Analysis

All of the Fourier techniques (series and transform), except the pitch synchronous harmonic analysis approach (paha), show considerable overlap for the H/N ratio values reported for jitter and additive noise. In general the methods are somewhat less sensitive to shimmer i.e. they reflect shimmer levels accurately. The harmonic to noise ratios for seven of the techniques (Kitajima method not shown because scale is different and two cycle analysis omitted because of it's similarity to the Kojima technique) are shown plotted with respect to the four perturbation measures in figures 5.58 to 5.61. The trend of the harmonic to noise ratio with respect to the perturbation measures is readily explained by referring to the spectral characterisation development in section 5.2.2.

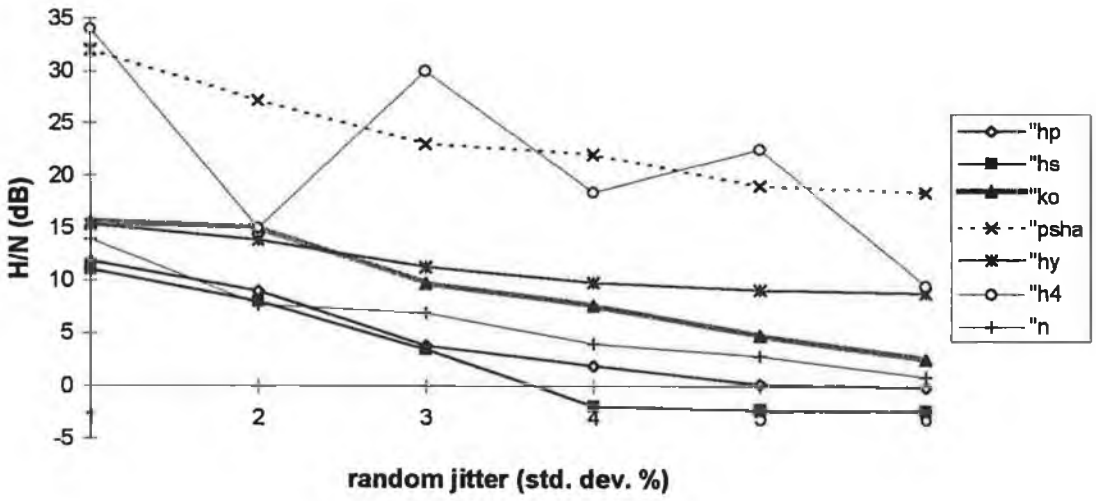


fig.5.58 *Response of seven of the analysis techniques to random jitter. The pitch synchronous harmonic analysis (psha) technique and modified Yumoto (hy) technique show least sensitivity to jitter. hp-periodogram, hs-Hiraoka, ko-Kojima, h4-Muta, n-Kasuya's NNE (inverted).*

The harmonic structure for the random jitter signal is completely missing even for 2 % standard deviation jitter. Therefore, in estimating the harmonic levels the programs acquire reduced energy values at n×f0 locations and because the 'noise' energy in jittered signals follow the signal properties the energy at between harmonics is of a comparable level to the energy at harmonic locations. Both of these effects contribute to reduced harmonic to noise ratio estimates. Close examination of the spectra for the random jitter signals reveals that small amounts of periodicity reappears in the signal at locations determined by the standard deviation of the jitter and the actual fundamental frequency present. Therefore, some measurement reflecting the reappearance of periodicity would reveal whether the reduced harmonic to noise ratio was actually due to increased levels of noise or increased levels of jitter.

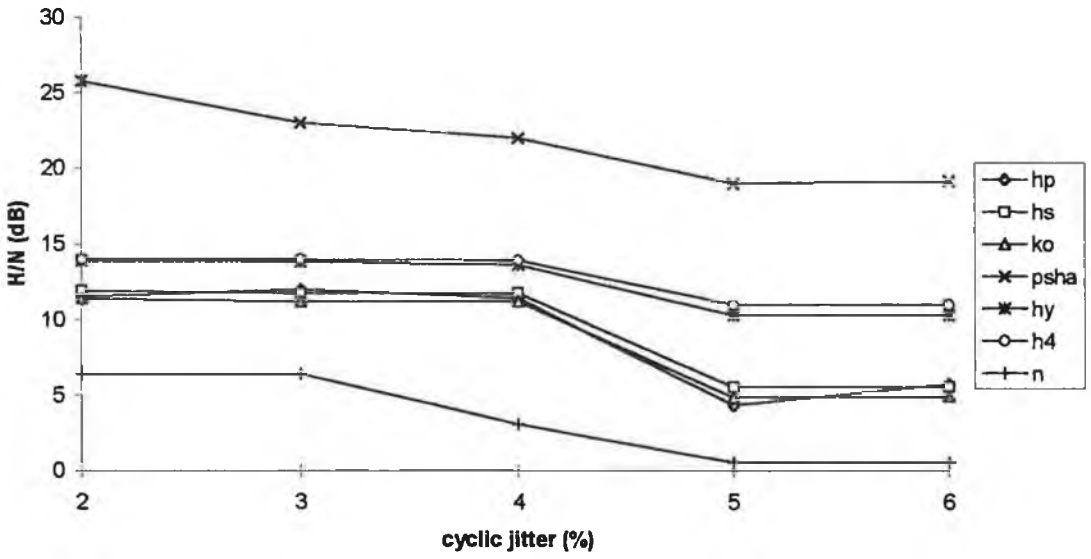


fig.5.59 *Response of seven of the analysis techniques to cyclic jitter. The pitch synchronous harmonic analysis (psha) technique, modified Yumoto (hy) technique and Muta method show least sensitivity to cyclic jitter. hp-periodogram, hs-Hiraoka, ko-Kojima, h4-Muta, n-Kasuya's NNE (inverted).*

For cyclic jitter, which is a cardinal symptom of 'creaky' vocal quality, subharmonics appear in the spectrum. As stated by Fujimura[39] if there is a discontinuous shift in fundamental frequency as sometimes evidenced in creaky voice production, traditional pitch trackers will try to fit a smoothed curve between f0 estimates. This therefore

does not reflect the source of the perturbation very accurately. An alternative approach is to base the pitch extraction on the spectral properties of the perturbation as suggested by Fujimura. Equations 5.10 and 5.11 provide the basis for quantifying the characteristics of the subharmonic regimes. Further developments, involving the application of these equations to successive spectra could provide an indication of the onset and offset of subharmonic production. Note amplitude modulation or cyclic shimmer would similarly produce subharmonic regimes.

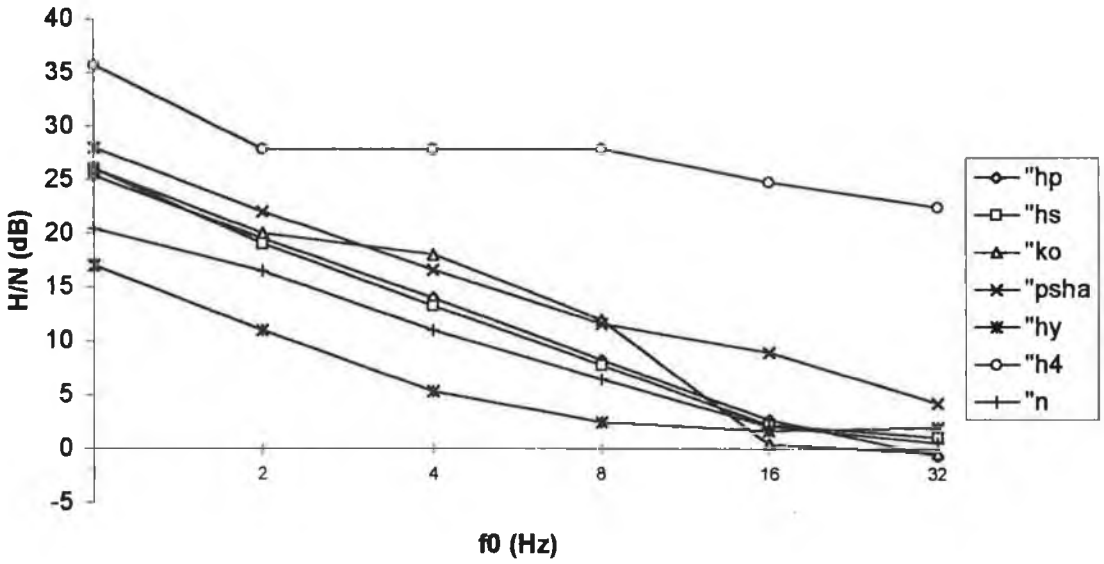


fig.5.60 *Response of seven of the analysis techniques to additive noise. All methods reflect the noise levels accurately except for the Muta technique. hp-periodogram, hs-Hiraoka, ko-Kojima, h4-Muta, n-Kasuya's NNE (inverted).*

All analysis programs reflect the level of additive noise accurately, but as we have seen in section 5.2.2, the noise levels introduced using random Gaussian noise are independent of the signal properties and also have a flat frequency characteristic which moves upwards (with respect to amplitude). Therefore all methods give false estimates of harmonic energy at n×f0 locations due to noise contributions at these locations. Observation of the noise spectra motivates other possible strategies for differentiating between additive noise and jitter e.g. a calculation of noise i.e. spectral estimates at (n+1/2)×f0, in the upper frequency region would reveal the source of the perturbation.

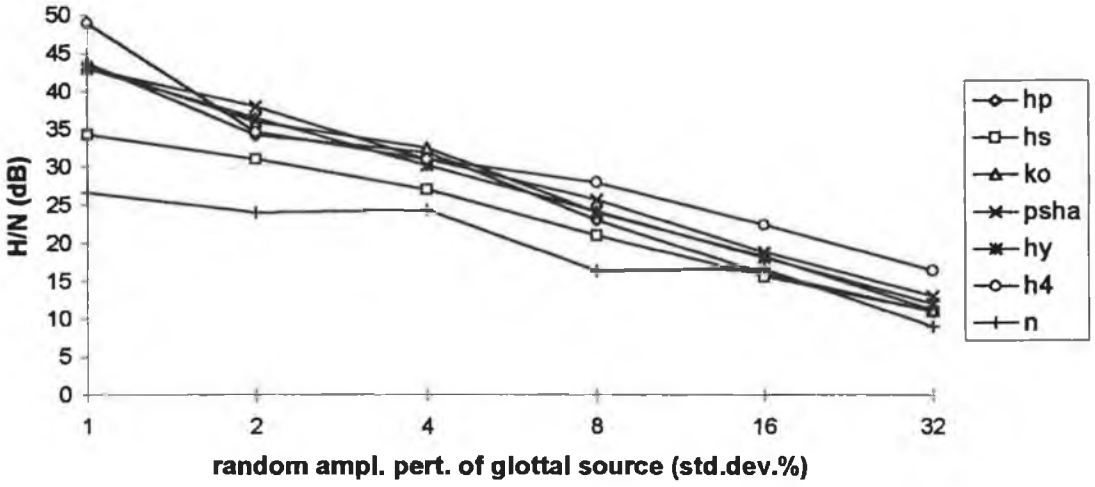


fig.5.61 *Response of seven of the analysis techniques to shimmer. All methods reflect the shimmer levels accurately except for the NNE technique which shows some variability. hp-periodogram, hs-Hiraoka, ko-Kojima, h4-Muta, n-Kasuya's NNE (inverted).*

The analysis programs reflect the levels of shimmer accurately. All methods show a linear (i.e. dB is linear with respect to doubling of noise) response to noise levels as expected. A scheme for detecting shimmer should check to see if the H/N ratio is constant for all frequencies. To compare the sensitivity of a given technique to jitter relative to it's sensitivity to additive noise the response of that technique should be compared by viewing fig.5.58 and fig.5.60. The pitch synchronous harmonic analysis technique and the modified Yumoto technique show least relative (to additive noise) sensitivity to random jitter.

5.6 Conclusion:

If the ultimate goal is to detect or specify the vibratory pattern of the vocal folds from spectral measures then the contaminating effects of jitter, shimmer and additive noise must be removed. Presently used harmonic to noise ratio methods provide a 'catch all' criterion in evaluating vocal pathologies i.e. the presence of jitter and shimmer contribute to the reduced harmonic to noise ratios. An advantage of this fact is that it might be useful for characterising the overall state of the voice. However, it reduces the specificity of the measure in terms of describing laryngeal activity.

This problem was addressed using a pitch synchronous harmonic analysis (psha) technique. The sensitivity of the resulting H/N ratio to jitter was much less than for other non-pitch synchronous methods. Similar reductions in jitter sensitivity were obtained using an adaptation of the Yumoto technique which employed a median period for the signal averaging scheme (eqtn.5.38). Jitter was not completely removed due to harmonic-formant interactions. However, when 'psha' was applied to the glottal waveform the effects of jitter were almost completely removed. Shimmer contributions to the H/N ratio were entirely eliminated in this manner. The advantage of frequency domain scaling was introduced and it was suggested that more complicated variations in pitch period could similarly be accounted for by pertinent frequency domain adjustments.

Basic research is required in order to characterise the nature of pitch perturbation and resultant glottal flow characteristics. Examination of pitch synchronous inverse filtered glottal spectra and waveforms for patients with high jitter scores would lead to better classification of the anomaly. Gobl[40], Karlsson[41] and others have matched inverse filtered glottal flow waveforms with the flow parameters of LF model in order to characterise the flow. Recent research has attempted to develop a frequency domain parameter set of glottal flow which accepts tape recorded speech data and therefore avoids the need for phase sensitive recording which can be quite laborious and give discomfort to the patient. The development of the 'psha' technique, therefore, is a convenient compliment to these research efforts. Furthermore, a model for f0 control

based upon biomechanical considerations introduced by Titze[42] and expanded upon by Farley[43] has emerged. The ultimate combination of these research efforts would provide acoustic indices that relate to specific physiological function. The importance of this in respect to the present study is that it would help to provide more differential diagnoses and anatomical specific characterisations of vocal pathology. To this end, accurate determination of flow characteristics from phyiological function seems to be a particularly important area for consideration.

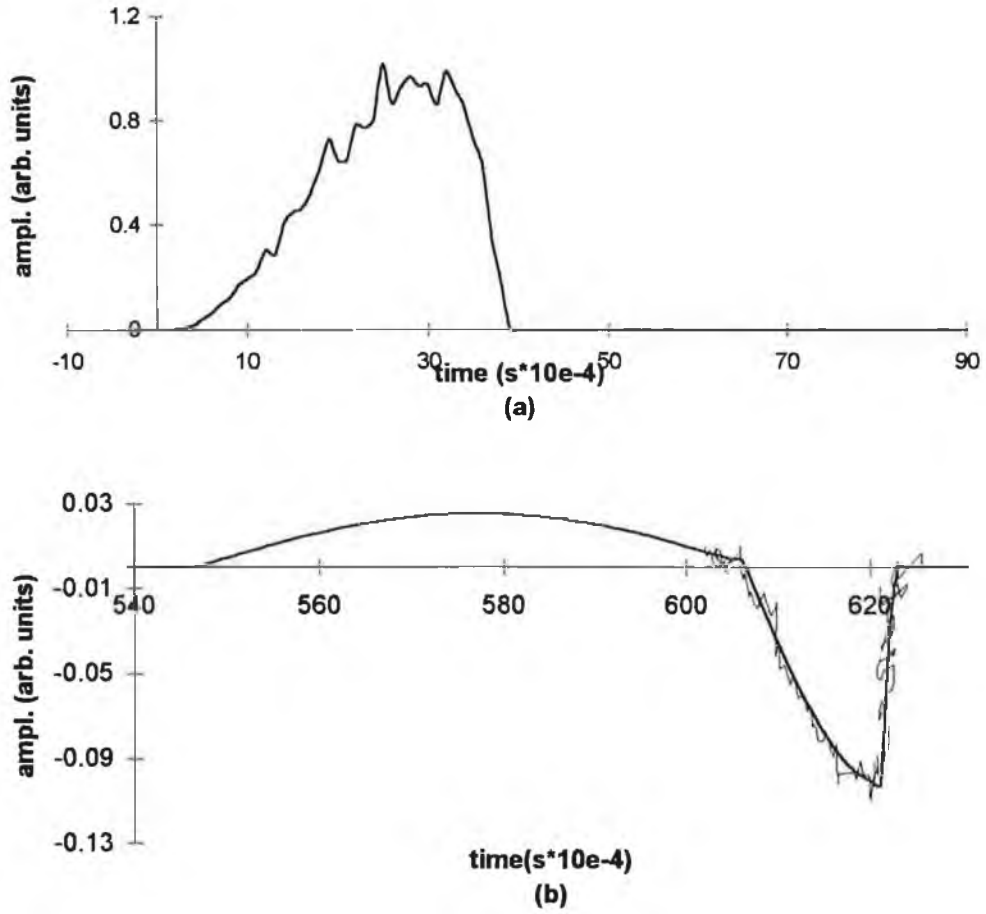


fig.5.62 (a) *Glottal Flow and* (b) *flow derivative indicating two possible conditions for turbulent noise generation*

Another facet requiring attention is an accurate determination and characterisation of the turbulent noise found in pathological voices. At present two distinct mechanisms seem to exist. The first involves turbulence concurrent with the moment of maximum glottal flow[9] (fig.5.62 (a)) and the second corresponds to turbulence associated with the moment of maximum glottal excitation[44,37] i.e. corresponding to the negative peak

in the flow derivative (fig.5.62 (b)). de Krom[37] states that "the energy of turbulent noise is inversely proportional to the cross sectional area of the glottal slit", while Stevens[9] reports "the amplitude of the turbulent noise at the glottis is expected to increase approximately in proportion to $A_g^{0.5}$, where A_g is the average glottal area during a cycle of vibration". This apparent contradiction is readily explained by considering the Reynolds number (eqtn.1.1), where a reduced glottal width or increased airflow (and corresponding increase in glottal width) can give rise to turbulent flow. Imaizumi[45] has observed turbulent noise both characteristics using a comb filtering technique[46] to extract the noise component. Yet another condition for turbulent flow is satisfied for patients containing mass lesions or nodules.

The development of alternative ratios to the traditional harmonic to noise ratio merits further study requiring more refined ratios and accurate perceptual determinations in order to provide correct correlations. The perceptual examination of the patient and normal data set for the present study (Table.2.3), rated all patients as dysphonic (degree not specified) and all the 'normal' group were rated 'normal' except for two who showed mild hoarseness and breathiness. Therefore, in the absence of a graded scale the measure could not be accurately assessed. However, in consideration of the fact that the geometric dB mean consistently showed good separability of the patient/normal data set we suggest that the ratio shows much promise and that further research is definitely merited. Similar arguments hold true for the source related ratio, requiring correlation to accurately determined physical data such as EEG or stroboscopy ratings. We quote Rothenberg[33] on the need for such measures:

"In general there are two types of purposes, namely, to measure or explicate the physiological basis of the voice characteristics (often the physician's goal) or to measure or explicate the perceptual characteristics of the voice (often the therapist's goal). For example if the vibratory characteristics of the vocal folds were of interest, a measure emphasising the periodicity in the fundamental frequency and lower partials might be of interest. If, on the other hand, the perceived pitch were of primary interest, the periodicity in the energy near the first formant might be a better measure."

Finally, although a total of nine separate methods were examined for determining the harmonic to noise ratio this list is by no means exhaustive. Other methods have been proposed by the following authors: Ladefoged[34] (time domain), Milenkovic[47]

(time domain-comb filter), Imaizumi[46] (time domain-comb filter (see appendix A for source code)), Kasuya (periodicity model)[48], deKrom[37] (cepstrum), Qi[49] (wavelet) and Gavidia-Ceballos[50]. The best strategy for introducing new methods is to have a specific goal in mind. The motivation for each of the three methods introduced here followed from independent objectives. The two cycle analysis was introduced in order to spectrally characterise perturbation, the periodogram method was introduced to motivate the idea of a noise signal characterising a system and the possibility of source derived ratios (as well as providing more consistent spectral estimates). The pitch synchronous harmonic analysis approach was developed in order to provide a jitter (and shimmer) free measurement of the harmonic to noise ratio. It is hoped that these techniques, particularly the later, will prove complimentary to the research efforts of others e.g. Hanson[51], Holmberg[52], deKrom[53] who have begun to take more diverse spectral measurements from the acoustic spectra. It is suggested that the 'long term harmonic spectrum' ($h_{i(AV)}$-eqtn.5.51) may provide more accurate H1(amplitude of first harmonic) to H2 (amplitude of second harmonic) ratios.

To surmise,

1. A general definition of noise has been discussed.
2. The spectral properties of random jitter, cyclic jitter, shimmer and additive noise have been characterised based on Fourier series, Fourier transform and periodogram estimation.
3. The harmonic to noise ratio of the output radiated speech waveform has been related to the harmonic to noise ratio of the source.
4. New ratios have been proposed and tested that relate more specifically to source and perceptual information regarding the voice.
5. Six presently available methods for determining the harmonic to noise ratio have been successfully programmed and tested with many alterations introduced.
6. Three new methods for determining the harmonic to nose ratio have been introduced, namely, periodogram averaging, two cycle analysis and pitch synchronous harmonic analysis.

7. Each new method was developed with a particular emphasis in mind.

   Two Cycle - to examine the spectral characteristics of perturbation.

   Periodogram - to show how random noise can be used to characterise a system.

   Pitch Synchronous Harmonic Analysis - to provide a jitter and shimmer free estimate of the harmonic to noise ratio.

8. The results from all (except two) of the analysis techniques show that jitter and shimmer are included in H/N ratio measurements. The modified Yumoto approach and the specifically developed pitch synchronous harmonic approach (psha) showed considerable less sensitivity to the jitter and shimmer artifacts. The 'psha' method was jitter and shimmer insensitive for the source data.
9. The 'psha' approach showed a lot of promise and many possible developments were suggested. The technique is complementary to similar methodologies employed by Fant et al.
10. The results of the various analyses show that presently used H/N ratios are useful in determining abnormal voice, especially if the H/N ratio is bandlimited.
11. The variation of the H/N ratio for the synthesis data with f0 for three levels of additive noise has been explained. The use of harmonic number as opposed to frequency location seems to be merited for studying source characteristics.
12. Quantitative spectral measurements have been proposed based on the spectral characterisation results.
13. Future research directions have been considered.

5.7 Bibliography

1. Hillenbrand, J. A methodological study of perturbation and additive noise in synthetically generated voice signals. J. Speech and Hear. Res. 1987; **30**: 448-461
2. Bielamovicz, S. Comparison of voice analysis systems for perturbation measurement. J. Speech Hear. Res. 1996; **39**: 126-134
3. Yanaghihara, Significance of harmonic changes and noise components in hoarseness. J. Speech Hear. Res. 1967; **10**: 531-541
4. Martin, D. et al Pathological voice type and the acoustic prediction of severity. J. Speech Hear. Res. 1995; **38**: 765-771
5. Pruszewicz, A. et al Usefulness of acoustic studies on the differential diagnostics of organic and functional dysphonia. Acta. Otolaryngol. 1991; **111**: 414-419
6. Valencia-Naranjo, N. et al Diagnostic voice disorders. Current opinion in head and neck surgery 1995 **3**: 164-168
7. Titze, IR. Towards standards in acoustic analysis of voice. NY: Raven Press J. Voice 1994; **8**: 1-7
8. Rontal, E. et al Objective evaluation of vocal pathology using voice spectrography. Ann. Otol. 1975; **84**: 662-671
9. Stevens, KN. Airflow and turbulent noise for fricative and stop consonants , J. Acoust. Soc. Am. **50**:1180:1192
10. Muta, H. et al Analysis of hoarse voices using the LPC method. Laryngeal function in phonation and respiration. Boston : Little Brown, 1987 pp. 463-474
11. Titze, IR. and Liang. H. Comparison of f0 extraction methods for high precision voice perturbation measurements. 1993; **36**: 1120-1133
12. Harris, FJ. On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. of the IEEE, 1978 **66**:51:142
13. Blackman, RB. and Tukey, JW. The measurement of power spectra. Toronto, Ontario: General Publ. Company, 1959
14. Klingholtz, M. and Martin, F. Quantitative spectral evaluation of shimmer and jitter. J. Speech Hear. Res. 1985; **28**: 169-174

15. Gauffin, J. et al Irregularities in the voice: A perceptual experiment using synthetic voices with subharmonics. In PJ Davis and NH. Fletcher (Eds.) Controlling complexity and chaos. San Diego: Singular Publ. Co., 1996

16. Fant, G. Acoustic theory of speech production. Mouton, The Hague, 1970.

17. Deller JR. et al Discrete time processing of speech signals, New York:Macmillan, 1993

18. Fant, G. and Lin, Q. Frequency domain interpretation and derivation of glottal flow parameters. STL-QPSR 1988, **2-3**:1-23

19. Kitajima, K. Quantative evaluation of the noise level in the pathologic voice. Folia phoniat. 1981; **33**: 115-124

20. Hiraoka, N. et al Harmonic intensity analysis of normal and hoarse voices. J. Acoust. Soc. Am. 1984; 76: 1648-1651

21. Oppenheim AV. and Schafer RW. Discrete-time signal processing. Englewood Cliffs, N.J.: Prentice Hall, 1989

22. Jenkins, GM. and Watts DG. Spectral analysis and it's applications. San Francisco: Holden-Day, 1968

23. Kasuya, H. et al Normalised noise energy as an acoustic measure to evaluate pathologic voice. J. Acoust. Soc. Am. **80**: 1329-1334

24. Muta, et al A pitch synchronous analysis of hoarseness in running speech. J. Acoust. Soc. Am. 1988; **84**: 1292-1301

25. Kojima, H. et al Computer analysis of hoarseness. Acta. Otolaryngol. 1980; **89**: 547-554

26. Yumoto, E. et al Harmonic-to-noise ratio as an index of the degree of hoarseness. J. Acoust. Soc. Am. **71**: 1544-1549

27. Hammarberg, B. and Gauffin, J. Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects. In O. Fujimura and M. Hirano (Eds.) Vocal fold physiology: Voice quality control, 1995

28. Rabiner L. and Schafer R. Digital processing of speech signals. Englewood Cliffs, N.J.: Prentice Hall, 1978

29. Fant, G. and Lin, Q. Frequency domain interpretation and derivation of glottal flow parameters. STL-QPSR 1988, **2-3**:1-23

30. Imaizumi, S. et al Harmonic analysis of the singing voice: - Acoustic characteristics of vibrato. Proc. of the Stockholm Music Acoustics Conference, Stockholm, (SMAC 93)
31. Horii, Y. Fundamental frequency perturbation observed in sustained phonation. J. Speech and Hearing Res. 1979; **18**: 19-201
32. Mathews, MV. et al Pitch synchronous analysis of voiced sounds. J. Acoust. Soc. Am. 1971; **33**: 179-186
33. Rothenberg, M. In O. Fujimura (Ed.) Vocal fold physiology: Voice production, mechanisms and functions. NY: Raven Press, 1988
34. Ladefoged, P. In O. Fujimura (Ed.) Vocal fold physiology: Voice production, mechanisms and functions. NY: Raven Press, 1988
35. Koike, Y. and Kohda, J. The effect of vocal fold surgery on the speech cepstrum. In J. Gaufin and B. Hammarberg (Eds.) Vocal fold physiology: Acoustic, perceptual and physiologic aspects of voice mechanisms. San Diego: Singular Publ. 1991
36. Titze, IR. Three models of phonation. J. Acoust. Soc. Am. Suppl. 1 **79**: S81
37. de Krom, G. A cepstrum based technique for determining a harmonics-to-noise ratio in speech signals. J. Speech Hear Res. 1993; **36**: 254-266
38. Kasuya, H. and Endo, Y. Analysis, synthesis and perception of brathy voice. In J. Gaufin and B. Hammarberg (Eds.) Vocal fold physiology: Acoustic, perceptual and physiologic aspects of voice mechanisms. San Diego: Singular Publ. 1991
39. Fujimura, O. In O. Fujimura (Ed.) Vocal fold physiology: Voice production, mechanisms and functions. NY: Raven Press, 1988
40. Gobl, C. Voice source dynamics in connected speech. STL-QPSR 1:123-159 (Dept. of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm).
41. Karlsson, I. Glottal waveform parameters for different speaker types. Proc. SPEECH '88 (7th FASE Symp.), 1:225-231. (Institute of Acoustics, Edinburgh).
42. Titze, IR. Mechanisms underlying the control of fundamental frequency. In J. Gaufin and B. Hammarberg (Eds.) Vocal fold physiology: Acoustic, perceptual and physiologic aspects of voice mechanisms. San Diego: Singular Publ. 1991
43. Farley, GR. A biomechanical laryngeal model of voice f0 control and glottal width control. J. Acoust. Soc. Amer. 1996, **100**:3794-3812

44. Lee, CK. Childers, DG. Some acoustical, perceptual and physiological aspects of vocal quality. In J. Gaufin and B. Hammarberg (Eds.) Vocal fold physiology: Acoustic, perceptual and physiologic aspects of voice mechanisms. San Diego: Singular Publ. 1991

45. Imaizumi, S. A preliminary study on the generation of pathological voice qualities. In O. Fujimura (Ed.) Vocal fold physiology: Voice production, mechanisms and functions. NY: Raven Press, 1988

46. Gauffin, J. et al, S. A microcomputer based system for acoustic analysis of voice characteristics. ICAPP 1986 IEEE 677-684

47. Milenkovic, P. et al Acoustic and perceptual characterisation of vocal nodules. In J. Gaufin and B. Hammarberg (Eds.) Vocal fold physiology: Acoustic, perceptual and physiologic aspects of voice mechanisms. San Diego: Singular Publ. 1991

48. Kasuya, H. and Endo, Y. Acoustic analysis, conversion and synthesis of the pathological voice. In O. Fujimura and M. Hirano (Eds.) Vocal fold physiology: Voice quality control, San Diego: Singular Publ., 1995

49. Qi, Y. Time normalisation in voice analysis. J. Acoust. Soc. Am. 1992; **92**: 1569-76

50. Gavidia-Ceballos L. Hansen JH. Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. IEEE Trans. on Biomedical Eng. 1996; **43**: 373-83

51. Hanson, HM. Glottal characteristics of female speakers; Acoustic correlates. J. Acoust. Soc. Am. 1997; **101:**466:481

52. Holmberg, EB. Comparisons among aerodynamic, electroglottographic and acoustic spectral measures of female voice. J. Speech Hearing Res. 1995, **38**:1212:1223

53. de Krom, G. Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. J. Speech Hear. Res. 1995; **38**: 794-811

Chapter 6

Long Term Average Spectrum Analysis

6.1 Introduction

The long term average spectrum (LTAS) is defined as the ensemble average of successive spectral estimates for a given sample of speech. The speech material in this case might typically consist of reading a paragraph of newspaper text. The use of LTAS derived measures has been the primary method of choice for speaker verification systems for some time now[1,2]. Part of their attraction is that they offer the possibility of text independent measurements of speaker identifying features i.e. if a sufficiently long sample of speech is taken, the resulting averaged spectrum contains information pertinent to speaker identification as opposed to the words spoken.

The LTAS has also been well documented in the speech pathology literature although opinions on it's potential use as an objective indicator of dysphonia have been divided.[3,4] Hammerberg et al[5] successfully used the LTAS to differentiate breathiness conditions of both hypofunction and hyperfunctional voice. Wendler also found it to be a very promising measure but later tailored his initial optimistic opinions on it's use[6].

In a study performed by Lofqvist[7] (1986) an attempt was made to differentiate a set of 37 clinical voices from a set of 36 normal voices. The two measures taken from the LTAS were the ratio of the energy below 1 kHz to the energy from 1 to 5 kHz and the energy level between 5 and 8 kHz. The result of the analysis produced almost a complete overlap of patient and normal data and Lofqvist therefore expressed a pessimistic view for it's use in clinical investigations. In another study by Lofqvist[8] he states that, in applying the LTAS

"... the short term variations due to phonetic structure will be averaged out and the resulting spectrum can be used to obtain information on the sound source: if the analysis is restricted to voiced sounds, the sound source is the vibrating glottis."

He therefore felt that the two LTAS derived measures were indicative of the sound source. He also investigated the effect of the time length on the LTAS and noticed that reducing a sample of voiced speech from 20 to 10 seconds had little effect, however, further reduction in the sample length made the LTAS variations unpredictable. In a study performed by Kitzing[9] at the same time as the Lofqvist work, an extensive set of LTAS derived measured were investigated. Their measures in particular proved useful for separating strained and sonorous vocal qualities: (1) the ratio of energy below and above 1 kHz , 2) a measure of the spectral slope inclination in the first formant range and (3) the ratio of the peak level of the fundamental and the first formant region. In a later study[10] he found that LTAS derived measures correlated moderately well with perceived improvements in patients undergoing voice therapy.

6.2 Analysis

As stated above, LTAS is generally used on connected speech where the vocal tract filter function varies with time. Advantages[11] and disadvantages[12] of using connected speech have both been reported in the speech pathology literature. From the point of view of general speech disorders or disorders associated with articulatory dynamics connected speech is the preferred choice. An obvious example would be spasmodic

dysphonia[13], where the aberrant vocal quality might not show up during the phonation of a sustained vowel but is easily detected in running speech.

Although a phonetically balanced sentence of about 2 seconds duration was recorded for all patients in this study and a program was successfully coded (thres.m - appendix x) to remove unvoiced segments of speech based on amplitude and zero-crossing rate considerations, it was decided not to use connected speech for the following reasons:

1. Lofqvist[7] reported variable LTAS results if the speech sample used was less than 10 seconds.
2. The result of a study by Anathapadmanabha[15] (1992), based on LTAS derived from a two second sample of connected speech showed LTAS to be a poor indicator of vocal quality.
3. We do not agree with Lofqvist's assumption that averaging a sufficient number of spectra will cancel the overall formant contributions and therefore leave a spectrum directed related to the voice source. This would require that the sum of the vocal tract spectra (second term of the third expression in eqtn.6.1) would add to zero.

$$\mathrm{LTAS}(f) = \frac{1}{N}\sum_{i=1}^{N} S_i(f) = \sum_{i=1}^{N} E_i(f) \times V_i(f) = E(f) \times \sum_{i=1}^{N} V_i(f) \qquad \text{eqtn.6.1}$$

$E_i = i^{th}$ source spectrum

$V_i = i^{th}$ vocal tract spectrum

$S_i = i^{th}$ speech spectrum

Where it is assumed that $E_i(f)$ is constant.

Therefore measures of fundamental frequency (f0) and first formant (f1) levels taken from the LTAS should be treated with caution in respect to how they relate to the voice source. Or as pointed out by Kitzing[10],

"... even if it may be possible to neutralise the influence of isolated vowel articulation by averaging a sufficient number of spectra, there is always still a substantial influence from articulation and the resonators of the vocal tract on the spectrum."

A sustained phonation of the vowel a/ is used in the analysis. This was felt to represent the voice source more directly, even though the influence of the vocal tract resonating cavities are of course present in order to produce the vowel resonances, the configuration is fixed and therefore the spectra do not add in an unpredictable manner. Averaging several short time spectra across a given phonation reduces the variance of the spectral estimates which can be quite high for non deterministic or random signals. This so called periodogram[15] averaging was thought useful in anticipation of the aperiodicities and random noise found in pathological voices.
The program ltas.m was coded in the Matlab high level language with the following input parameters.

1. window size 256
2. hop size 100
3. total length 0.75 sec

Another version (ltashfe.m) of the program was also written in order to provide high frequency emphasis in the spectrum. The source code is given in appendix A. The LTAS for two normals and two patients of the present study are shown in fig.6.1.
Four measures were taken from the resultant spectra :

1. The ratio of the energy below 1 kHz to that above 1kHz. (R_{14})
2. The ratio of the energy below 2 kHz to that above 2kHz. (R_{24})
3. The ratio of the energy below 1 kHz to that above 1kHz taken from a dB-averaged spectra.
4. The ratio of the energy below 2 kHz to that above 2kHz taken from dB-averaged spectra.

Where R_{14} and R_{24} were calculated form the following equations

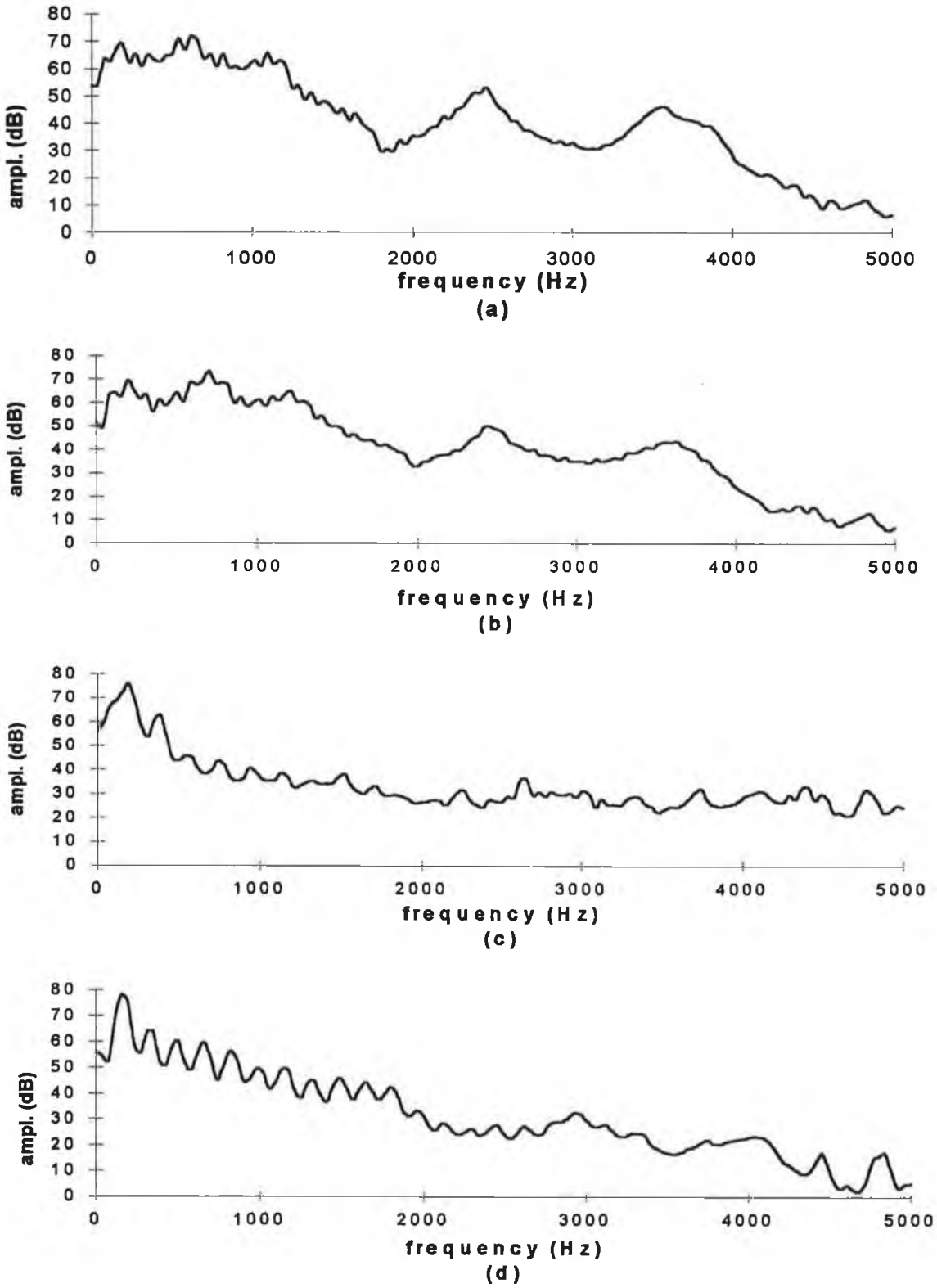


fig.6.1 *LTAS for* (a), (b) *two 'normals' and* (c), (d) *two patients of the present study with window length 25.6 ms, hop size 10 ms and total analysis length 0.75 second*

$$R_{14} = \frac{\sum_{i=1}^{N/5} A_i^2}{\sum_{i=N/5}^{5\times N/4} A_i^2} \qquad \text{eqtn.6.2}$$

$$R_{24} = \frac{\sum_{i=1}^{N\times 2/5} A_i^2}{\sum_{i=N\times 2/5}^{5\times N/4} A_i^2} \qquad \text{eqtn.6.3}$$

where A_i is the spectral amplitude at the i^{th} frequency
location taken from the LTAS.
N = number of spectral estimates of the LTAS (128) covering up to 5 kHz.

6.3 Results

The above program was run on all the synthesis data files detailed in table 2.5 of chapter 2. LTAS are show for the four different perturbation measures in fig.6.2. The results for (1) and (2) above are given in fig.6.3 and fig.6.4 for the 110 Hz synthesis of the vowel a/. As can be seen in the figure, as the additive noise level increases the R_{14} and R_{24} ratios decrease. As the noise is mean zero, Gaussian noise it has a flat spectral characteristic. In considering the source spectrum, equal amounts of energy are added to the signal but because the lower partials of the source are significantly greater in magnitude the influence of the additional noise has a much lesser effect on the numerator in eqtn.6.2 and eqtn.6.3 than it has on the denominator. The magnitude of the resonant contributions of the vocal tract are also a consideration. The results for the jitter set signals and the shimmer signals are shown in fig. 6.5 and 6.6 respectively. An examination of these graphs reveals that the R_{14} and R_{24} ratios are somewhat insensitive to jitter and shimmer and even for the maximum amounts of jitter and shimmer added, the ratios are above the corresponding ratios obtained for the additive noise signals.

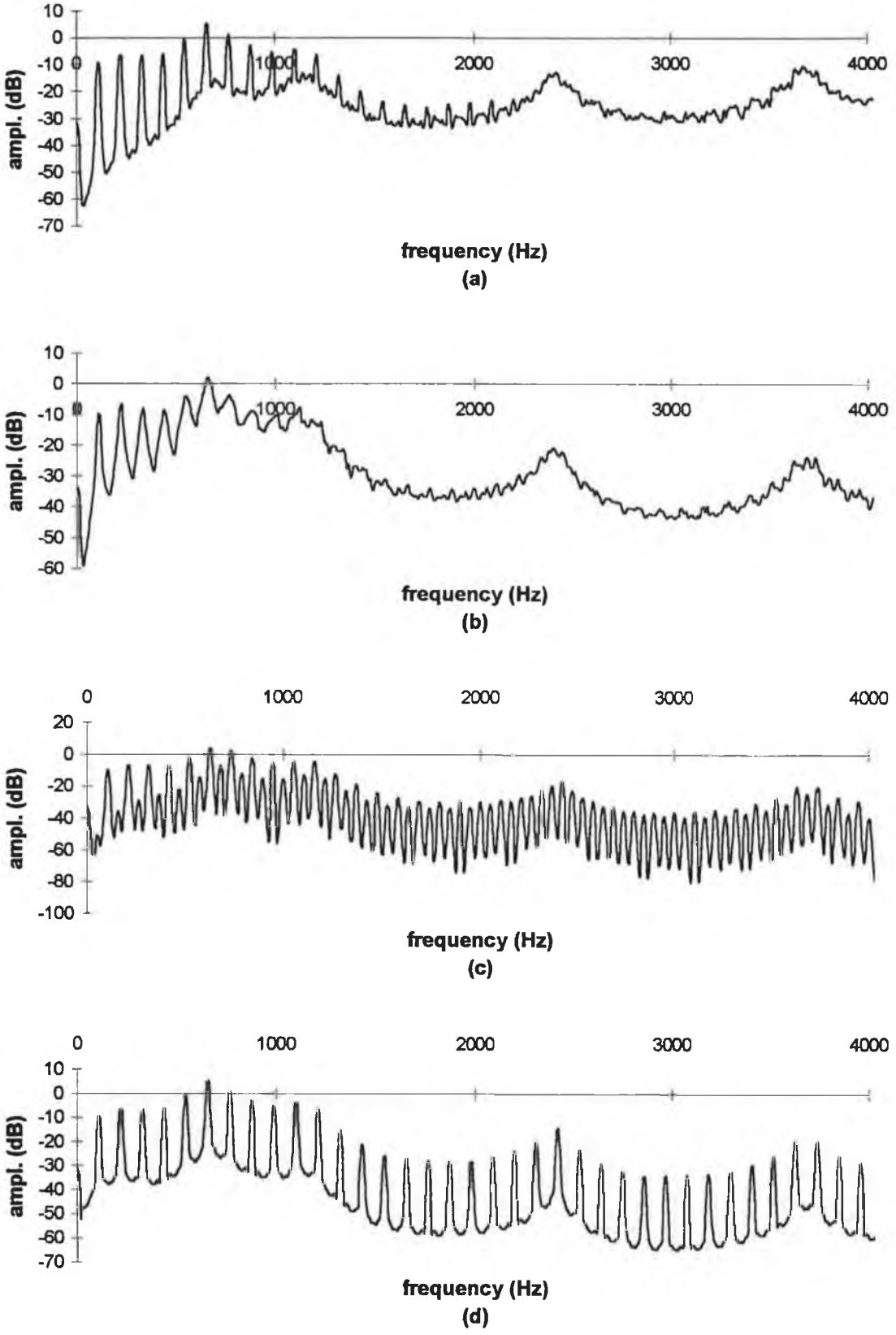


fig.6.2 *LTAS for* (a) *8 % std. dev. additive noise*, (b) *4% std.dev. random jitter*, (c) *4% cyclic jitter and* (c) *std.dev.8 % random shimmer. (LTAS window length = 1024)*

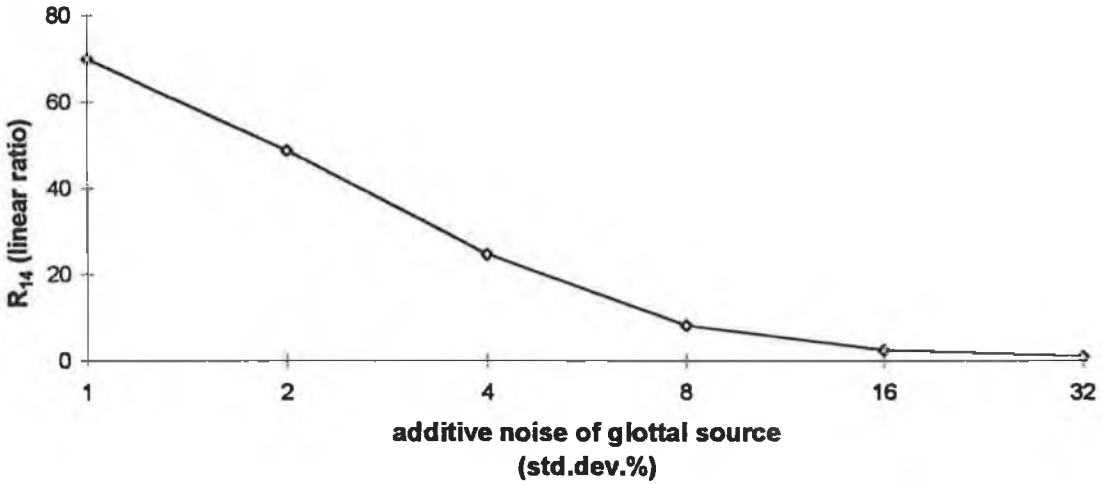


fig.6.3 *R_{14} vs random additive noise of the glottal source*

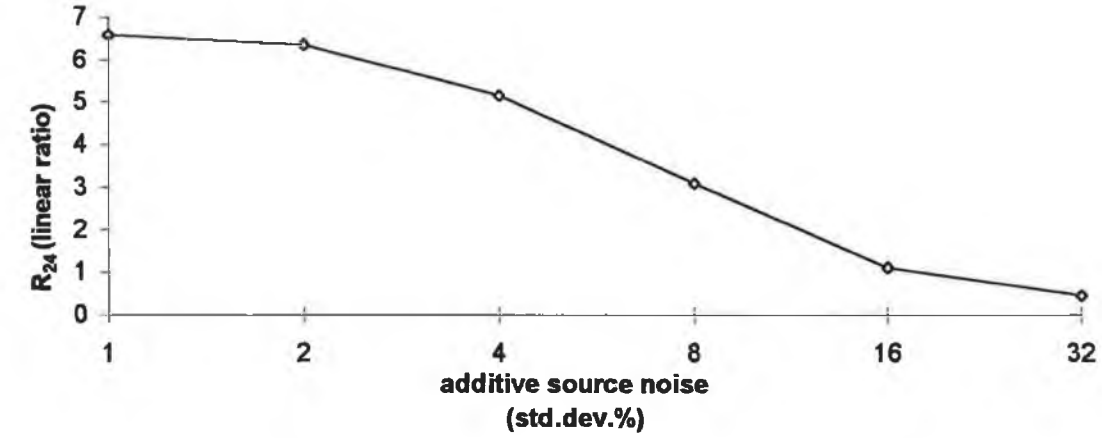


fig.6.4 *R_{24} vs random additive noise of the glottal source*

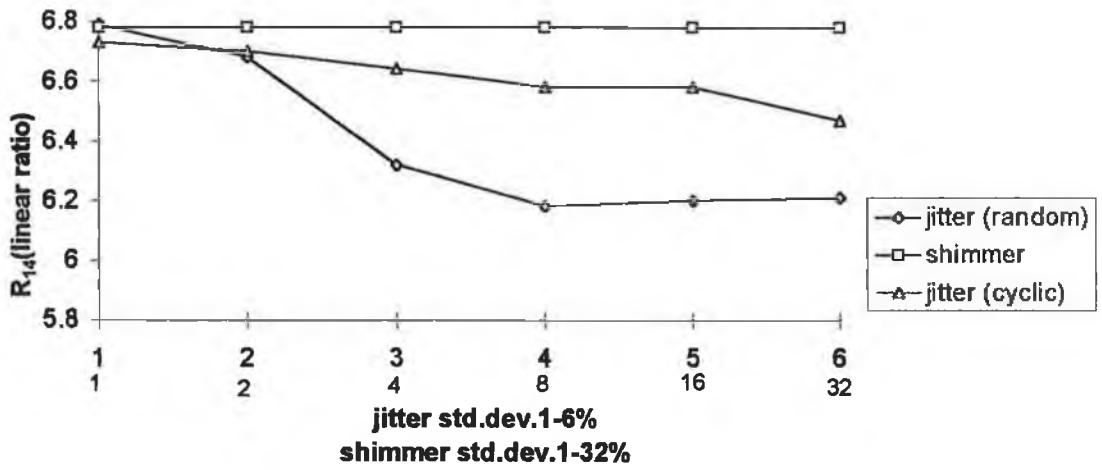


fig.6.5 *R_{14} vs random and cyclic jitter and shimmer of the glottal source*

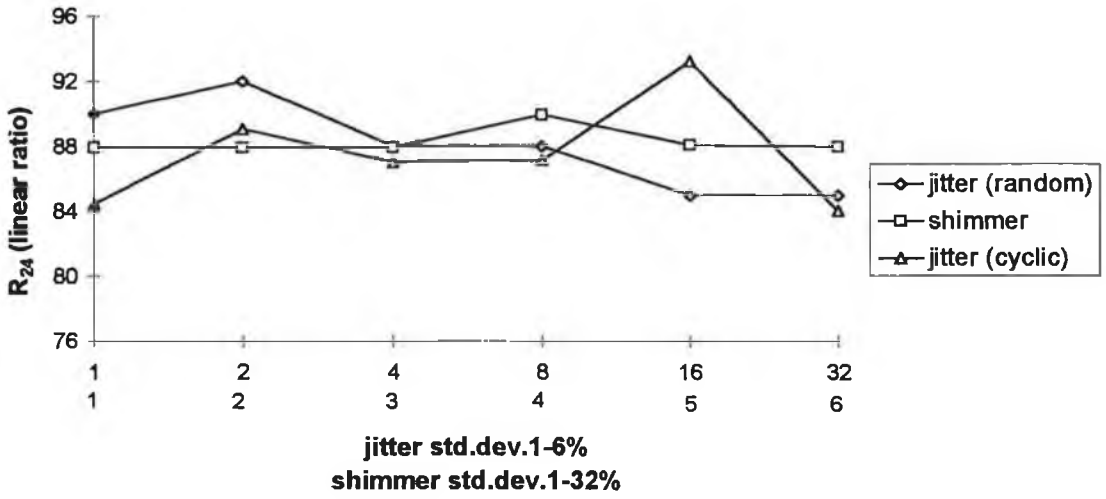


fig.6.6 *R_{24} vs random and cyclic jitter and shimmer of the glottal source*

This is an interesting result in that we now have a ratio that is representative of additive noise and independent of jitter and shimmer values. Although, of course we should keep in mind that this result is critically dependent on how well we have modelled the noise signal. The results obtained for (3) and (4) above proved inconclusive and were not considered further as were the results from the high frequency emphasis program.

The ltas.m program was next run on the patient/normal data files for the vowel a/. Histograms of the results is given in fig.6.7 and fig.6.8. Referring to these figures, it can be seen that the R_{24} ratio shows poor separability, whereas the R_{14} ratio has separated all but one of the patients from the normal data. Interestingly, however the results are in complete opposition to the ratios obtained with the additive noise (fig. 6.2). In respect to the R_{14} ratio, the 'normal' data are in agreement with values obtained from the low additive noise or perturbation results whereas the patient data shows a marked increase in this ratio. Firstly, it is encouraging that the ratio, which was shown to reflect additive noise levels, independent of jitter and shimmer, has separated the real data. However, it also raises questions about our simplistic model for simulating pathologies. In our model the glottal waveform maintains it's shape with the open and closed periods remaining fixed and signal dependent, mean zero, Gaussian noise is introduced. In this manner no noise is added during the closed phase as the airflow is assumed to be zero during this period. In more realistic models of glottal flow we would expect the closed phase to be less pronounced in many cases

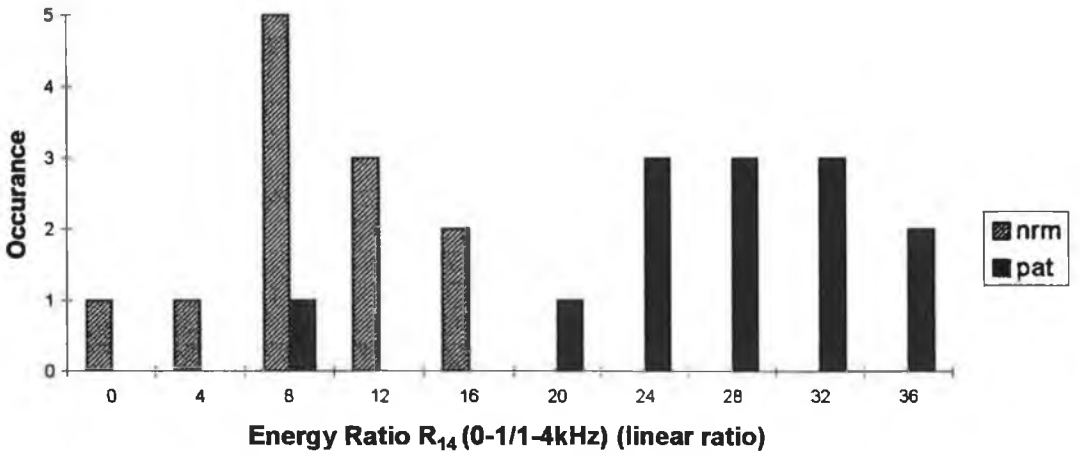


fig.6.7 *Histogram of R_{14} for the patient/normal data set. Highly significant at the 5 % level (one tailed, two sample, equal variance, student's t-test).*

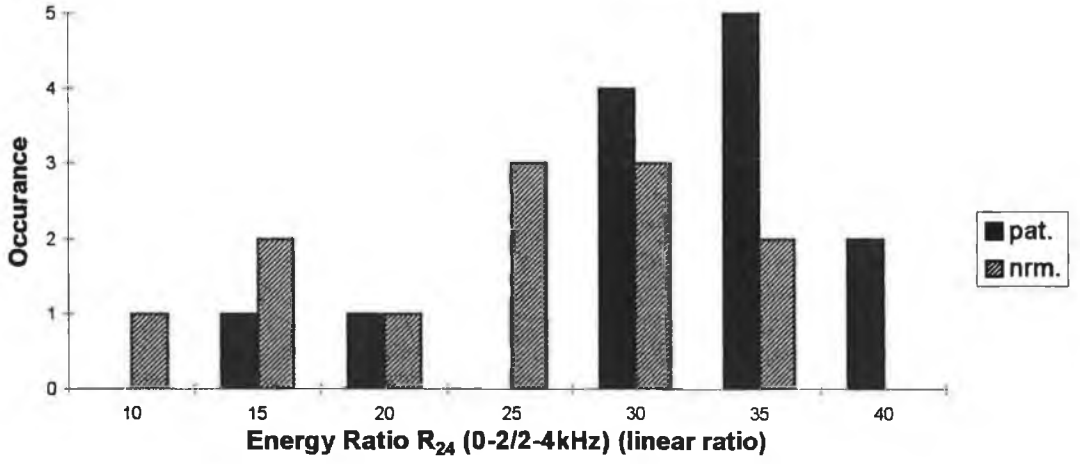


fig.6.8 *Histogram of R_{24} for the patient/normal data set. Not significant at the 5 % level (one tailed, two sample, equal variance, student's t-test)*

involving voice pathologies. This might be due, for example, to the effect of mass lesions such as nodules, polyps etc. which would result in appreciable airflow leakage even when the vocal processes come into contact during the attempted close phase of the cycle. Incomplete or reduced closed phase also occurs in hypofunctional and breathy voices or as a result of vocal cord paralysis or paresis. The acoustic effect of this, is a reduction in harmonic structure in the higher frequency end of the spectrum which leads to an overall reduction in energy in the upper part of the spectrum despite

having an increase in noise energy in this region. Also, the lower frequency region may have an increase in the lower frequency components due to the more sinusoidal nature of the glottal flow waveform. The overall result of this is an increase in the R_{14} ratio (also referred to as called spectral tilt).

6.4 Discussion and Conclusion:

The LTAS has been obtained from sustained phonations rather than from connected speech for the reasons outlined in section 6.2. Averaging (n) the spectrum of a sustained phonation reduces the variance of the spectral estimates by 1/n at the expense of broader bandwidths and hence reduced resolution. Two measures, the ratio of the energy below 1 kHz to the energy above 1 kHz (R_{14}) and the ratio of the energy below 2 kHz to the energy above 2 kHz (R_{24}) were considered for analysis. Both the R_{14} and R_{24} ratios decreased with increasing levels of additive noise. Both methods were also relatively insensitive to jitter and shimmer. Furthermore, R_{14} has been shown to be a useful indicator of vocal pathology.

A number of studies have attempted to relate the spectral effects of varying parameters in the glottal flow waveform. Most of this work is based on Fant's four parameter LF-model of glottal flow[16] (fig. 6.8) where the four parameters are derived from the three basic time events that occur during the glottal cycle, 1) the location of peak flow, T_p 2) the discontinuity point at glottal closure, T_e and 3) the return phase, T_a. In relation to the source spectrum, these studies have revealed that :

1) the level of the fundamental is closely related to the rising portion of the flow glottogram.

2) the rate of closure corresponds to the level of all upper partials, i.e. a higher closing speed gives rise to an increase to all higher harmonics.

3) the spectral tilt is very dependent on the final part of the closing phase that appears after the instant of maximum airflow decrease.

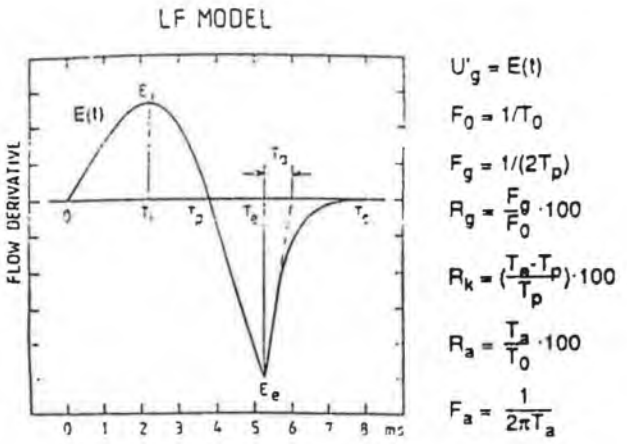


fig.6.8 *Four parameter model of differentiated glottal flow illustrating the three main time events that occur during the glottal cycle - 1) the location of peak flow, T_p 2) the discontimuity point at glottal closure, T_e and 3) the return phase, T_a.*

The R_{14} measure incorporates all of these source spectrum characteristics. Therefore to investigate the measures independently, other spectral measurements are required. The ratio of the amplitude of the first harmonic to the amplitude of the second harmonic is thought to relate to abductory behaviour[17] whereas the ratio of the amplitude of the first harmonic to the amplitude of the third formant is thought to relate more closely to the closing phase[18,19]. In our glottal flow waveform, adductory and abductory behaviour were purposely not altered in order to firstly examine the gross spectral characteristics of jitter, shimmer and additive noise. There are obvious advantages of modelling in this fashion, in that there is a strict control on the variables, a situation always absent, even for 'normal' voice. The disadvantage is that these situations may, in fact, not be physically realisable. Independent model parameters may not behave in an unconstrained manner in practical situations. Gauffin and Sundberg[20] made a similar comment when comparing the flow glottograms of singers and non-singers:

> "As the Fant model is theoretical, it will consider cases, regardless of whether or not they occur in reality. In our material on the other hand, we have included only normal or trained voices. In pathological voices, glottogram characteristics may be combined in other ways.".

So, in detailing the spectrum in respect to it's ability to extract salient cues to vocal pathology we must consider, a) the spectral consequences of the variable parameters that occur in the four parameter model and b) the spectral consequences of additional features important to vocal pathology. Additional factors for consideration (to those mentioned above) include the type of phonation, such as soft, normal and loud and in respect to the spectral analysis, how many frames to average, overlap etc. or whether and when to use a pitch synchronous harmonic analysis.

As previously mentioned, use of the LTAS on short segments of speech, has not been very successful in making correlations with vocal qualities[14]. An interesting alternative to using the LTAS, would be to investigate the long term harmonic spectrum, LTHS, defined as the ensemble average of successive harmonic estimates for a given sample of speech i.e. spectra are extracted pitch synchronously and subsequently averaged according to harmonic number as opposed to frequency location. Alternatively, pitch synchronous inverse filtering, based on a spectral matching procedure, followed by averaging facilitates a more direct comparison to the parameters included in the LF model of glottal flow.

A further advantage of making calculations based on harmonic number as opposed to frequency location is that the ratio of the number of harmonics within the 0-1 kHz range to the 1-4 kHz range changes with fundamental frequency. Table 6.1 illustrates these ratios for the synthesis data files. Using harmonic numbers the ratio is of course constant. An obvious objection to this approach is that the harmonics receive different resonant contributions if the fundamental frequency is different. A possible solution would be to calculate the ratio between harmonic to noise ratio from 0-1 kHz to the harmonic to noise ratio from 1-4 kHz. Another factor for consideration is the relationship between the harmonic locations in relation to the position of the formant peaks. Figure 6.8 (a) and (b) show how the R_{14} and R_{24} ratios vary with fundamental frequencies. The fact that the harmonics in the spectrum of the low fundamental frequency data file have attenuated by 2 kHz and the harmonics for the higher fundamental frequency data are still very prominent at 5 kHz is also an important consideration.

Finally, basic research is required in order to relate glottal flow (and hence spectral) characteristics to specific vocal pathologies of aberrant vocal fold vibrations. The

Vowel a/ with first three resonances at ~660, ~1100 and ~2400 Hz						
fundamental frequency / number of harmonics	80	110	160	220	290	350
a) 0-1 kHz	12	9	6	4	3	2
b) 1-4 kHz	38	27	19	14	10	9
ratio a) / b)	0.316	0.333	0.316	0.286	0.3	0.22

Table.6.1 *Ratio of number of harmonics from 0-1 kHz to number of harmonics to 1-4 kHz.*

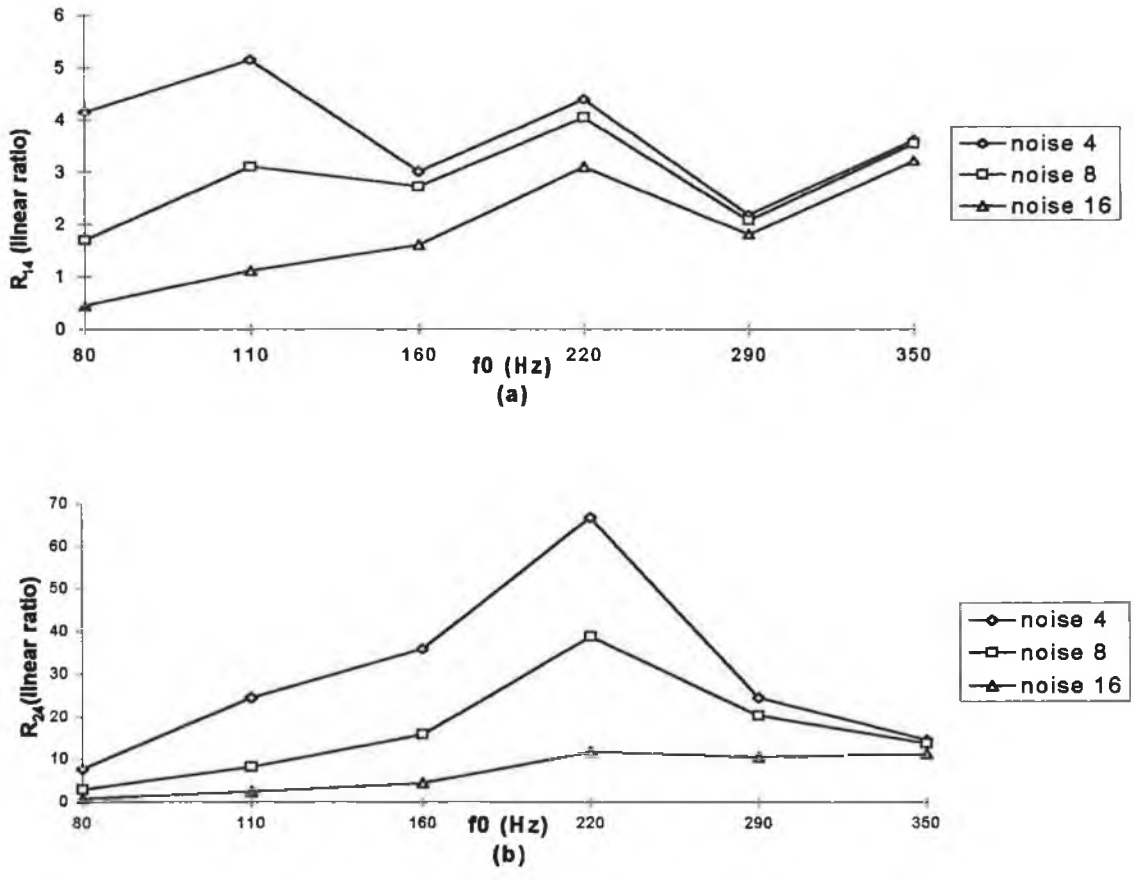


fig.6.9 *Variation of* (a) R_{14} *and* (b) R_{24} *with fundamental frequency for three levels of additive random noise of std. dev. 4, 8 and 16 %.*

connection between flow characteristics is not always obvious and this link in the speech chain may prove to be the limiting factor in the potential usefulness of acoustic analyses with regard to differential and accurate diagnoses of varying pathology types. A couple of examples of possible inferences based on flow characteristics are given. From a knowledge of the rate of decrease of airflow before closure we may be able to infer how the closure occurred along the length of the vocal process, beginning anteriorly and continuing until the posterior processes approximate (if they do). Subsequent to this is the closed phase. The length of this period after a given rate of decrease of volume velocity might provide information regarding the tissue properties of the folds (elasticity and impact stress). The next example helps illustrate the possible ambiguities that could possibly arise from inferences regarding vibratory characteristics based on the flow characteristics. When assymmetry of phase occurs, the vocal fold consistently move in the same direction (Diane Bless[21] illustrates a nice example on her instructional stroboscopy video cassette). This results in a constant glottal area with no open or closed phase. Greater airflow with possible turbulent characteristics therefore results. It remains to be seen if the acoustic analyses including LTAS and associated parameters can differentiate the resultant airflow characteristics from this type of dysfunction with turbulent flow due to the presence of polyps or mass lesions. Hence the need for basic research correlating acoustic analysis parameters to alternatively assessed specific pathology types.

6.5 Bibliography

1. Rosenberg, AE. Automatic speech verification : a review, Proc. IEEE 1976; **64**: 475-487
2. Hollien, H. et al Speaker identification by long term spectra under normal, stess and disguised conditions. J. Acoust. Soc. Am. 1974; **55**: S-20(A)
3. Wendler, J. et al: Voice calssification by means of long term spectra. Folia Phoniat. 1980 **32**: 51-60
4. Prytz, Long time average spectra (LTAS) of normal and pathological voices; Proc. 17th Int. Congr. Logopedics Phoniat. 1977; **1**: 459-475
5. Hammarberg, B. et al erceptual and acoustic correlates of abnormal voice qualities. Acta. Oto-lar. 1980; **90**: 441-451
6. Wendler, J. et al, Classification of voice qualities. Journal of Phonetics 1986 **14**: 483-488
7. Lofqvist, A. Mandersson, B. Long-Time Average Spectrum of Speech and Voice Analysis Folia phoniat. 1987; **39**: 221-229
8. Lofqvist, A. The Long Time Average Spectrum as a tool in voice research, J. Phonetics 1986; **14**: 471-475
9. Kitzing, P. LTAS criteria pertinent to the measurement of voice quality. J. Phonetics 1986; **14**: 477-482
10. Kitzing, P. and Akerlund, L. Long time average spectrograms of dysphonic voices before and after therapy. Folia Phoniatr. 1993; **45**:53-61
11. Pruszewicz, A. et al Usefulness of acoustic studies on the differential diagnostics of organic and functional dysphonia. Acta. Otolaryngol. 1991; **111**: 414-419
12. Fex, B. Acoustic analysis of functional dysphonia: before and after voice therapy (accent method) J. Voice 1995 **8**: 163-167
13. Blitzer, A. et al Clinical and laboratory characteristics of focal laryngeal dystonia: study of 110 cases. Laryngoscope 1988; **98**: 636-640
14. Ananthapadmanabha, TV. Acoustic factors determining perceived voice quality. In O. Fujimura and M. Hirano (Eds.) Vocal fold physiology: Voice quality control, San Diego: Singular Publ., 1995

15. Rabiner, L. and Schafer, R. Digitial processing of speech signals, Englewood Cliffs, N.J.: Prentice Hall, 1978

16. Fant, G. and Lin, Q. Frequency domain interpretation and derivation of glottal flow parameters. STL-QPSR 1988, **2-3**:1-23

17. Klatt, DH. and Klatt, LC. Analysis, synthesis and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 1990; **87**: 820-857

18. Holmberg, EB. Comparisons among aerodynamic, electroglottographic and acoustic spectral measures of female voice. J. Speech Hearing Res. 1995, **38**:1212:1223

19. Hanson, HM. Glottal characteristics of female speakers; Acoustic correlates. J. Acoust. Soc. Am. 1997; **101:**466:481

20. Gauffin, J. and Sundberg, J. Spectral correlates of glottal source waveform characteristics. J. Speech Hear. Res. 1989; **32**: 56-565

21. Bless, D. et al Comparison of vibratory characteristics of young adult males and females. In M. Hirano and R. Hibi (Eds.) Proc. of the Int. Conf. on Voice. Kurume, Japan. 1986; **2**: 46-54

Chapter 7

Cepstral Analysis Techniques

7.1 Introduction

The term cepstrum first appeared in the scientific literature in a paper with the unusual title: "The Quefrency Alanysis of Time Series for Echos: Cepstrum, Pseudoautocovariance, Cross-Cepstrum, and Saphe Cracking"[1]. The paper was published in 1963 by Bogert, Healy and Tukey in which they observed that the logarithm of the power spectrum of a signal containing an echo has an additive periodic component due to the echo, and thus the Fourier transform of the logarithm of the power spectrum should exhibit a peak at the echo delay. They called this function the 'cepstrum', reversing the order of the first four letters in the word spectrum because according to Bogert et al "we find ourselves operating on the frequency side in ways customary on the time side and vice versa." Tukey went on to liberally define a rich vocabulary of terms reflecting the fact that the resultant functions lay neither properly in the time or frequency domains. However, only the words cepstrum, rahmonics, quefrency and liftering have found popular usage.

The use of the cepstrum function by the above authors to locate echos in seismic data was not very successful. Schroeder[2] suggested it's use for analysing speech signals, given that the log of the short time spectrum for voiced speech exhibits an envelope corresponding to the vocal tract transfer function with a superimposed periodic component due to the glottal source. The Fourier transform of this spectrum should therefore lead to a prominent peak corresponding to the pitch period and a large signal based around zero Hertz reflecting the spectral envelope. The cepstral technique proved very robust in detecting the pitch period of voiced speech, even in the presence of additive noise: a detailed account is given in Noll[3]. The ability of the cepstrum to separate the source and envelope characteristics of the speech was next put to use in formant extraction schemes[4] and more recently cepstral coefficients have replaced LPC coefficients in advanced speech recognition systems[5]. Before going on to review cepstral methods as applied to speech pathology we should note that the cepstrum comprises one of several methods that fall under the general heading of 'Homomorphic Deconvolution'. This new class of systems was proposed by Oppenheim[6] shortly after the paper by Bogert et al. They are nonlinear systems in a classical sense but they do follow a type of generalised superposition principle i.e. input signals and their corresponding responses are superimposed by an operation having the same algebraic properties as addition.

Despite the above mentioned success of cepstral methods in speech analysis they have received very little attention in the vocal pathology literature. A complete review of the literature reveals only three independent authors to have investigated it's use as an indicator of vocal pathology. Koike[7] has used the method on three separate occasions to assist patient diagnosis. The height of the first cepstral peak (first rahmonic) was used as an indicator of good periodicity and the location of this peak on the quefrency axis was used to determine the pitch period. Positive, objective assessment was reported through using the method and a call for further studies was made. Hillenbrand[8] also calculated the height of the first cepstral peak, using a normalisation procedure in which he did not check for pitch tracking errors and yet found the method very useful in predicting breathiness. By far the most comprehensive assessment of the technique with respect to it's application to speech pathologies was the work by de Krom[9], "A Cepstrum-Based Technique for Determining an Harmonic-to-Noise

Ratio in Speech Signals", in which he removed the rahmonics in the cepstrum, Fourier transformed the resulting liftered cepstrum to provide a noise spectrum which was subtracted from the original log spectrum. This resulted in, what we have termed, a source related spectrum. After performing a baseline correction procedure on this spectrum, the modified noise spectrum was subtracted from the original log spectrum in order to provide the harmonic to noise ratio estimate. This work used speech samples synthesized to simulate various amounts of jitter and additive noise conditions in a manner similar to the files produced here. de Krom pointed out the absence of a database of pathological voices samples from which researchers might investigate new analysis techniques and citing the dangers of using pathological voice samples with vague assessments suggested the use of synthetic signals as an objective, quantifiable alternative. This is the procedure followed throughout this study. In addition to this, for the cases outlined here, once a method appeared successful based on the synthesis data, it was then used in an attempt to discriminate between a group of thirteen patients with various voice disorders (Table.2.1 Chapter 2) and twelve normal speakers. The method detailed in this chapter is developed along the same lines as the de Krom technique, but the present procedure follows fewer steps, leading to a source related harmonic to noise ratio (H/N_S). Hillenbrand's normalisation scheme and bandlimited analysis was also tested.

7.2 Method

Noll pointed the way forward towards an easy explanation of the seemingly complicated cepstrum technique, stating that "the spectrum itself can be regarded as a signal and can be processed by standard signal-analysis techniques". We follow this intuitive approach to the cepstrum of voiced speech in the outline that follows.

For voiced speech we know that s(t) can be represented as the convolution of the excitation signal and the impulse response of the vocal tract transfer function (fig. 7.1).

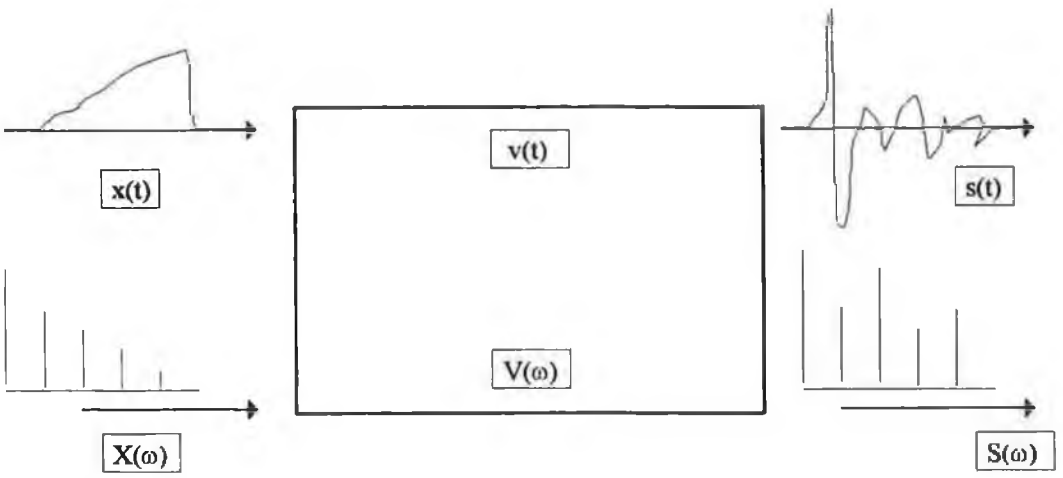


fig.7.1 *Source/filter model of speech production illustrating impulse and frequency responses.*

$$s(t) = x(t)*v(t) \qquad \text{eqtn.7.1}$$

s(t) = output radiated speech waveform

x(t) = glottal source

v(t) = transfer function of the vocal tract

* indicates convolution

This is equivalently represented by the frequency response of the vocal tract as

$$S(\omega) = |X(\omega)| \times |V(\omega)| \qquad \text{eqtn.7.2}$$

where $S(\omega) = F(s(t))$, $X(\omega) = F(x(n))$, $V(\omega) = F(v(n))$

where F represents Fourier transformation.

Now, by simply taking the logarithm of the spectrum, we obtain

$$\log|S(\omega)| = \log|X(\omega)| \times \log|V(\omega)| \qquad \text{eqtn.7.3}$$

$$\log|S(\omega)| = \log|X(\omega)| + \log|V(\omega)|$$

Therefore the multiplicative components consisting of the source excitation (fast varying) and filter function (slow varying) have been changed into additive components. The motivation for doing this is that we can now apply a linear operator, i.e. the Fourier transform, knowing that the transform operates individually on the two additive components and that the transform will conveniently separate the slowly varying part from the fast varying part. In this way we are considering the signal log|S(w)| as a standard "time" signal with one "high frequency" and one "low-frequency" component, the Fourier transform of which, gives a high amplitude at locations in the "frequency domain" corresponding to these frequencies. Since, we were in the frequency domain to begin with the new terminology (rahmonics, quefrency, etc.) was employed to reflect this distinction between the resultant and that which would have occurred with using real time domain signals. This process is illustrated in fig.7.2. Therefore taking the inverse Fourier transform of the log magnitude spectrum yields the real cepstrum

$$C(\tau)= \text{IDFT}[\log|S(\omega)|] \qquad \text{eqtn.7.4}$$

where τ represents quefrency
and IDFT is the inverse discrete Fourier transform.

Note the complex cepstrum is obtained by simply replacing the magnitude spectrum, $|S(\omega)|$, with the Fourier spectrum, $S(\omega)$. The complex cepstrum is rarely used in practice and is only of interest where knowledge of the original phase of the signal is important. A similar argument holds for the appropriateness of taking the forward Fourier transform or it's inverse: it is in fact not that important, so long as the phase is not a major concern.
The inverse Fourier transform is conventionally taken although Noll[3] developed his ideas based on the forward Fourier transform.

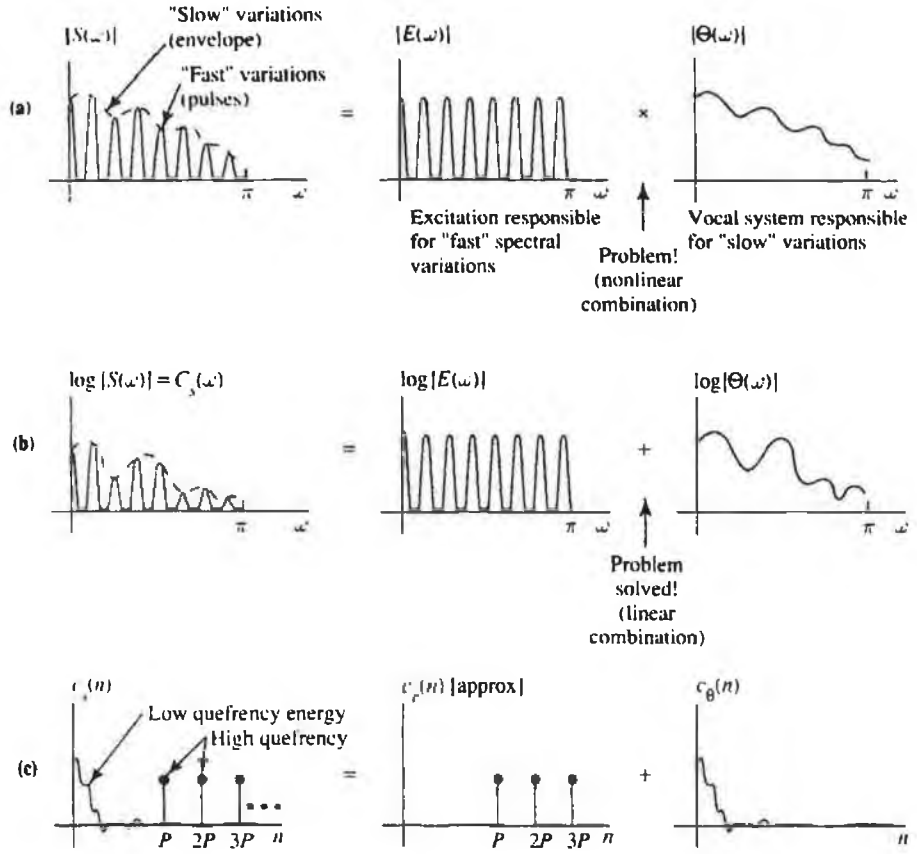


fig.7.2 *Illustration of the cepstral technique whereupon applying the log operation to the speech waveform, the source and filter contributions are separated. Deller/Proakis/Hansen, Discrete Time Processing of Speech Signals,©1993,p.361. Reprinted by permission of Prentice Hall, Upper Saddle River, New Jersey.*

In the development so far we have ignored the fact that in any practical implementation of the above we are required to limit the signal length through applying a window to the signal.

$$s_w(t) = x(t)*v(t)\times w(t) \qquad \text{eqtn.7.5}$$

w(t) = window function

This potentially complicating factor is avoided if we can take the window function inside the convolution. Oppenheim and Schafer[10] have shown that this can be achieved, giving

$$s_w(t) = e(t) \times w(t) * v(t) \qquad \text{eqtn.7.6}$$

on the condition that the window function is sufficiently long. As the method develops we will see in fact that the length, type and placement of the window are very important considerations. Therefore, although the window must be long in order to satisfy eqtn.7.6 there are other considerations relating to harmonic resolution and smoothness of the spectrum, that limit it's length.

Looking at figure 7.2 (c) it can be seen that the vocal tract filter contribution and the periodic glottal excitation have been separated in the cepstrum. This gives rise to the possibility of various liftering operations. The cepstral rahmonics could be masked and the resulting spectrum Fourier transformed to reveal the spectral envelope. Alternatively, the low quefrency signal could be masked and the resultant, Fourier transformed to reveal the glottal source excitation harmonic spectrum. de Krom took advantage of this fact to provide an estimate of the harmonic to noise ratio of the signal. The main point of the analysis is that after inverse Fourier transforming the comb-liftered cepstrum and subtracting the resultant spectral envelope (noise floor estimate) from the original spectrum, an harmonic source spectrum remains. Following application of a baseline correction procedure to this spectrum, a corrected noise floor spectrum is obtained which is subtracted from the original log spectrum in order to acquire the harmonic to noise ratio estimate.

Although an excellent spectral match between the noise and original spectra is obtained using this procedure, it involves a number of steps including three Fourier transforms, log, subtraction and masking operations and a baseline correction which involves taking peaks and between peaks. An alternative approach, requiring one less step, might be to mask all but the rahmonics in the cepstrum and Fourier transform the result to give the source spectrum directly and then continue as above.

However, the basis for the technique outlined here is derived directly from Noll's suggested heuristic approach to the cepstrum in conjunction with considerations of

traditional harmonic to noise ratio estimates based on spectral measurements e.g. Kasuya's NNE[11]. A typical estimate of the harmonic to noise ratio based on spectral calculations involves summing the energy at harmonic locations and dividing by the summed energy of between harmonic locations (eqtn.7.7).

$$\frac{H}{N}(\text{waveform}) = 10 \times \log 10 \left[\frac{\sum_{\omega} X(\omega)V(\omega)}{\sum_{\omega} N(\omega)V(\omega)} \right] \qquad \text{eqtn.7.7}$$

In obtaining the cepstrum, the height of the rahmonic peaks are dependent on the depth of the valleys between adjacent harmonic locations (consider fig.7.3).

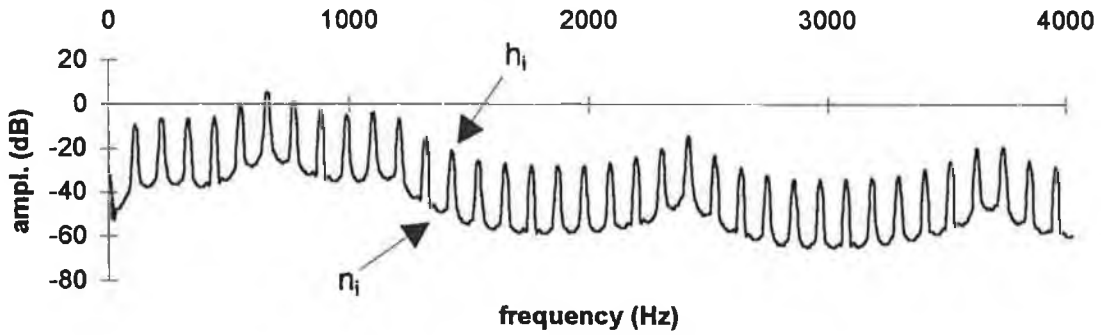


fig.7.3 *Location of i^{th} harmonic and i^{th} noise estimates as per traditional H/N ratio calculation. The spectrum illustrated is for a 110 Hz file with std. dev. 8 % random shimmer.*

In this manner, all the noise contributions can be considered to be contained in the height of the cepstral rahmonic peaks only i.e. they limit the height. Consequently, the height of the rahmonics reflect the harmonic to noise ratio of the source related spectrum (H/N_S), and hence provide an alternative approach for extracting a noise index based on the cepstrum. The height of the rahmonics are not directly related to the H/N ratio of the output radiated speech waveform because they are independent of the actual 'DC' component in the original spectrum. The H/N_S ratio can similarly be obtained from the original spectrum according to eqtn.7.8

$$\frac{H}{N}(\text{source}) = 10 \times \log 10\left[\frac{1}{M}\sum_{\omega=0}^{M}\frac{X(\omega)V(\omega)}{N(\omega)V(\omega)}\right] \qquad \text{eqtn.7.8}$$

where M is the number of harmonics.

It should be noted that eqtn.7.8 only gives the true harmonic to noise ratio of the source when the noise is random i.e. constant at all frequencies. This is also true for the rahmonic peaks and hence the use of the terms 'source related harmonic to noise ratio' and 'source related spectrum'.

Of course, two Fourier transforms are still required as opposed to one but the resultant rahmonic peaks are genearlly fewer in number and more easily located.

Consider, for comparison, obtaining the H/N_S ratio from some form of direct calculation.

1. Log Magnitude Spectrum
2. Locate harmonic peaks (35×110 Hz up to 3.8 kHz)
3. Locate between harmonics (35×110 Hz up to 3.8 kHz)
4. Sum the original ratio at each frequency location

And from the cepstrum

1. Log Magnitude Spectrum
2. Cepstrum
3. Locate rahmonic peaks (11×9.1ms up to 1024 points (one-sided))
4. Sum each rahmonic in order to directly obtain the ratio

So, two advantages of the cepstral technique are readily evident from the above comparison in that there are less points to compute and the ratio is obtained directly from summing these points. The second point is easily explained by considering fig. 7.2 (c) once again and realising that the rahmonic peaks in the quefrency domain

provide a direct representation of the periodicity (and amplitude) of the signal in the frequency domain which, in turn is of course a direct measure of the harmonic to noise ratio of the source related spectrum in dB. There is no problem with respect to adding the rahmonics as these themselves are linear in amplitude even though their overall sum represents a dB ratio.

7.3 Analysis and Results

Points (1) to (4) in the second list is in fact an outline of the method that was actually used for analysis. The source code was for the program (cpphnr.m given in appendix A) was written in the Matlab high level language. Another program was written to implement Hillenbrand's normalisation scheme (cpp.m). Band pass and high pass versions using a 250^{th} order, finite impulse response filter, were also coded. In the cpphnr.m file a window length of 2048 points was used (fig.7.4). This followed from actual investigation of different window lengths, a consideration of eqtn.7.6 and de Krom's observations :

> "The potential positive influence of a longer analysis window on HNR, related to a higher frequency resolution, has a negative side effect. At higher perturbation levels, the harmonic bandwidth increases. If we now increase the frequency resolution by increasing the length of the analysis window, we will observe the emergence of spiky subharmonics, rather than a mere broadening of the harmonic bandwidth. This breaking apart of harmonics in distinct energy spikes results in a less coherent harmonic structure, with a negative influence on HNR."

In order to test the programs in a systematic way they were applied to the synthetically generated signals listed in Table.2.5 (chapter 2). The set consists of three different noise levels for six different fundamental frequencies ranging from 80 Hz to 350 Hz and therefore covering the extremes of the expected vocal pitch range. If the method reflected the relative noise levels correctly for these signals, it was subsequently tested on the jitter (both random and cyclic) and shimmer signals. Figure 7.5 shows the H/N_S ratios obtained for the variation with additive noise of the glottal source with a std. dev. of 4 %, 8 % and 16 % for various fundamental frequencies. The trend of

increasing harmonic to noise ratio with increasing fundamental frequency, for the synthesis data, is explained in section 5.5.1 (chapter 5).

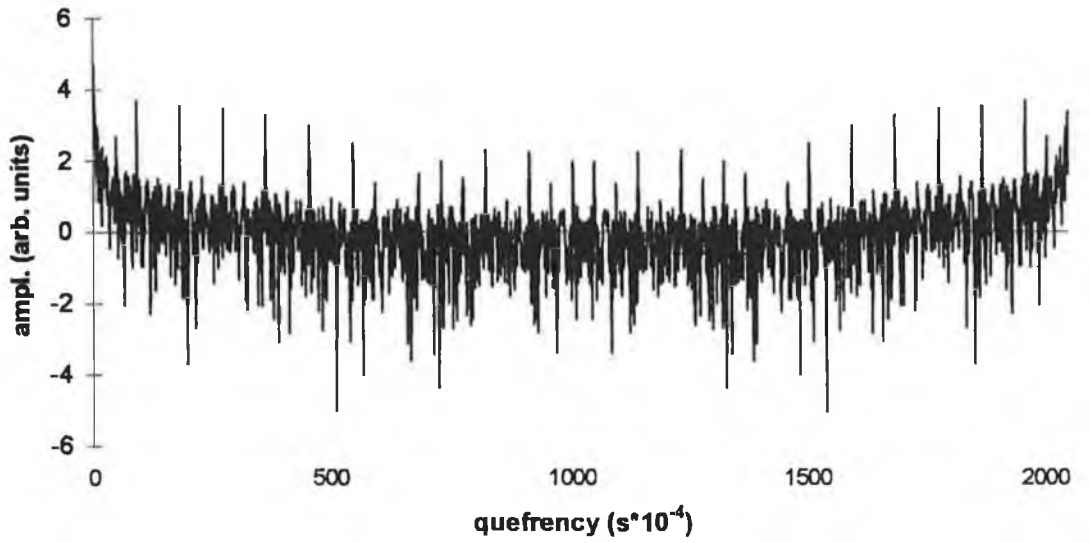


fig.7.4 *A 2048 point real cepstrum of unperturbed 110 Hz synthesis file.*

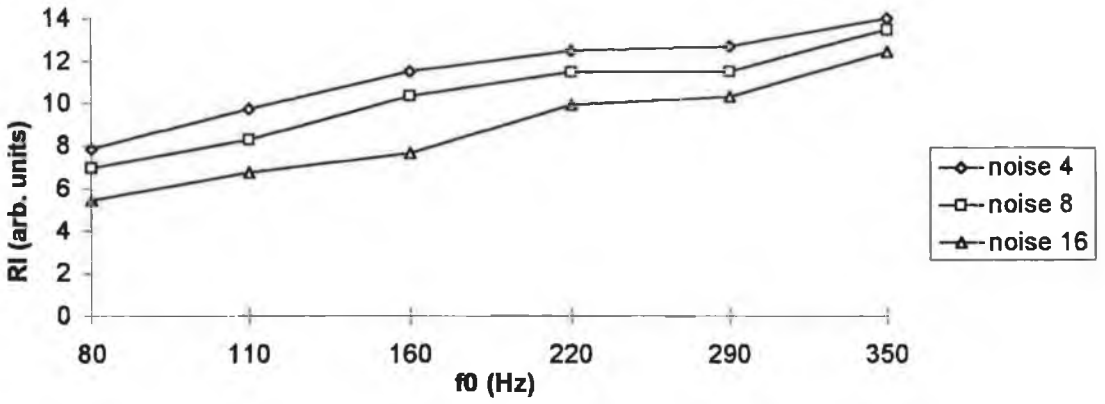


fig.7.5 *Source related harmonic to noise ratio (H/N_S) vs f0 for three levels of random source noise.*

Hillenbrand's normalisation scheme is shown in fig.7.6 where a regression line is fitted to the dB noise level. The level of the first rahmonic with respect to this regression line was taken by Hillenbrand to be an indicator of periodicity. This was called the cepstral peak prominence (CPP) and he found a good correlation between this measure and the perceived breathiness of his subjects. Fig.7.7 shows this measure plotted against f0

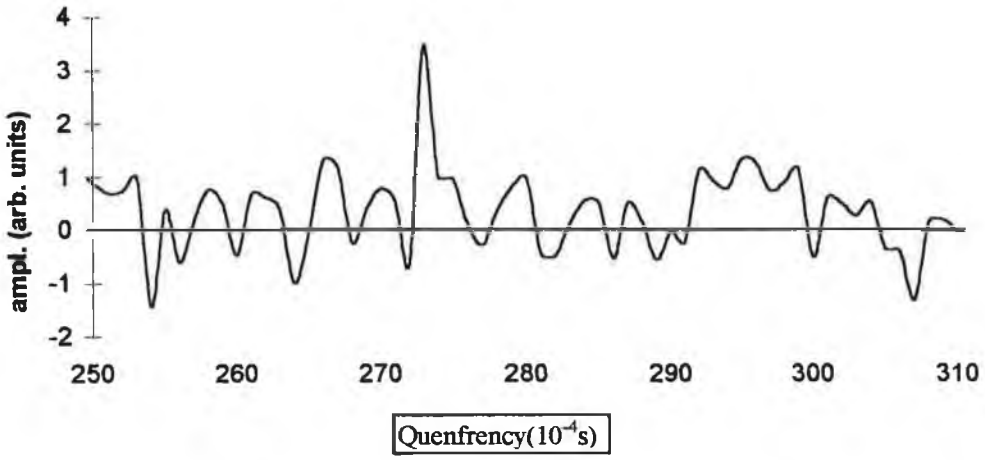


fig.7.6 *Regression line fitted to cepstrum. The first rahmonic peak is calculated in dB with respect to the regression line.*

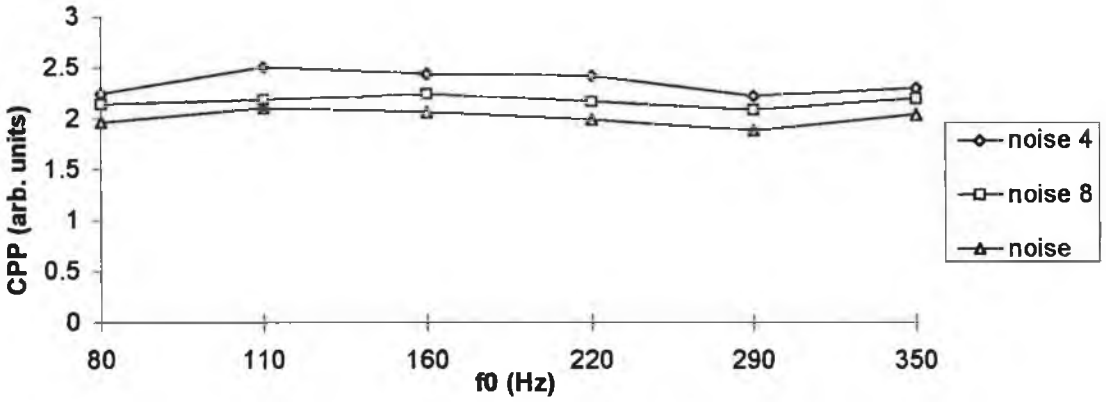


fig.7.7 *CPP plotted against f0 for three levels of additive noise.*

with the three additive noise levels. It can be seen that the increase in additive noise is reflected faithfully at each frequency whereas the f0 dependence of the measure is not evident. This is due, in part at least, to the fact that we are using a single cepstral peak. A 51.2 ms window was used in our study as opposed to Hillenbrand's 25.6 ms window (recall the requirement for eqtn.7.6 to hold). The same process was carried out measuring the level of all rahmonics with respect to the regression line and a similar curve to the CPP measure was obtained. However these dB rahmonic values were simply averaged using an arithmetic mean when perhaps a geometric mean would have been more appropriate. Band pass and high pass versions of these programs were also

run on the data but for the CPP measure the filtered versions did not reflect the noise increases. Hillenbrand did not employ a pitch tracking routine in implementing his technique and found that not only did this have no adverse effects it actually increased the breathiness prediction indicator. However, we found it essential to locate the actual rahmonic peaks, especially in the case when using all rahmonics, otherwise the method was found to give more erratic results (fig.7.8).

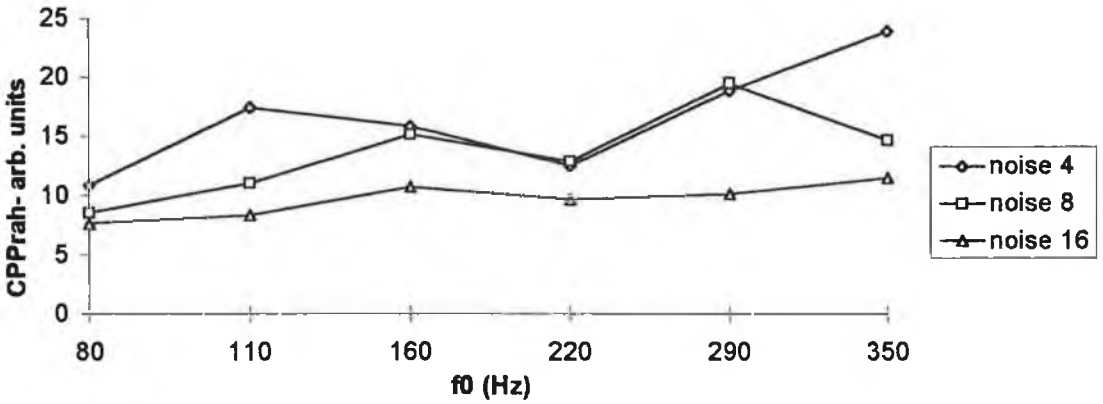


fig.7.8 *CPPrah vs f0 showing increased variability due to absence of pitch tracking.*

Hillenbrand interpreted his results to mean that since reduced CPP correlated very well with breathiness, that aperiodicity is a strong indicator of breathiness. Although not explicitly stated the implication is strongly taken to refer to periodicity of the signal in the time domain. However, we feel that the reduced rahmonic peak in breathiness is not due to aperiodicity of the time domain waveform. One of the main indicators of breathiness has often been reported to be an increased first harmonic amplitude[12] along with aspiration noise. The former acoustic parameter (increased f0 amplitude) has direct aerodynamic and physiological correlations in the form of increased volume velocity and more abducted vocal folds respectively. So, it is a well accepted breathiness indicator. In the case of breathy signals the reduced amplitude of the cepstral peak is also primarily due to the increase in the amplitude of the first harmonic. To understand this, we see that the periodicity in the frequency domain is offset by this increase in f0 amplitude, leading to a less obvious 'separation of the log'. We note then that periodicity in one domain does not necessarily indicate periodicity in another. On

the contrary, we observe that a sharp peak in one domain corresponds to a more broadened (sinusoidal) event in the other domain (in fact, this is the reverse of the cepstrum). Therefore, perfect periodicity (sinusoid) can exist in the time domain and yet no cepstral peak is found at the expected quefrency location. This in no way limits the cepstrum for investigating breathy signals, on the contrary, Hillenbrand found an excellent correlation but our inference is that this is due to the exploitation of the reduced ability to 'separate the log' and has little to do with aperiodicity in the time domain. The response of the H/N_S, CPPrah and CPP indices to all perturbation measures is shown in fig.7.9 (a), (b) and (c).

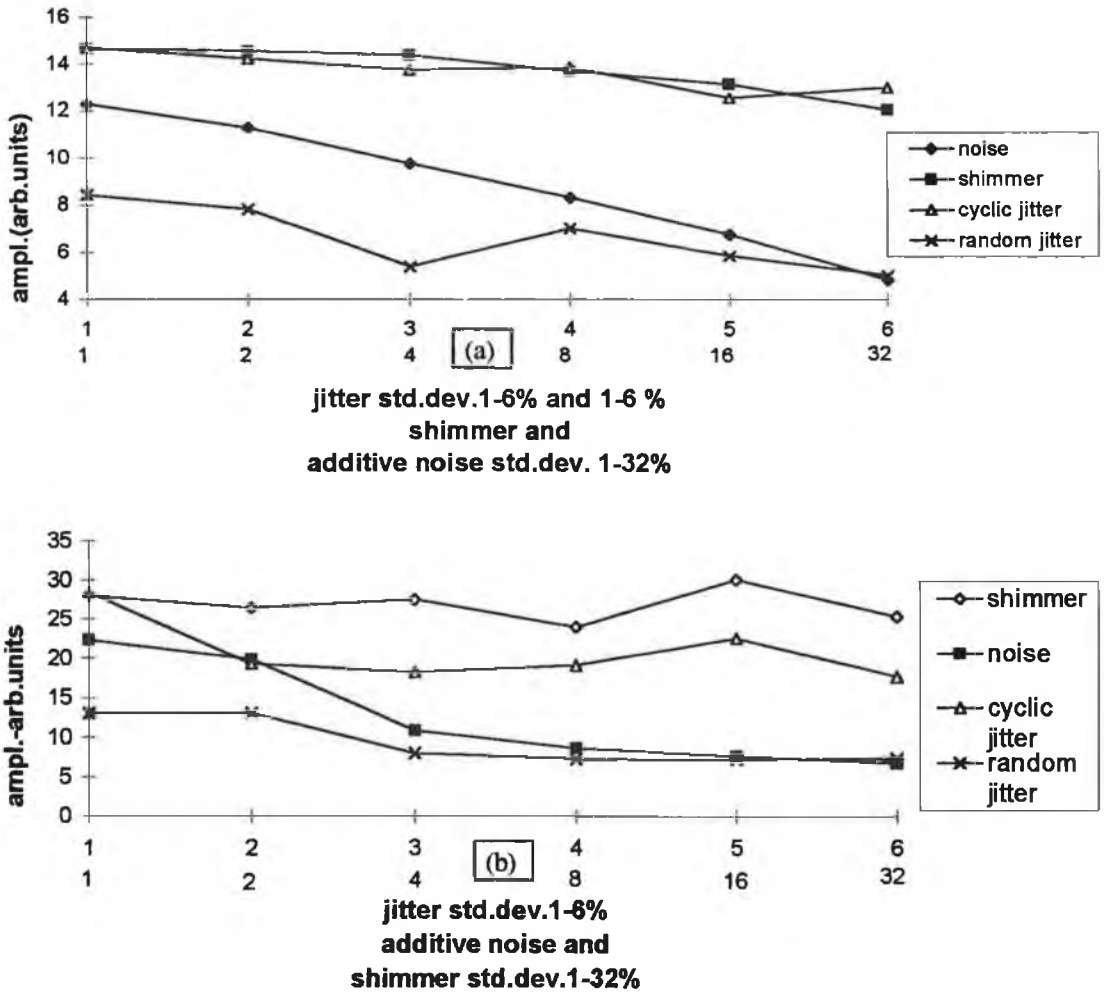


fig.7.9 (a) *H/N_S and* (b) *CPPrah vs perturbation measures, where noise is more linearly reflected for H/N_S with both methods being sensitive to random jitter.*

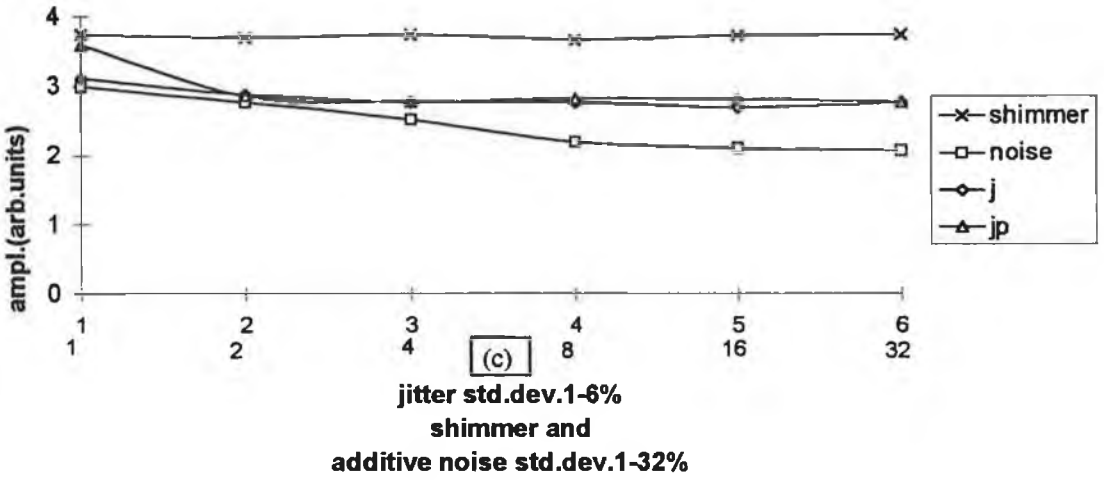


fig.7.9 (c) *CPP vs perturbation measures, showing insensitivity to shimmer and somewhat less sensitivity to random jitter that H/N_S and CPPrah, with increases in noise levels not linearly reflected.*

With reference to part (a) and (b) of fig.7.9, particularly part (a), the relative insensitivity of the measures to both cyclic jitter and random shimmer is very obvious. In consideration of the spectral characteristics of each of these sources of perturbation, with cyclic jitter containing subharmonics and shimmer resulting in an H/N ratio that is equal at all frequencies, it can be seen (fig.7.10 (c) and (d)) that good harmonic structure remains throughout the spectrum. In the case of shimmer, the increased height of the valleys between harmonic locations is seen to have a relatively small effect on the cepstrum calculation. In contrast to this are the random jitter and additive noise spectra (fig.7.10 (a) and (b)) which still show early harmonic structure which quickly deteriorates with increasing frequency. The cepstrally based indices reflect these more severe alterations in harmonic structure.

In summary, the source related index (H/N_S) seems to give a good estimate of the signal to noise ratio and it is also affected by jitter. The CPP measure also reflects the H/N_S ratio quite well but the trend seems less reliable. The regressed rahmonics also show some indication of the H/N_S ratio but a detrimental effect is found rather than an improvement in the method. The filtered versions were unsuccessful in following the trend of fig.7.5.

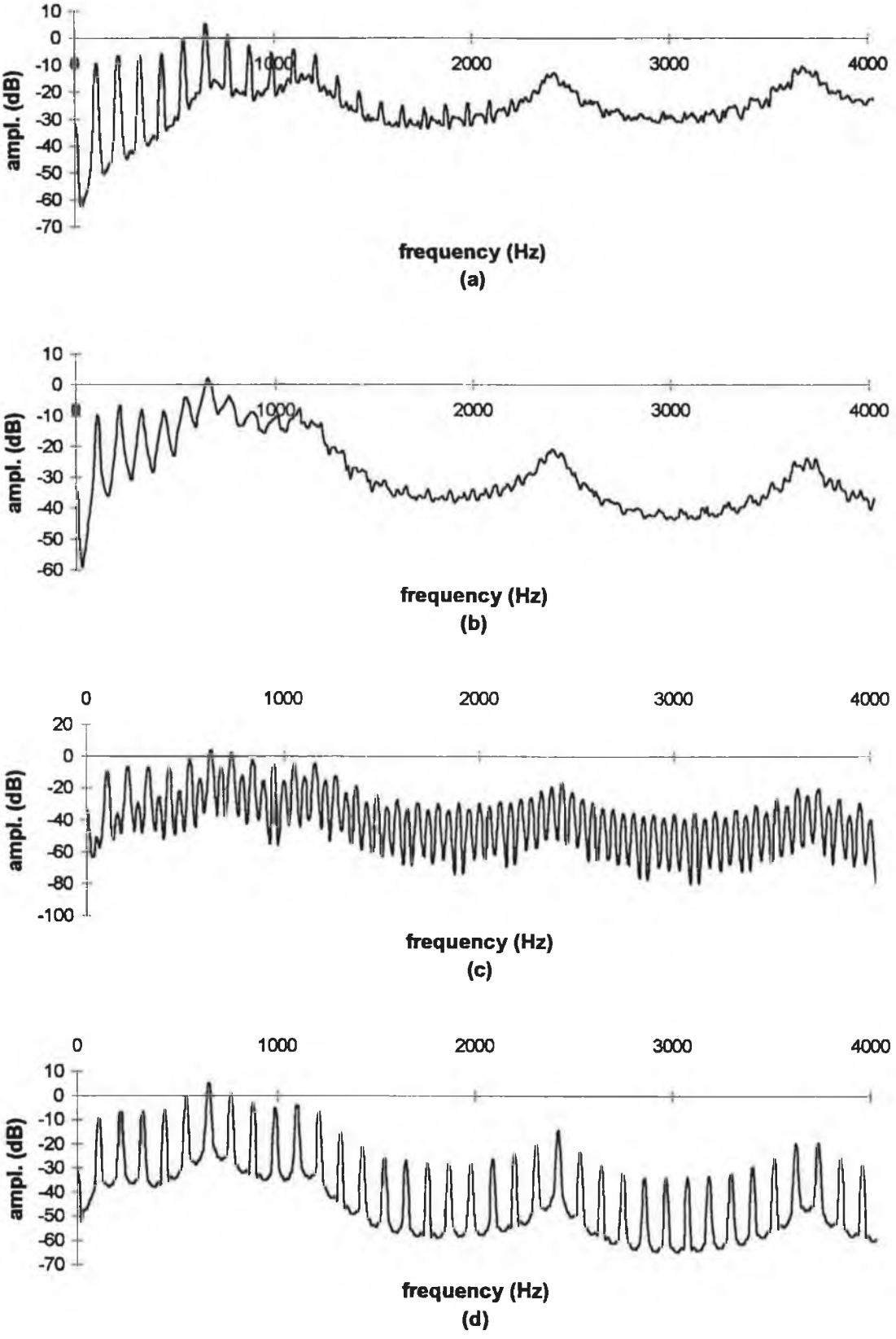


fig.7.10 *Spectra for* (a) *8 % std. dev. additive noise*, (b) *4% std.dev. random jitter*, (c) *4% cyclic jitter and* (c) *std.dev.8 % random shimmer.*

All methods were applied to the patient data in attempt to separate the patients from the normals (fig.7.11 and fig.7.12). The CPP method shows reasonable ability in separating the patient/normal data set with it's filtered versions showing no discriminatory ability. The regressed rahmonics show some degree of separability. However, the source related harmonic to noise ratio (H/N_S) gives the best overall discrimination, being highly significant at the 5 % level (one tailed, equal variance, two sample mean, student's t-test). Two encouraging hypotheses are made based on these results : firstly, the method is potentially a good indicator of vocal pathology (fig.7.11) and secondly, the synthesis files are in some way representative of the artifacts found in actual vocal pathologies.

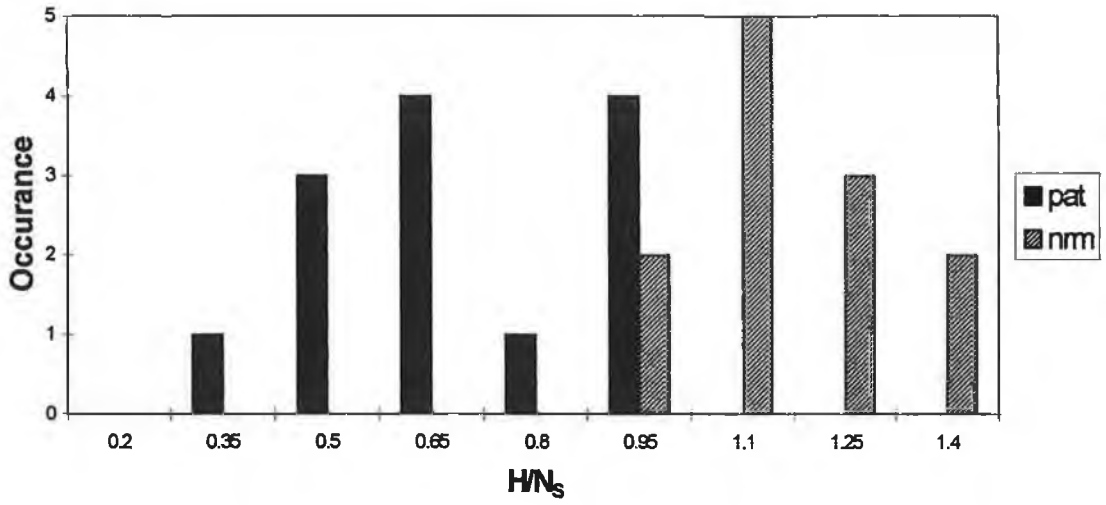


fig.7.11 *Source related H/N_S index showing good separability of the patient/normal data set. (Highly significant at the 5% level using a one tailed, two sample, equal variance, student's t-test).*

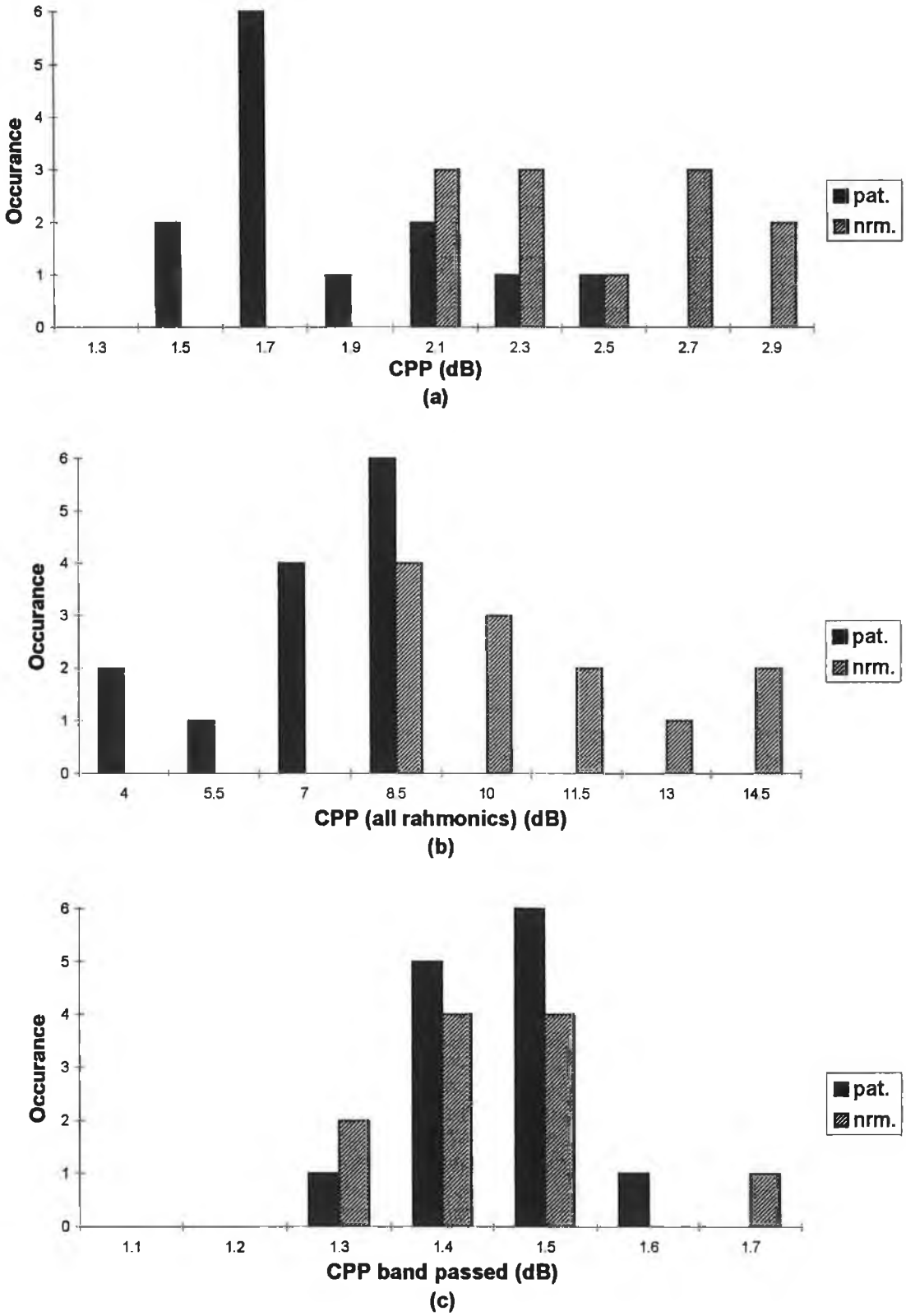


fig.7.12 *CPP measures* (a) *first rahmonic (CPP),* (b) *all rahmonics (CPPrah) and* (c) *bandpassed, CPPb, with (a) and (b) showing significant (at the 5% level) separability of the patient/normal data set but not as high as* H/N_S.

7.4 Conclusion:

The main points can be summarised as follows:

1. A new cepstral measure, the source related harmonic to noise ratio (H/N_S) has been defined and tested on tested on synthetically generated signals and also on a set of patient and normal productions of the vowel a/.
2. The height of the cepstral rahmonics have been shown to be directly related to the harmonic to noise ratio of the 'source related spectrum', where the source related spectrum can be obtained, for example, by comb liftering the cepstrum and subtracting the result from the original log spectrum.
3. The reduction in the ability to 'separate the log' when signals become more dominantly sinusoidal (i.e. reduced richness of harmonics) has been proposed as an index for breathiness.
4. The H/N_S has been shown to be a potentially useful indicator of vocal pathology, reflecting additive noise levels accurately and discriminating between a set of 13 patients with varying vocal pathologies and a group of 12 'normals' with statistical significance.
5. Absence of pitch tracking and band passing result in less reliable indices.

In conclusion, the cepstrum seems to offer three (although perhaps not completely independent) indices for evaluating vocal pathology, namely, the H/N ratio as implemented by de Krom, the H/N_S ratio as developed in this chapter and the 'separation of the log' factor taken advantage of by Hillenbrand. Future studies might include other measures taken from the 'source derived spectrum'.

7.6 Bibliography

1. Bogert, BP. et al In M. Rosenblatt (Ed.) Proc. of the Symp. On time series analysis. NY: John Wiley and Sons 1963, pp.209-243
2. Schroeder, MR. Vocoders: Analysis and synthesis of speech. Proc. IEEE 1966; **54**: 720-734
3. Noll, AM. Cepstrum pitch determination. J. Acoust. Soc. Am. 1966; 293-309
4. Deller, J. Proakis J. and Hansen, J. Discrete time processing of speech signals, NY: Macmillan, 1989
5. Furui, S. Cepstral analysis technique for automatic speaker verification. IEEE Trans. on Acoust., Speech and Signal Processing 1981; **29**: 254-272
6. Oppenheim, A.V. and Schafer, R.W. Discrete-time signal processing. Englewood Cliffs, N.J.: Prentice Hall, 1989
7. Koike, Y. and Kohda, J. The effect of vocal fold surgery on the speech cepstrum. In J. Gaufin and B. Hammarberg (Eds.) Vocal fold physiology: Acoustic, perceptual and physiologic aspects of voice mechanisms. San Diego: Singular Publ. 1991
8. J. Hillenbrand, J. et al Acoustic correlates of breathy vocal quality. J. Speech and Hear. Res. 1994; **37**: 769-777
9. de Krom, G. A cepstrum based technique for determining a harmonics-to-noise ratio in speech signals. J. Speech Hear Res. 1993; **36**: 254-266
10. Deller JR. et al Discrete time processing of speech signals, New York:Macmillan, 1993
11. Kasuya, H. et al Normalised noise energy as an acoustic measure to evaluate pathologic voice. J. Acoust. Soc. Am. **80**: 1329-1334
12. Klatt, DH. and Klatt, LC. Analysis, synthesis and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 1990; **87**: 820-857

Conclusion

Presently used acoustic indices provide useful supplementary evidence for documenting pathological voice types. The finding as laid out in the present thesis support this fact, showing that indices do exist that separate pathologic voice types from 'normal' voices. For example, bandlimiting the frequency range from 1-4 kHz and calculating the harmonic to noise ratio appears to provide a reliable indicator of dysphonia. Also, new indices have been introduced in this thesis that successfully separate pathologic voice types from 'normal' voices. However, as presently implemented, these indices offer only limited information regarding the exact physical nature of the voice disorder.

In order to achieve the ultimate goal of providing accurate clinical diagnoses of voice disorders, further basic research is required i.e. research that relates more specifically to the relationship between anatomical events and the resultant acoustic sound pressure waveform. This requires better knowledge of physiological function during phonation, the extraction of pertinent information from the acoustic speech waveform and a clear understanding of the relationship between the acoustic speech waveform and the underlying vibratory pattern.

In spectral analysis of pathological voice types, the gross spectral features of jitter, shimmer and additive noise, contaminate useful spectral information relating to the vibratory pattern of the vocal folds. This problem has been addressed in two ways. Firstly, the spectral characteristics of jitter, shimmer and additive noise have been determined. Therefore, based on this information, pertinent spectral measures will reveal what perturbation type is present. Secondly, a pitch synchronous harmonic

intensity analysis approach has successfully been developed to eliminate the effects of the perturbation measures and therefore provide more reliable information regarding the vibratory pattern of the vocal folds. Simple models have proven very useful for providing quantitative information regarding speech-like material. However, in conjunction with modelling, more research regarding the physiological and neurological bases of voice disorders is required, as well as improved correlations between specific voice pathology types and acoustic findings. The final word is left to Ingo Titze:

"There is a fallacy in trying to characterise the voice by a single number. There is a very complex pattern in the voice signal, and this is the thing that we should be attempting to describe. After viewing the voice in all it's complexity and leaning more about it's subtleties, then we may be in a position to return and attempt to describe the voice through one or two measures. Prior to this, we have to spend time just looking at the voice, say as one views a picture of the vocal folds or of vocal fold movements. After we look at enough pictures, maybe then we will be in a better position to come up with the few quantitative measures that are the most useful for describing the vocal folds during voice production"[1].

Bibliography:

1. Titze, IR. In PJ. Davis and NH. Fletcher (Eds.) Vocal Fold Physiology: Controlling Complexity and Chaos, San Diego: Singular Publ., 1996 pp.417

Appendix A

Source Code for Principal Matlab Program Files

A.1 Time Domain Analysis

A.1.1 ppitch3.m

```
%%%%%%%%%%
%
% File:   PPitch3.m
%
% Name:   Peter Murphy
%
% Date:   Mon. 24_02_'97
%
% Descr.  A time domain pitch extraction method based on zero crossings, +ve & -ve %          peaks
%          from a low passed filtered version of the acoustic speech signal.
%          The pitch ampl. is also returned via PPs
% Call:   [Noutp,Noutn,Nout1,Nout2,fointlpp,fointlnp,folpp,folnp,sPPs ] =
%          PPitch3(sp);
%
%%%%%%%%%%

  function  [Noutlp,Noutln,Noutl1,Noutl2,fointlpp,fointlnp,folpp,folnp,fointpp,fointnp,fopp,fonp,sPPs]
= PPitch3(sp);

 hop  = 100;
 fsam = 10000;
```

```
 PERC = 0.4;
 fsam=10000;
%%%%
%   Band pass filter the waveform (FIR (200-300) ). (60Hz-hp)
%   Set the rough markers (-ve going zero crossings) for pitch extraction.
%%%%

  [sect] = sonasect(sp);

  [f0_est,nout] = fester2(sect,hop);

% 1=fs/2 i.e. 5000 Hz ... 0.1=500 Hz.

  b = fir1(250,[2*60/fsam,2*1.5*f0_est/fsam]);
  l = filtfilt(b,1,sect);


  length(l);

  j=1;
  for i=1:length(l)-1              % Obtain the number of
  if l(i)>0&l(i+1)<0               % negative going zero crossings.
                                % NZC.
   zc_in(j) = i ;
   j=j+1;
  end
  end

  NZC=j-1

 P(1)=0;
 PP(1) = 0;
 PPs(1)=0;
 for j=2:NZC-1;
 [y,in] =sort(l(zc_in(j-1):zc_in(j)));
 [sy,sin] = sort(sect(zc_in(j-1):zc_in(j)));
 P(j)=zc_in(j-1)+in(1);
 PP(j)=zc_in(j-1)+in(length(l(zc_in(j-1):zc_in(j))));
 Ps(i)=zc_in(j-1)+sin(1);
 PPs(j) = zc_in(j-1)+sin(length(l(zc_in(j-1):zc_in(j))));
 PPrl(j) = PP(j)+(-0.5*(l(PP(j)+1)-l(PP(j)-1)))/(l(PP(j)+1)-2*l(PP(j))+l(PP(j)-1));
 Prl(j) = P(j)+(-0.5*(l(P(j)+1)-l(P(j)-1)))/(l(P(j)+1)-2*l(P(j))+l(P(j)-1));
 PPsrl(j)  =  PPs(j)+(-0.5*(sect(PPs(j)+1)-sect(PPs(j)-1)))/(sect(PPs(j)+1)-2*sect(PPs(j))+sect(PPs(j)-
1));
% Psrl(j) = Ps(j)+(-0.5*(sect(Ps(j)+1)-sect(Ps(j)-1)))/(sect(Ps(j)+1)-2*sect(Ps(j))+sect(Ps(j)-1));
 end
                                  % Poly interpolation
                                  % PERC error detection

for j=6:NZC-4
 fintpp(j-5) =fsam/(PPsrl(j)-PPsrl(j-1));
% fintnp(j-5) =fsam/(Psrl(j)-Psrl(j-1));
 fnp(j-5)   =fsam/(Ps(j)-Ps(j-1));
 fpp(j-5)   =fsam/(PPs(j)-PPs(j-1));
 fintlpp(j-5)=fsam/(PPrl(j)-PPrl(j-1));
 fintlnp(j-5)=fsam/(Prl(j)-Prl(j-1));
```

```
 flnp(j-5)   =fsam/(P(j)-P(j-1));
 flpp(j-5)   =fsam/(PP(j)-PP(j-1));
end

sPPs=sect(PPs(2:length(PPs)));

% Last check!!!
% Make sure that the f0s fall within an acceptable level.
% PERC

fointlpp=fintlpp(                    fintlpp<mean(fintlpp)+mean(fintlpp)*PERC&fintlpp>mean(fintlpp)-
PERC*mean(fintlpp));
disp('no. of outliers');
Noutp = length(fintlpp)-length(fointlpp)
fointlnp=fintlnp(                    fintlnp<mean(fintlnp)+mean(fintlnp)*PERC&fintlnp>mean(fintlnp)-
PERC*mean(fintlnp));
disp('no. of outliers');
Noutn = length(fintlnp)-length(fointlnp)
folpp=flpp( flpp<mean(flpp)+mean(flpp)*PERC&flpp>mean(flpp)-PERC*mean(flpp));
disp('no. of outliers');
Nout1 = length(flpp)-length(folpp)
folnp=flnp( flnp<mean(flnp)+mean(flnp)*PERC&flnp>mean(flnp)-PERC*mean(flnp));
disp('no. of outliers');
Nout2 = length(flnp)-length(folnp)
fointpp=fintpp(                      fintpp<mean(fintpp)+mean(fintpp)*PERC&fintpp>mean(fintpp)-
PERC*mean(fintpp));
disp('no. of outliers');
Noutp = length(fintpp)-length(fointpp)
%fointnp=fintnp(                     fintnp<mean(fintnp)+mean(fintnp)*PERC&fintnp>mean(fintnp)-
PERC*mean(fintnp));
%disp('no. of outliers');
%Noutn = length(fintnp)-length(fointnp)
fopp=fpp( fpp<mean(fpp)+mean(fpp)*PERC&fpp>mean(fpp)-PERC*mean(fpp));
disp('no. of outliers');
Nout1 = length(fpp)-length(fopp)
fonp=fnp( fnp<mean(fnp)+mean(fnp)*PERC&fnp>mean(fnp)-PERC*mean(fnp));
disp('no. of outliers');
Nout2 = length(fnp)-length(fonp)
plot(fointpp);
hold on
plot(fointlpp,'r');
hold off
pause


Hf0=sect(PPs(6:NZC-4));        % index of original waveform +ve peaks
Hf0l=l(PP(6:NZC-4));           % index of low pass +ve peaks

plot(Hf0);
hold on
plot(Hf0l,'r');
hold off
disp('Perturbation measures from unfiltered waveform');
[app,rapp,appp,stdndf0,PF1,PF2,DPF,stdnd2f0,stddf0] = supperb1(fointpp);
disp('Perturbation measures from filtered waveform');
[app,rapp,appp,stdndf0,PF1,PF2,DPF,stdnd2f0,stddf0] = supperb1(fointlpp);
```

```
disp('Ampl.Perturbation measures from unfiltered waveform');
[aap,raap,apap,stdndHf0,HPF1,HPF2,DHPF,stdnd2Hf,stddHf0,dBdHf0] = ampere1(abs(Hf0));
%disp('Ampl.Perturbation measures from filtered waveform');
%[aap,raap,apap,stdndHf0,HPF1,HPF2,DHPF,stdnd2Hf,stddHf0,dBdHf0] = ampere1(abs(Hf01));
% 5. Mean first order perturbation

%   dHf01=diff(Hf01);
%   HPF1 =  mean(abs(dHf01)./two_pt(Hf01))*100;
%   fprintf('HPF1(Hf01) = %.4f\n',HPF1);


 %   PF1 =  mean(abs(diff(fointpp))./two_pt(fointpp))*100;
 %   fprintf('PF1pp = %.4f\n',PF1);

 %   PF1 =  mean(abs(diff(fointlpp))./two_pt(fointlpp))*100;
 %   fprintf('PF1ppl = %.4f\n',PF1);
%[l,fowav] = wavmat(sect,f0_est);
[Nout1,Nout2,fintlpc,flpc] = pzclpit(sect,f0_est);
```

A.1.2 pperb.m

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Program: pperb.m Matlab program to calculate the pitch variation
%                    or perturbation factor.
%
% Name:    Peter Murphy
%
%
% Date:    Mon. 24-02-'97
%
% Aim:     To calculate the pitch perturbation in the speech signal
%
% Call:    [app,rapp,appp,stdndf0,PF1,PF2,DPF,stdnd2f0,stddf0] = pperb(f0);
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% To calculate an indices for jitter (period/f0 perturbation).


function [app,rapp,appp,stdndf0,PF1,PF2,DPF,stdnd2f0,stddf0] = pperb(f0);


% pitch measurements

%disp('f0 measurements');
 mf0=mean(f0);
 sf0=std(f0);
 mdf0=median(f0);

% fprintf('aver. fundamental freq. = %6.3f\n', mf0);
% fprintf('std. f0  = %6.3f\n', sf0);
```

```
% fprintf('median f0 = %6.3f\n', mdf0);

df0=diff(f0);
d2f0 =diff(df0);

% perturbation analysis
% 1. average pitch perturbation (app)

   app=mean(abs(df0));

% 2. relative average pitch perturbation

   rapp = mean(abs((three_pt(f0))-f0(2:length(f0)-1)))/mean(f0)*100;

% 3. average percentage pitch perturbation

   appp = mean(abs(df0)./f0(2:length(f0)))*100;

% 4. standard deviation of the pitch perturbation divided by f0

    stdndf0 = std(df0./two_pt(f0));

% 5. Mean first order perturbation

    PF1 =  mean(abs(df0)./two_pt(f0))*100;

% 6. Mean 2nd order perturbation


    PF2 = mean(abs(d2f0)./three_pt(f0))*100;


% 7 Directional perturbation factor
       k=0;
   for i = 1:length(df0)-1
     if df0(i)>0&df0(i+1)<0|df0(i)<0&df0(i+1)>0
       k=k+1;
     end
   end


  DPF = k/length(df0)*100;


% 8 standard deviation of second order pitch perturbation divided by f0


   stdnd2f0 = std(d2f0./three_pt(f0));


% 9 standard deviation of df0
```

```
  stddf0 = std(df0);
%%%%%
%
% Display results in column form
%
%%%%

% disp('Perturbation measures');
  disp(' app    rapp    appp    stdndf0    PF1    Pf2 ');
  disp( [app        rapp        appp        stdndf0        PF1        PF2 ] );
  disp(' DPF   stdnd2f0   stddf0   mf0   sf0    mdf0');
  disp([ DPF         stdnd2f0         stddf0      mf0      sf0     mdf0 ]);
```

A.1.3 amperb.m

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Program: amperb.m Matlab program to calculate the pitch amplitude
%          perturbation factor.
% Name:  Peter Murphy
% Date:  Mon. 24-02-'97
% Aim:   To calculate the pitch amplitude perturbation in the speech signal
%
%    Call:      [aap,raap,apap,stdndHf0,HPF1,HPF2,DHPF,stdnd2Hf,stddHf0,dBdHf0]  =  %
amperb(Hf0);
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% To calculate an indices for shimmer (f0 amplitude perturbation).

function [aap,raap,apap,stdndHf0,HPF1,HPF2,DHPF,stdnd2Hf,stddHf0,dBdHf0] = amperb(Hf0);
 % pitch amplitude measurements

 disp('Hf0 measurements');
```

```
mHf0=mean(Hf0);
sHf0=std(Hf0);
mdHf0=median(Hf0);
% fprintf('ampl. fundamental freq. = %6.3f\n', mHf0);
% fprintf('std. Hf0  = %6.3f\n', sHf0);
% fprintf('median Hf0 = %6.3f\n', mdHf0);

 dHf0=diff(Hf0);
 d2Hf0 =diff(dHf0);

 % amplitude perturbation analysis

 % 1. average amplitude perturbation (aap)

   aap=mean(abs(dHf0));

% 2. relative average pitch amplitude perturbation

   raap = mean(abs(three_pt(Hf0)-Hf0(2:length(Hf0)-1)))/mean(Hf0)*100;

% 3. average percentage pitch amplitude perturbation

   apap = mean(abs(dHf0)./Hf0(2:length(Hf0)))*100;

% 4. standard deviation of the pitch amplitude perturbation divided by f0

    stdndHf0 = std(dHf0./two_pt(Hf0));
% 5. Mean first order perturbation

    HPF1 =  mean(abs(dHf0)./two_pt(Hf0))*100;

% 6. Mean 2nd order perturbation

    HPF2 = mean(abs(d2Hf0)./three_pt(Hf0))*100;

% 7 Directional amplitude perturbation factor
        k=0;
```

```
  for i = 1:length(dHf0)-1
    if dHf0(i)>0&dHf0(i+1)<0|dHf0(i)<0&dHf0(i+1)>0
      k=k+1;
    end
  end
  DHPF = k/length(dHf0)*100;

% 8 standard deviation of second order pitch amplitude perturbation divided by Hf0

  stdnd2Hf = std(d2Hf0./three_pt(Hf0));

% 9 standard deviation of dHf0

  stddHf0 = std(dHf0);

% 10 Average power differences-dB

   dBdHf0 = mean(diff(20*log10(Hf0)));


% Display results in row form
 disp('Perturbation measures');
 disp('  aap     raap     apap    stdndHf0    HPF1    HPf2 ' );
 disp(  [aap     raap     apap    stdndHf0     HPF1     HPF2 ]);
 disp('  DHPF   stdnd2Hf  stddHf0   dBdHf0 ' );
 disp([ DHPF   stdnd2Hf   stddHf0  dBdHf0 ]);
 disp(' mHf0   sHf0    mdHf0 ');
 disp([mHf0   sHf0     mdHf0  ]);
```

A.2 Harmonic Intensity Analysis

A.2.1 Noise Reducing Filter

```
% Program: Kitnos5.m
%
% Date :  Thurs. 21-04-'97
%
% Call: [NR,dBNR,geoNR]=kitnos5(sp);
%
% Name: Peter Murphy
%
% Aim: A modified version of Kitajima's method of using
%      a mov-av filter to estimate the noise levels.
%      The no. to aver depends on f0.
%      Combines kit3&kit4 to give 3 ratios.
%
%%%%%%%%%%%%

   function [fsect,hsect,mavstd,nsect,NR,dBNR,geoNR]=kitnos5(sp);


     pad  = 2048;
     fsam = 10000;
     olap = 1024;
     f_cut= 3800;
     df   = fsam/pad;
     len  = 2048;


 %  [sect1] = sonasect(sp);
   [sect2] = sonastar(sp,3*len);


   fsect = psd(sect2,pad,fsam,len,olap)';
   fsect2= psd(sect2,pad,fsam,512)';
   dBfsect = 10*log10(fsect);
   [f0_est,nout] = fester2(sect2,olap);
   f0_est
   m_avlen=round(f0_est/df)-2;
   if rem(floor(m_avlen),2)==0
   m_avlen=m_avlen+1;
   end
%    m_avlen=81;
   [m_stdsect , m_avfsect] = mov_av(fsect,m_avlen);



%%%%%%%%%%
%
```

```
% Use the filtered spectrum to eliminate the
% noise energy.
%
%%%%%%%%%%

 mavstd=m_avfsect+m_stdsect;

 for i=1:length(fsect)

  if fsect(i)>(mavstd(i));

    hsect(i) = fsect(i)-(mavstd(i));

  else

    hsect(i)=0;

  end
 end


%%%%%%%%%%
%
% Now, use the filtered spectrum to estimate the
% noise energy.
%
%%%%%%%%%%

 for i=1:length(fsect)

  if fsect(i)>(mavstd(i));

    nsect(i) = (mavstd(i));

  else

    nsect(i)=0;

  end
 end


   clf
   plot(dBfsect(1:500));
   pause
   hold on
 dBfsect2=10*log10(fsect2);
   plot(dBfsect2(1:500),'r');
   hold on

plot(10*log10(hsect(1:500)),'g');
title('Noise Reduced Periodogram');
hold on
plot(10*log10(nsect(1:500)),'b');

NR =   ( (mean(hsect.^2)).^0.5/mean(nsect.^2).^0.5);
```

```
dBNR=10*log10(NR);
hold off
pause
clf
[hsect,mavstd,geoNR]=kitdB(dBfsect,m_avlen);
fprintf('NR=%6.3f\t',NR);
fprintf('dBNR=%6.3f\n',dBNR);
```

A.2.2 Harmonic Intensity (Hiraoka)

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%  Program:  harmony4.m
%
%  Date:    03-03-'97
%
%  Call:    [Hr,Sr,dBH,dBS] = harmony4(sp_data);
%
%  Note:    make len =2048 and
%           padded to 4096. Test db ratio.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

  function [spec_amp,Hr,Sr,dBH,dBS] = harmony4(sp_data);

 % Let the operator choose the region for analysis as usual.
 % variables:


   olap=100;
  f_cut=3800;
   fsam=10000;
    pad=4096;
    len=2048;
     df=fsam/pad;
     bw=3*pad/len;

% Plot the speech sample
% Choose the region for analysis
% (Less than .4096 s)
% Plot the spectrum for this region
% (padded out to 4096)

  subplot(3,1,1);
  sonagram(sp_data,fsam);
  [x y] = ginput(1);
  data= sp_data((x(1)*fsam):(x(1)*fsam+len-1));
  subplot(3,1,2);
  sonagram(data,fsam);
  [spec_amp,spam] = specam2(data,pad);
  subplot(3,1,3);
  specplot(spec_amp,pad);
  disp('Press return to close figure window');
```

```
pause
clf
% length(spec_amp)
% subplot(3,1,1);
% dBplot(

% Plot the spectrum from 70 to 400 Hz
% Determine f0 from the usual cepstral analysis.

speclow = spec_amp((70/df):(400/df));
subplot(3,1,2);
plot(speclow);

[f0,nout] = fester2(data,olap);
f0

% Find the harmonic frequencies and sum the energies at these frequency
% locations (exclude f0).
% Find the total signal energy.
% Hence determine the relative harmonic intensity (Hr).


% Total Harmonic Amplitude (THA1-incl. f0, THA-excl. f0)
% Total Signal Amplitude (TSA)


spec_har=zeros(size(spec_amp(1:f_cut/df+bw)));

THA1=0;
THA1dB=0;
for i=1:(f_cut/f0)                    % Determine no. harmonics
for n=-bw:1:bw                        % Determine bw
 THA1 = THA1+spec_amp(i*f0/df+n).^2;
 spec_har(i*f0/df+n)=spec_amp(i*f0/df+n).^2;
 THA1dB = THA1dB+10*log10(spec_amp(i*f0/df+n).^2);
end
end

disp('lengths');
length(spec_amp)
length(spec_har)
length(spec_amp(1:f_cut/df+bw))

spec_nos=spec_amp(1:f_cut/df+bw).^2-spec_har;
spec_nos=spec_nos(spec_nos~=0);
spec_har=spec_har(spec_har~=0);
mHA=mean(spec_har);
mHAdB=mean(10*log10(spec_har));
mNA=mean(spec_nos);
mNAdB=mean(10*log10(spec_nos));

TSA=sum(spec_amp(1:f_cut/df).^2);
mSA=mean(spec_amp(1:f_cut/df).^2);
TSAdB=sum(10*log10(spec_amp(1:f_cut/df).^2));
mSAdB=mean(10*log10(spec_amp(1:f_cut/df).^2));
THA= THA1-sum(spec_amp(f0/df-bw:f0/df+bw).^2);
```

```
subplot(2,1,2);
[dbspec] = dBplot(spec_amp,pad);


% disp('Relative Harmonic Intensity =');
 disp('Hr=');
 Hr = THA/TSA*100 ;
% disp('Sr=');
 Sr = THA1/TSA*100;

 HN  = 10*log10(THA1/(TSA-THA1));
 H2N = 10*log10(THA/(TSA-THA));
 SN  = 10*log10(TSA/(TSA-THA1));

 for i=1:(fsam/2/f0)-1
 z(i) = max(spec_amp);
 end
 pause
 clf
% subplot(2,1,1);
 plot(300:600,10*log10(spec_amp(300:600)),'g*');
 hold on
 plot(300:600,10*log10(spec_amp(300:600)));
 hold off



  %%%%%%
  % S/N ratios
  %%%%%%

   HNgeo = mHAdB-mNAdB;


  % 1-3.8kHz
 har_14=zeros(size(spec_amp(1:f_cut/df)));

 THA14=0;
 THAdB14=0;
 for i=round(1000/f0):round(f_cut/f0)    % Determine no. harmonics
 for n=-bw:1:bw
% Determine bw
 THA14=THA14+spec_amp(i*f0/df+n).^2;
 THAdB14=THAdB14+10*log10(spec_amp(i*f0/df+n).^2);
 har_14((i-1000/f0+1)*f0/df+n)=spec_amp(i*f0/df+n).^2;
 end
 end
 THA14
 THAdB14
length(spec_amp(1000/df:f_cut/df));
length(har_14);

% nos_14=spec_amp(1000/df:f_cut/df).^2-har_14;
 %nos_14=nos_14(nos_14~=0);
```

```
har_14=har_14(har_14~=0);
mHA14=mean(har_14);
mHAdB14=mean(10*log10(har_14));
%mNA14=mean(nos_14);
%mNAdB14=mean(10*log10(nos_14));

TSA14=sum(spec_amp(round(1000/f0)*f0/df-bw:round(f_cut/f0)*f0/df+bw).^2)
TSAdB14=sum(10*log10(spec_amp(round(1000/df):round(f_cut/df)).^2));

HN_14 = 10*log10(abs(THA14)/abs(TSA14-THA14));
SN_14 = 10*log10(abs(TSA14)/abs(TSA14-THA14));
% HN14geo = mHAdB14-mNAdB14;

Sr_14=THA14/TSA14*100;


%%%%
% Display ratios
%%%%

%OUT=[Hr;Sr;HN;H2N;SN;HNgeo;Sr_14;HN_14;SN_14;HN14geo];
%disp(['Hr      '; 'Sr      '; 'HN      '; 'H2N     ';'SN      '; 'HNgeo   '; 'Sr_14   '; 'HN_14   ';'SN_14
';'HN14geo ']);
%sprintf('                    %f\n',OUT)
fprintf('Hr=%.3f\t',Hr);
fprintf('Sr=%.3f\t',Sr);
fprintf('HN=%.3f\t',HN);
fprintf('H2N=%.3f\t',H2N);
fprintf('SN=%.3f\t',SN);
%fprintf('HNgeo=%.3f\n',HNgeo);
%fprintf('Sr_14=%.3f\t',Sr_14);
%fprintf('HN_14=%.3f\t',HN_14);
%fprintf('SN_14=%.3f\t',SN_14);
%fprintf('HN14geo=%.3f\n',HN14geo);
```

A.2.3 Periodogram Averaged Analysis (PAHA)

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Program: harmper2.m
%
% Date:    03-03-'97
%
% Call:    [hsdB,Hr,Sr] = harmper2(sp_data);
%
% Note:    make len =2048 and
```

```
%           padded to 4096. Test db ratio.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

  function [pdgram,hsdB,Hr,Sr,dBH,dBS] = harmper2(sp_data);

 % Let the operator choose the region for analysis as usual.
 % variables:

  olap=100;
  f_cut=3800;
  fsam = 10000;
  pad = 4096;
  len = 2048;
  df=fsam/pad;
  bw=6*pad/len;

% Plot the speech sample
% Choose the region for analysis
% (0.2048 s)
% Plot the spectrum for this region
% (padded out to 4096)
  [spdata]=sonastar(sp_data,5000);
 [f0,nout] = fester2(spdata,olap);
f0

  [data] = sonasect(sp_data);   % Mouse click about 1 second in length
  [Pxx,f] = psd(data,pad,fsam,len,len/2);
  clf
  plot(f,Pxx);

 % Plot the spectrum from 70 to 400 Hz
 % Determine f0 from the usual cepstral analysis.

 speclow = Pxx((70/df):(400/df));
 subplot(3,1,2);
 plot(speclow);
```

```
% Find the harmonic frequencies and sum the energies at these frequency
% locations (exclude f0).
% Find the total signal energy.
% Hence determine the relative harmonic intensity (Hr).

 for i=1:f_cut/f0
  [peaks(i),locs(i)] = pkpicker( Pxx( i*round(f0/df)-bw:i*round(f0/df)+bw),1e-300,1);
   locs(i)=locs(i)+round(i*(f0/df)-bw)-1 ;
 end

% Total Harmonic Amplitude (THA1-incl. f0, THA-excl. f0)
% Total Signal Amplitude (TSA)

spec_har=zeros(size(Pxx(1:f_cut/df+bw)));

 THA1=0;
 THA1dB=0;
 for i=1:(f_cut/f0)                    % Determine no. harmonics
 for n=-bw:1:bw                         % Determine bw
 THA1 = THA1+Pxx(locs(i)+n);
 THA1dB = THA1dB+10*log10(Pxx(locs(i)+n));
 spec_har(locs(i)+n)=Pxx(locs(i)+n);
 end
 end

 for i=1:(f_cut/f0)
 har_min(i) = min(Pxx(locs(i)-bw:locs(i)+bw));      % spectrum derived Source spectrum
 har_max(i) = max(Pxx(locs(i)-bw:locs(i)+bw));
 hs(i) = har_max(i)/har_min(i);
 end
 length(hs);
 hsdB=10*log10(hs);
 HNS=10*log10(mean(hs));
 HNS01=10*log10(mean(hs(1:1000/f0)));
 HNS14=10*log10(mean(hs(1000/f0:f_cut/f0-1)));
 RHNS14=10*log10(mean(hs(1:1000/f0)/mean(hs(1000/f0:f_cut/f0-1))));  % subtract HNS01-HNS14
(same)
```

```
HNSh1_5=10*log10(mean(hs(1:5)));
HNSh6_11=10*log10(mean(hs(6:11)));
HNSh11=10*log10(mean(hs(1:11)));
RHNSh6_5=10*log10(mean(hs(1:5))/mean(hs(6:11)));
disp('lengths');
length(Pxx(1:f_cut/df+bw));
length(spec_har);

% spec_nos=Pxx(1:f_cut/df+bw)-spec_har;
% spec_nos=spec_nos(spec_nos~=0);
% spec_har=spec_har(spec_har~=0);

subplot(2,1,1)
title('source spectrum derived from output waveform spectrum');
plot(hsdB);
subplot(2,1,2)
('isolation of harmonics from output waveform spectrum')
plot(10*log10(spec_har));

% mHA=mean(spec_har);
% mHAdB=mean(10*log10(spec_har));
% mNA=mean(spec_nos);
% mNAdB=mean(10*log10(spec_nos));

TSA=sum(Pxx(1:f_cut/df+bw));
TSAdB=sum(10*log10(Pxx(1:f_cut/df+bw)));
THA= THA1-sum(Pxx(locs(1)-bw:locs(1)+bw));
THAdB=THA1dB-sum(10*log10(Pxx(locs(1)-bw:locs(1)+bw)));
subplot(2,1,2);
plot(f,Pxx);
title('Periodogram averaged spectrum');
xlabel('freq.(Hz)');
ylabel('amplitude (arb units)');

disp('Relative Harmonic Intensity =');
disp('Hr=');
Hr = THA/TSA*100;
```

```
disp('Sr=');
Sr = THA1/TSA*100;

HN  = 10*log10(THA1/abs(TSA-THA1));
H2N = 10*log10(THA/abs(TSA-THA));
SN  = 10*log10(TSA/abs(TSA-THA1));

for i=1:(fsam/2/f0)-1
z(i) = max(Pxx);
end

% subplot(2,1,1);
pdgram=10*log10(Pxx);
 plot(f(300:600),10*log10(Pxx(300:600)),'*');
% hold on
% plot(f(300:600),10*log10(Pxx(300:600)),'g');
 title('modified periodogram estimate');
 xlabel('freq.(Hz)');
 ylabel('ampl.(dB)');

 %%%%%%
 % S/N ratios
 %%%%%%

%  HNgeo = mHAdB-mNAdB;


 % 1-3.8kHz

har_14=zeros(size(Pxx(1000/df:f_cut/df+bw)));

THA14=0;
THAdB14=0;
for i=round(1000/f0):(f_cut/f0)                    % Determine no. harmonics
for n=-bw:1:bw                                  % Determine bw
THA14=THA14+Pxx(locs(i)+n);
THAdB14=THAdB14+10*log10(Pxx(locs(i)+n));
```

```
har_14((locs(i)-1000/f0+1)+n)=Pxx(locs(i)+n);
end
end
disp('lengths');
length(Pxx(1000/df:f_cut/df+bw));
length(har_14);

% nos_14=Pxx(1000/df:f_cut/df)-har_14;
% nos_14=nos_14(nos_14~=0);
% har_14=har_14(har_14~=0);
% mHA14=mean(har_14);
% mHAdB14=mean(10*log10(har_14));
% mNA14=mean(nos_14);
% mNAdB14=mean(10*log10(nos_14));

TSA14=sum(Pxx(round(1000/f0)*f0/df-bw:round(f_cut/f0)*f0/df+bw));
TSAdB14=sum(10*log10(Pxx(1000/df:f_cut/df)));

HN_14 = 10*log10(THA14/(TSA14-THA14));
SN_14 = 10*log10(TSA14/(TSA14-THA14));
% HN14geo = mHAdB14-mNAdB14;
Sr_14=THA14/TSA14*100;

%%%%
% Display ratios
%%%%

fprintf('Hr=%.3f\t',Hr);
fprintf('Sr=%.3f\t',Sr);
fprintf('HN=%.3f\t',HN);
fprintf('H2N=%.3f\t',H2N);
fprintf('SN=%.3f\n',SN);
%fprintf('HNgeo=%.3f\n',HNgeo);
fprintf('Sr_14=%.3f\t',Sr_14);
fprintf('HN_14=%.3f\t',HN_14);
fprintf('SN_14=%.3f\t',SN_14);
fprintf('HNS=%.3f\t',HNS);
```

```
fprintf('HNS14=%.3f\t',HNS14);
fprintf('HNS01=%.3f\n',HNS01);
fprintf('RHNS14=%.3f\t',RHNS14);
fprintf('HNSh1_5=%.3f\t',HNSh1_5);
fprintf('HNSh6_11=%.3f\t',HNSh6_11);
fprintf('HNSh11=%.3f\t',HNSh11);
fprintf('RHNSh6_5=%.3f\n',RHNSh6_5);
%fprintf('HN14geo=%.3f\n',HN14geo);
```

A.2.4 Pitch Synchronous (Four Periods)

```
%%%%%%%%%%%%%%%%%%
%
%
% Date:     Tues. 04-03-'97
%
% Program:  Harm4_2.m
%
% Call:   [LMIN,H,LOCS,dBH,NS] = harm4_2(sp);
%
% Descr.    Finds the noise to harmonic ratio for a signal by examining four
%           pitch periods. The ratio is in dBs. f0 is first calculated from
%           the cepstrum and then a better estimate is attained using zcs on
%           a lp filtered waveform.
%
%           Name:    Peter Murphy
%
%%%%%%%%%%%%%%%%%%



function [LMIN,H,LOCS,dBH,NS] = harm4_2(sp);


 fsam=10000;
 hop=100;
 PERC=0.4;
```

```
nh=16;

    %%%%%%%% Choose region for intended analysis %%%%%%%%%%

 [sect]=sonastar(sp,7500);

     %%%%%%%%   Obtain initial f0 estimate    %%%%%%%%%%

 [f0_est,nout] = fester2(sect,hop);
 f0_est

% 1=fs/2 i.e. 5000 Hz ... 0.1=500 Hz.

  b = fir1(250,[2*60/fsam,2*1.5*f0_est/fsam]);
  l = filtfilt(b,1,sect);
   length(l);
  j=1;
  for i=1:length(l)-1              % Obtain the number of
  if l(i)>0&l(i+1)<0               % negative going zero crossings.
                                % NZC.
   zc_in(j) = i ;
   j=j+1;
  end
  end

  NZC=j-1

  P(1)=0;

  for j=2:NZC-1;
  [y,in] =sort(l(zc_in(j-1):zc_in(j)));
  P(j)=zc_in(j-1)+in(1);
  end
                                % Poly interpolation
                                % PERC error detection

for j=5:NZC-1
```

```
 flnp(j-4)=fsam/(P(j)-P(j-1));
end

% Last check!!!
% Make sure that the f0s fall within an acceptable level.
% PERC

folnp=flnp( flnp<mean(flnp)+mean(flnp)*PERC&flnp>mean(flnp)-PERC*mean(flnp));
disp('no. of outliers');
Nout2 = length(flnp)-length(folnp)

P=P(4:NZC-1);              % Indices for the pitch periods
NPM=NZC-4;                 % in sect.


    %%%%%%%%%  Calculate the Power Spectrum    %%%%%%%%%%

  for i=1:NPM-4
  plen = P(i+4)-P(i);
  harms = abs(fft((sect(P(i):P(i+4)-1)).*hamming(plen) ) ).^2/plen;
  H(1:plen/2+1,i) = harms(1:plen/2+1)';
  df(i)=fsam/plen;
  end

%%%%%%%%%%%%%%%%%%%%  Calculate the N/H ratio for the first
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%                16                harmonics.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%

  for j=1:NPM-4
  for i=1:nh
   [peaks(i),locs(i)] = pkpicker( H( i*round(f0_est/df(j))-2:i*round(f0_est/df(j))+2,j),1e-300,1);

   LOCS(i,j)=locs(i)+round(i*(f0_est/df(j))-2)-1 ;
  end
  end
```

```
for j=1:NPM-4
for i=1:nh

  [pmin(i),lmin(i)] =  min( H(LOCS(i,j)-1:LOCS(i,j)+2) );
  LMIN(i,j)=lmin(i)+LOCS(i,j)-2;

end
end


N=sum(sum(H(LMIN)));

S=sum(sum(H));

NS =10*log10(N/S);

for j=1:size(H,2)
Nsseg(j) =10*log10( sum(H(LMIN(:,j)))/sum(H(:,j)) );
end
Nin=length(Nsseg);
Nsseg=Nsseg(
abs(Nsseg)<median(abs(Nsseg))+median(abs(Nsseg))*PERC&abs(Nsseg)>median(abs(abs(Nsseg)))-
PERC*median(abs(Nsseg)));
disp('no. of outliers');
Nout=Nin-length(Nsseg)

Nsseg;
NSseg=mean(Nsseg);
NSsegstd=std(Nsseg);
H=H+1e-100;                % Avoid log10(0)
dBH=10*log10(H);
%waterfall(dBH(3:66,:)')
%pause
plot(dBH(3:66,1),'r*')
hold on
plot(dBH(3:66,1))
pause
```

```
sum(dBH);
mean(dBH);
dBS=mean(mean(dBH));
dBN=mean(mean(dBH(LMIN)));
dBNS=dBN-dBS;

for j=1:size(H,2)
dBNsseg(j) =( mean(dBH(LMIN(:,j) ) )-mean(dBH(:,j)) ) ;
end
dBNSseg=mean(dBNsseg);
dBNSsstd=std(dBNsseg);
gSN=10*log10( sum(sum(dBH))/sum(sum(dBH(LMIN))) );
for j=1:size(H,2)
gSnseg(j) =10*log10( sum(dBH(:,j))/sum(dBH(LMIN(:,j))) );
end
gSnseg;
gSNseg=mean(gSnseg);
gSNsstd=std(gSnseg);


%OUT=[NS;NSseg;NSsegstd;dBNS;dBNSseg;dBNSsstd;gSN;gSNseg;gSNsstd];
%disp(['NS ';'NSseg ';'NSsegstd';'dBNS ';'dBNSseg ';'dBNSsstd';'gSN ';'gSNseg ';'gSNsstd
']);
%sprintf('                %f\n',OUT)
fprintf('NS=%.3f\t',NS);
fprintf('NSseg=%.3f\t',NSseg);
fprintf('NSsegstd=%.3f\t',NSsegstd);
fprintf('dBNS=%.3f\t',dBNS);
fprintf('dBNSseg=%.3f\t',dBNSseg);
fprintf('dBNSsstd=%.3f\n',dBNSsstd);
fprintf('gSN=%.3f\t',gSN);
fprintf('gSNseg=%.3f\t',gSNseg);
fprintf('gSNsstd=%.3f\t',gSNsstd);
```

A.2.5 Normalised Noise Energy

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Program: Noise6.m
%
% Date: wed. 22-01-'97
%
% Call: [NS] = noise6(sp);
%
% Name: Peter Murphy
%
% Aim:  To estimate the ratio of noise to total signal energy
%       for a speech signal. This is a revamp of noise2.m in
%       to evaluate the NNE at each segment, not just the
%       last.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


function [dB_Mmean,NS,M1,M2,M3,M4] = noise6(sp);


pad = 2048;
fsam = 10000;
olap = 200;
f_cut = 3800;
df = fsam/pad;
kilo = 1000;
hop = 100;
[sect2] = sonasect(sp);
[T0_est,nout] = fester1(sect2,hop);       % Make an initial estimate of f0
f0_est=1/T0_est;                          % using ckcorr.m
TI=7*T0_est*fsam;                         % From this est. we take 7 pitch
n0_segs=(length(sect2)-TI)/olap;          % periods-the analy.length.
BW=round(2*pad/TI);

%f0_est=109.89;
```

```
    for i=1:n0_segs

   seg= sect2(1+i*olap:TI+i*olap);
   fseg = (abs(fft(seg.*hamming(TI),pad)).^2)';
   M(:,i) = (fseg(1:f_cut/df+BW));
   dB_M(:,i) = 10*log10(fseg(1:f_cut/df+BW+5));          % dB_M contains the spectra in
                                                         % columns.
    end

   clf
%  waterfall(dB_M');
%  pause
%  clf
     dB_Mmean=mean(dB_M(100:200,:)');
    plot(mean(dB_M(100:200,:)'));
%    hold on
%    plot(mean(dB_M(200:300,:)'),'r*');
   title('dB spectrum of mean frequencies taken over the section');

%    [peaks,locs] = pkpicker(dBfseg,0.001,f_cut/f0_est);
   for i=1:14
   z(i)= max( dB_M(1:130,2) );
   end
%   clf
%   plot(dBfseg,'r');
%   hold on
%   stem(locs,z);
%   pause
%   hold off

   for j=1:n0_segs
    for n=1:f_cut/f0_est

  [peaks(n),locs(n)] = pkpicker( dB_M( (round(1+n*(f0_est/df)-BW ):round(1+n*(f0_est/df)+BW)
),j),-100,1);

    LOCS(n,j)=locs(n);
```

```
 end
end

for j=1:n0_segs
 for n=1:f_cut/f0_est
 LOCS(n,j) = round( 1+n*(f0_est/df)-BW-1+LOCS(n,j) );   % shift=index into array -1.
 end
end
                                        % Location of harmonic peaks
                                        % at j-th spectrum.
 %%%%%%%%%
 %
 % Plot one of the resulting spectra to
 % show that the peaks have been deter-
 % mined correctly.
 %
 %%%%%%%%%

%  clf
%  plot(dB_M((1:130),2),'r*');
%  hold on
%  plot( dB_M((1:130),2) );
%  hold on
%  stem(LOCS(1:14,2),z(1:14));
%  pause
%  hold off


 %%%%%%%
 %
 % Establish the harmonic peak regions. i.e. Take the BW eith
 %  -er side of n harmonic peaks.
 %
 %%%%%%%
```

```
for j=1:size(LOCS,2)

  for i=1:size(LOCS,1)

   P((i-1)*(2*BW+1)+1:i*(2*BW+1),j) = (LOCS(i,j)-BW:LOCS(i,j)+BW)';

  end
end

 %%%%%%%
 %
 % Establish the between harmonic regions. i.e. Take from
 % k+BW to k+1-BW.
 % This region is made up of the pad/2+1 points less the
 % P points (Also less the point up to the 1st harmonic)
 %
 %%%%%%%

 M1=M;
 M2=M;
 M2(P)=zeros(size(P));

 %%%%%%%%%%%
 %
 % Now estimate the noise energy in the
 % harmonic regions.
 %
 %%%%%%%%%%%

for j=2:size(LOCS,2)

 for i=2:size(LOCS,1)-1


  M1(LOCS(1,j)-BW:LOCS(1,j)+BW,j)=mean(M(LOCS(1,j)+BW+1:LOCS(2,j)-BW-
1)).*ones(size(LOCS(1,j)-BW:LOCS(1,j)+BW))';
```

```
    if( M(LOCS(i-1,j)+BW+1:LOCS(i,j)-BW-1) )==[]|(M(LOCS(i,j)+BW+1:LOCS(i+1,j)-BW-1))==[]

  M1(LOCS(i,j)-BW:LOCS(i,j)+BW,j)=mean(mean((M(LOCS(i-1,j)+BW+1:LOCS(i-
1,j)+BW+1)))+mean((M(LOCS(i,j)+BW+1:LOCS(i,j)+BW+1))).*ones(size(LOCS(i,j)-
BW:LOCS(i,j)+BW))';

     else
  M1(LOCS(i,j)-BW:LOCS(i,j)+BW,j)=mean(mean((M(LOCS(i-1,j)+BW+1:LOCS(i,j)-BW-
1)))+mean((M(LOCS(i,j)+BW+1:LOCS(i+1,j)-BW-1))).*ones(size(LOCS(i,j)-BW:LOCS(i,j)+BW))';
 end
end
end

%%%%
% Same for 1st column.
%%%%

 for i=2:size(LOCS,1)-1

  M1(LOCS(1,1)-BW:LOCS(1,1)+BW,1)=mean(M(LOCS(1,1)+BW+1:LOCS(2,1)-BW-
1)).*ones(size(LOCS(1,1)-BW:LOCS(1,1)+BW))';

    if(    M(LOCS(i-1,1)+BW+1:LOCS(i,1)-BW-1)    )==[]|(M(LOCS(i,1)+BW+1:LOCS(i+1,1)-BW-
1))==[]

  M1(LOCS(i,1)-BW:LOCS(i,1)+BW,1)=mean(mean((M(LOCS(i-1,2)+BW+1:LOCS(i-
1,2)+BW+1)))+mean((M(LOCS(i,2)+BW+1:LOCS(i,2)+BW+1))).*ones(size(LOCS(i,2)-
BW:LOCS(i,2)+BW))';

     else
  M1(LOCS(i,1)-BW:LOCS(i,1)+BW,1)=mean(mean((M(LOCS(i-1,1)+BW+1:LOCS(i,1)-BW-
1)))+mean((M(LOCS(i,1)+BW+1:LOCS(i+1,1)-BW-1))).*ones(size(LOCS(i,1)-
BW:LOCS(i,1)+BW))';
 end
end
```

```
%%%%%%
%
% If NaNs still exist remove that spectrum from
% both the signal & noise estimates.
%
%%%%%%

M3=M;
M4=M1;
for j=1:size(M1,2)
 x=M1(:,j);
 i=[];
 i=find(isnan(x));
 if i~=[];
 M1(:,j)=zeros(size(M1,1),1);
 M(:,j)=zeros(size(M,1),1);
 end
end

for j=1:size(M1,2)
 if M1(:,j)==0
 M3(:,j)=[];
 M4(:,j)=[];
end
end

%%%%%%%%%%%%%
%1.
% Sum the noise energy.
% Sum the signal energy.
% Calculate NNE.
%
%%%%%%%%%%%%%

 S = sum(sum(M3));

 W = sum(sum(M4));
```

```
NS=10*log10(W/S);

%%%%%%%
% 2. Limit 1-4kHz
%%%%%%%

   M1_4 = M3(kilo/df:f_cut/df,:);
   M11_4= M4(kilo/df:f_cut/df,:);

S14 = sum(sum(M1_4));

W14 = sum(sum(M11_4));

NS14=10*log10(W14/S14);

%%%%%%%
% Calculate 1.&2. above with dB ratios
%%%%%%%

dBM = 10*log10(M3);
dBM1= 10*log10(M4);

NSdB=mean(mean(dBM1-dBM));

dBM1_4 = 10*log10(M1_4);
dBM11_4= 10*log10(M11_4);

NS14dB=mean(mean(dBM11_4-dBM1_4));

%%%%%%
% Calculate segmental means
%%%%%%

Nsseg= 10*log10(sum(M4)./sum(M3));
NSseg=mean(Nsseg);
NSsegstd=std(Nsseg);
```

```
fsam = 10000;
olap = 100;
f_cut = 3800;
len = 7250;
PERC=0.4;
%%%%%%
%
% Make an initial estimate of f0
% This analysis is to be performed on 3 successive cycles
% for 325ms.
%
%%%%%%

[f0_est,nout] = fester2(sp,olap);
f0_est
df=fsam/(3*fsam/f0_est);

%%%%%%
%
% Choose region for analysis (325 ms).
%
%%%%%%

  [sect] = sonastar(sp,len);

   [in,pp,f,P,NPM] = zpitch3(sect,f0_est);
 %   [sect,f,pp,NPM] = zpitch2(sect,f0_est);
 %   M = zeros(max(diff(P))*3/2,NZC/3 );

 for j=2:3:NPM/3-3
  seg3pit = sect(pp(j):pp(j+3)-1);
  fseg3pit =  abs(fft(seg3pit)).^2';
 %   plot(fseg3pit(1:length(seg3pit)/2),'r');pause
 %   x = 1:length(seg3pit)/2; hold on
 %   stem(x,fseg3pit(1:length(seg3pit)/2) );
 %   pause
 %   hold off
```

```
M(1:length(fseg3pit)/2,(j+1)/3) = fseg3pit(1:length(seg3pit)/2)/length(seg3pit);
dB_M(1:length(fseg3pit)/2,(j+1)/3) = 10*log10( fseg3pit(1:length(seg3pit)/2)/length(seg3pit) );

                                % dB_M contains the spectra in
                                % columns.
end


waterfall(M(1:f_cut/df,:)');
pause
 clf
 plot(10*log10(mean(M(1:f_cut/df,:)')),'r*');
 hold on
 plot( 10*log10( mean( M(1:f_cut/df,:)')));
 hold off

 S  = [ M(1+3:3:f_cut/df ,:) ];
 N1 = [ M(1+1:3:f_cut/df-2 ,:) ];
 N2 = [ M(1+2:3:f_cut/df ,:) ];

 R = 10*log10( sum(sum(S))/( sum(sum(N1))+sum(sum(N2)) ) );

 S_dB  = [ dB_M(1+3:3:f_cut/df ,:) ];
 N1_dB = [ dB_M(1+1:3:f_cut/df-2,:) ];
 N2_dB = [ dB_M(1+2:3:f_cut/df,:) ];


 rseg=10*log10( mean(S)./(mean(N1)+mean(N2)) )

Nin=length(rseg);
rseg=rseg(        abs(rseg)<mean(abs(rseg))+mean(abs(rseg))*PERC&abs(rseg)>mean(abs(abs(rseg)))-
PERC*mean(abs(rseg)));
disp('no. of outliers');
Nout=Nin-length(rseg)

  Rseg=mean(rseg);
```

```
Rsegstd=std(rseg);

RdB =  10*log10( mean(mean(S_dB))-( mean(mean(N1_dB))+mean(mean(N2_dB)) ) );

rdbseg=10*log10( mean(S_dB)-(mean(N1_dB)+mean(N2_dB)) );

RdBseg=mean(rdbseg);

RdBsstd = std(rdbseg);

Rgeo= 10*log10( sum(sum(S_dB))/( sum(sum(N1_dB))+sum(sum(N2_dB)) ) );

rgeoseg= 10*log10( sum(S_dB)./(sum(N1_dB)+sum(N2_dB)) );

Rgeoseg= mean(rgeoseg);

Rgeosstd=std(rgeoseg);

%OUT=[R;Rseg;Rsegstd;RdB;RdBseg;RdBsstd;Rgeo;Rgeoseg;Rgeosstd];
%disp(['R       ';'Rseg    ';'Rsegstd ';'RdB     ';'RdBseg  ';'RdBsstd ';'Rgeo     ';'Rgeoseg ';'Rgeosstd']);
sprintf('                %f\n',OUT)
fprintf('R=%.3f\t',R);
fprintf('Rseg=%.3f\t',Rseg);
fprintf('Rsegstd=%.3f\t',Rsegstd);
fprintf('RdB=%.3f\t',RdB);
fprintf('RdBseg=%.3f\t',RdBseg);
fprintf('RdBsstd=%.3f\n',RdBsstd);
fprintf('Rgeo=%.3f\t',Rgeo);
fprintf('Rgeoseg=%.3f\t',Rgeoseg);
fprintf('Rgeosstd=%.3f\n',Rgeosstd);

  S_14  = [ S(1000/f0_est:size(S,1),:) ];
  N1_14 = [ N1(1000/f0_est:size(N1,1),:) ];
  N2_14 = [ N2(1000/f0_est:size(N2,1),:) ];

  R_14 = 10*log10( sum(sum(S_14))/( sum(sum(N1_14))+sum(sum(N2_14)) ) );
```

```
  S_dB14  = [ S_dB(1000/f0_est:size(S,1),:) ];
  N1_dB14 = [ N1_dB(1000/f0_est:size(N1,1),:) ];
  N2_dB14 = [ N2_dB(1000/f0_est:size(N2,1),:) ];

  rseg14=10*log10( mean(S_14)./(mean(N1_14)+mean(N2_14)) );

Nin=length(rseg14);
rseg14=rseg14(
abs(rseg14)<median(abs(rseg14))+median(abs(rseg14))*PERC&abs(rseg14)>median(abs(abs(rseg14))
)-PERC*median(abs(rseg14)));
%disp('no. of outliers');
%Nout=Nin-length(rseg14)

  Rseg14=mean(rseg14);

  Rseg14d=std(rseg14);


  RdB14 =  10*log10( mean(mean(S_dB14))-( mean(mean(N1_dB14))+mean(mean(N2_dB14)) ) );

  rdbseg14=10*log10( mean(S_dB14)-(mean(N1_dB14)+mean(N2_dB14)) );

  RdBseg14=mean(rdbseg14);

  RdBs14d = std(rdbseg14);

  Rgeo14=10*log10( sum(sum(S_dB14))/( sum(sum(N1_dB14))+sum(sum(N2_dB14)) ) );

  rgeos14= 10*log10( sum(S_dB14)./(sum(N1_dB14)+sum(N2_dB14)) );

  Rgeos14= mean(rgeos14);

  Rgeos14d=std(rgeos14);

%OUT=[R_14;Rseg14;Rseg14d;RdB14;RdBseg14;RdBs14d;Rgeo14;Rgeos14;Rgeos14d];
%disp(['R_14       ';'Rseg14   ';'Rseg14d  ';'RdB14     ';'RdBseg14';'RdBs14d  ';'Rgeo14    ';'Rgeos14
';'Rgeos14d']); sprintf('                %f\n',OUT)
```

```
fprintf('R_14=%.3f\t',R_14);
fprintf('Rseg14=%.3f\t',Rseg14);
fprintf('Rseg14d=%.3f\t',Rseg14d);
fprintf('RdB14=%.3f\t',RdB14);
fprintf('RdBseg14=%.3f\t',RdBseg14);
fprintf('RdBs14d=%.3f\n',RdBs14d);
fprintf('Rgeo14=%.3f\t',Rgeo14);
fprintf('Rgeos14=%.3f\t',Rgeos14);
fprintf('Rgeos14d=%.3f\t',Rgeos14d);
```

A.2.7 Partial Sum of the Fourier Series (Two Period)

```
%%%%%%%%%%%%%%%%%%%%
%
% Program:  psha2.m
%
% Date:     Mon. 24-06-'97
%
% Call:     psha2(sp);
%
% Name:     Peter Murphy
%
% Aim:      To estimate the harmonic to noise ratio
%           using Fourier Series Expansion. H/N cycle by cycle.
%           psha2 = pitch sync. harm. analysis (1=lths3)
%
%%%%%%%%%%%%%%%%%%%%


  function [S,N,P] = psha2(sp);

fsam = 10000;
olap = 100;
f_cut = 3800;
len = 7250;
```

```
PERC=0.4;
%%%%%%
%
% Make an initial estimate of f0
% This analysis is to be performed on 2 successive cycles
% for 725ms.
%
%%%%%%

[f0_est,nout] = fester2(sp,olap);
f0_est
df=fsam/(2*fsam/f0_est);

%%%%%%
%
% Choose region for analysis (725 ms).
%
%%%%%%

  [sect] = sonastar(sp,len);

   [in,pp,f,P,NPM] = zpitch3(sect,f0_est);
%   [sect,f,pp,NPM] = zpitch2(sect,f0_est);
%   M = zeros(max(diff(P))*3/2,NZC/3 );

 for j=2:1:NPM-3
  seg2pit = sect(pp(j):pp(j+2)-1);
  fseg2pit =  abs(fft(seg2pit)).^2';
%   plot(fseg3pit(1:length(seg3pit)/2),'r');pause
%   x = 1:length(seg3pit)/2; hold on
%   stem(x,fseg3pit(1:length(seg3pit)/2) );
%   pause
%   hold off


  M(1:length(fseg2pit)/2,j-1) = fseg2pit(1:length(seg2pit)/2)/length(seg2pit);
  dB_M(1:length(fseg2pit)/2,j-1) = 10*log10( fseg2pit(1:length(seg2pit)/2)/length(seg2pit) );
```

```
                                % dB_M contains the spectra in
                                % columns.
 end


%  waterfall(M(1:f_cut/df,:)');
%  pause
%   clf
%   plot(10*log10(mean(M(1:f_cut/df,:)')),'r*');
%   hold on
   stem( 10*log10( mean( M(1:f_cut/df,:)')));
%   hold off
pause
   S  = [ M(1+2:2:f_cut/df ,:) ];
   N = [ M(1+1:2:f_cut/df-2 ,:) ];
size(S)
size(N)
if (size(S)==size(N))
   hns= S(1:size(S,1)-1,:)./N(1:size(N,1)-1,:);
else
   hns= S(1:size(S,1)-2,:)./N(1:size(N,1)-1,:);
end
 HNS = 10*log10(mean(mean(hns)));
%   waterfall(10*log10(hns)');

%pause
%   plot(10*log10(hns(1:size(hns,1),1)));
%pause
%  plot(10*log10(hns(1:size(hns,1),2)));
%pause
% plot(10*log10(hns(1:size(hns,1),3)));
%pause
%  plot(10*log10(hns(1:size(hns,1),4)));
%pause
   plot(10*log10(mean(hns(1:size(hns,1),:)')));
   hns_av = mean((hns(1:size(hns,1),:))');
```

```
HNS01=10*log10(mean(hns_av(1:1000/f0_est)));
HNS14=10*log10(mean(hns_av(1000/f0_est:f_cut/f0_est-2)));
RHNS14=10*log10(mean(hns_av(1:1000/f0_est)/mean(hns_av(1000/f0_est:f_cut/f0_est-2))));      %
subtract HNS01-HNS14 (same)
HNSh1_5=10*log10(mean(hns_av(1:5)));
HNSh6_11=10*log10(mean(hns_av(6:11)));
HNSh11=10*log10(mean(hns_av(1:11)));
RHNSh6_5=10*log10(mean(hns_av(1:5))/mean(hns_av(6:11)));


  R = 10*log10( sum(sum(S))/sum(sum(N)) );

  S_dB  = [ dB_M(1+2:2:f_cut/df ,:) ];
  N_dB = [ dB_M(1+1:2:f_cut/df-2,:) ];

  rseg=10*log10( mean(S)./mean(N) )

Nin=length(rseg);
rseg=rseg(        abs(rseg)<mean(abs(rseg))+mean(abs(rseg))*PERC&abs(rseg)>mean(abs(abs(rseg)))-
PERC*mean(abs(rseg)));
disp('no. of outliers');
Nout=Nin-length(rseg)

  Rseg=mean(rseg);

  Rsegstd=std(rseg);

  RdB =  10*log10( mean(mean(S_dB))-( mean(mean(N_dB)) ) );

  rdbseg=10*log10( mean(S_dB)-(mean(N_dB)) );

  RdBseg=mean(rdbseg);

  RdBsstd = std(rdbseg);

  Rgeo= 10*log10( sum(sum(S_dB))/( sum(sum(N_dB)) ) );
```

```
  rgeoseg= 10*log10( sum(S_dB)./(sum(N_dB)) );
  Rgeoseg= mean(rgeoseg);
  Rgeosstd=std(rgeoseg);

%OUT=[R;Rseg;Rsegstd;RdB;RdBseg;RdBsstd;Rgeo;Rgeoseg;Rgeosstd];
%disp(['R       ';'Rseg    ';'Rsegstd ';'RdB      ';'RdBseg  ';'RdBsstd ';'Rgeo     ';'Rgeoseg ';'Rgeosstd']);
sprintf('                %f\n',OUT)
fprintf('R=%.3f\t',R);
fprintf('Rseg=%.3f\t',Rseg);
fprintf('Rsegstd=%.3f\t',Rsegstd);
fprintf('RdB=%.3f\t',RdB);
fprintf('RdBseg=%.3f\n',RdBseg);
fprintf('RdBsstd=%.3f\t',RdBsstd);
fprintf('Rgeo=%.3f\t',Rgeo);
fprintf('Rgeoseg=%.3f\t',Rgeoseg);
fprintf('Rgeosstd=%.3f\n',Rgeosstd);

  S_14  = [ S(1000/f0_est:size(S,1),:) ];
  N_14 = [ N(1000/f0_est:size(N,1),:) ];
  R_14 = 10*log10( sum(sum(S_14))/sum(sum(N_14))  );
  S_dB14  = [ S_dB(1000/f0_est:size(S,1),:) ];
  N_dB14 = [ N_dB(1000/f0_est:size(N,1),:) ];

  rseg14=10*log10( mean(S_14)./mean(N_14)  );

Nin=length(rseg14);
rseg14=rseg14(
abs(rseg14)<median(abs(rseg14))+median(abs(rseg14))*PERC&abs(rseg14)>median(abs(abs(rseg14))
)-PERC*median(abs(rseg14)));
%disp('no. of outliers');
%Nout=Nin-length(rseg14)

  Rseg14=mean(rseg14);
  Rseg14d=std(rseg14);
  RdB14 =  10*log10( mean(mean(S_dB14))-( mean(mean(N_dB14)) ) );
  rdbseg14=10*log10( mean(S_dB14)-(mean(N_dB14)) );
```

```
RdBseg14=mean(rdbseg14);
RdBs14d = std(rdbseg14);
Rgeo14=10*log10( sum(sum(S_dB14))/( sum(sum(N_dB14)) ) );
rgeos14= 10*log10( sum(S_dB14)./(sum(N_dB14)) );
Rgeos14= mean(rgeos14);
Rgeos14d=std(rgeos14);
```

A.2.8 Time Domain Averaging

```
%%%%%%%%%%%%%%%%%%%%%%
% program: harmyum1.m
%
% date:   Tues. 03-12-'96
%
% call:   [M] = harmyum1(sp);
%
% Descr.  finds the harmonic to noise ratio from the time domain signal.
%%%%%%%%%%%%%%%%%%%%%%

function [M,P] = harmyum1(sp,np);

f_cut=3800;
fsam=10000;
hop=200;
len=10500;
olap=100;

    %%%%%%%% Mouse pick the section again %%%%%%%%%%

[sect]=sonastar(sp,len);

   %%%%%%%%  Calculate an initial estimate of %%%%%%%%%%
   %%%%%%%%  f0 using autocorrelation.       %%%%%%%%%%
```

```
 [f0_est,nout] = fester2(sect,olap);
%  [sect1,f,pp,NPM] = zpitch2(sect,f0_est);
 [in,pp,f,P,NPM] = zpitch3(sect,f0_est);

   %%%%%%%%%%%%%    Attain an average waveform     %%%%%%%%%%%%
    for j=5:NPM-5
    M( 1:(pp(j+1)-pp(j)),j-4) = sect( pp(j):pp(j+1)-1 )';
    end
    M=M(:,1:size(M,2)-2);
    snratio(M,np);
```

A.2.9 Pitch Synchronous Harmonic Analysis

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Program:   LTHS4 - long time harmonic spectrum
%
% Date:      Thurs. 18-03-'97
%
% By:        Peter Murphy
%
% call:      [pp]=lths4(sp);
%
% Note:      This program provides a convenient way to analyse
%          how the ampl. of the 'harmonics' change with time.
%          The changes to lths1.m are 1. 1-3.8k range used.
%                                     2. global ratios are calc.
%                                        every 50 periods & std taken.
%           As per lths3 but padded to 512 pts.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 function [pp]=lths4(sp);

 fsam=10000;
 f_cut=3800;
 len=7500;
 hop=200;
 np=50;
```

```
sft=10;
pad=512;
%%%%%
%
% Long time harmonic spectra -estimate the gross spectral features for a sample
% of speech. Note the addition is done wrt the harmonic freqs.
% Make an initial estimate of f0
%
%%%%%%

[f0_est,nout] = fester2(sp,100);

%f0_est=220;

%%%%%%
% Choose region for analysis.
%%%%%%

 [sect] = sonastar(sp,len);
 [in,pp,f,P,NPM] = zpitch3(sect,f0_est);
% [sect,f,pp,NPM] = zpitch2(sect,f0_est);

 % M = zeros(min(diff(P))/2+1,NZC-1 );

for j=1:NPM-5

  segpit = sect(pp(j+2):pp(j+3)-1);
  fsegpit =  (abs(fft(segpit,pad))/pad)';   % Normalise wrt win. length
  fsegpit = (fsegpit(1:pad/2+1)*f_cut/fsam*2)/max(fsegpit);
 % stem(fsegpit); pause
 % hold on
  M(1:length(fsegpit),j) = fsegpit;
  dB_M(1:length(fsegpit),j) = 20*log10(fsegpit);

                                % dB_M contains the spectra in
                                % columns.
end
```

```
%clf
%fallspec(M');
%pause


%clf
avhspec = mean(M');
%plot(mean(M'),'r');
%hold on
%stem(mean(M'));
%hold off
%pause
%clf
stem(mean(dB_M'));
plot(mean(dB_M'),'r');
pause

%%%%%%%%%
% Amplitude Perturbation Ratios
%%%%%%%%%

APAP = mean(abs(diff(M(2,:))))/mean(M(2,:));   % Aver perc. pitch ampl. perturbation.
stdAPAP=std(abs(diff(M(2,:))))/mean(M(2,:));
SHR  = mean(abs(diff(dB_M(2,:))));            % Shimmer
stdSHR=std(abs(diff(dB_M(2,:))));

HAPAP = mean(abs(diff(M')))./mean(M');        % As per APAP but for all harmonics.
SHH  = mean(abs(diff(dB_M')));               % "   " SH  "  "  "  "      .

THAPAP = mean(HAPAP);                        % Aver over all freqs.
sTHAPAP= std(HAPAP);
TSHH  = mean(SHH);                          % "   "   "   "  .
sTSHH  = std(SHH);

%%%%%%%%%
% Signal to Noise Ratios
```

```
%%%%%%%%%

    for i=1:(size(M,2)-np)/sft
     Ms=M(:,1+(i-1)*sft:np+(i-1)*sft);
     As = mean(Ms');
     AVs=As';
     H = sum(sum(Ms.^2));
     plot(AVs');

     Ms_AV=[];
     for j=1:size(Ms,2)
     Ms_AV = [Ms_AV AVs];
     end

     N = sum( sum( (Ms-Ms_AV).^2) );
     Hn(i)=10*log10(H/N);
    end

     HN=mean(Hn);
     stdHN=std(Hn);                    %% HN

     A = mean(M');
     AV=A';

     Hnseg=0;
     for i=1:size(M,2)
     Hnseg(i)=10*log10(sum(M(:,i).^2)/sum((M(:,i)-AV).^2));
     end                               %% HNseg
     stdHNseg=std(Hnseg);
     HNseg=mean(Hnseg);

    for i=1:(size(dB_M,2)-np)/sft
     dB_Ms=dB_M(:,1+(i-1)*sft:np+(i-1)*sft);
     dBAs = mean(dB_Ms');              %% HNgeo
     dBAVs=dBAs';
     dBM_AVs=[];
     for j=1:size(dB_Ms,2)
```

```
dBM_AVs = [dBM_AVs dBAVs];
end


Hngeo(i) = 10*log10((mean(mean((abs(dB_Ms-dBM_AVs))))));

dBHs = sum(sum(dB_Ms.^2));
dBNs = sum( sum( (dB_Ms-dBM_AVs).^2) );

dBHn(i)=10*log10(dBHs/dBNs);          %% dBHN
end

HNgeo=mean(Hngeo);
stdHNgeo=std(Hngeo);
dBHN=mean(dBHn);
stddBHN=std(dBHn);

%%%%%%%
% Distortion Factor
%%%%%%%

for i=1:size(M,2)
d(i)=sum(M(3:size(M,1),i).^2)/M(2,i).^2;
end

DF1=mean(d);                          %% DF1
stdDF1=std(d);

for i=1:size(dB_M,2)
db(i)=sum(abs(dB_M(3:size(dB_M,1),i)))/abs(dB_M(2,i));
end

DF2=mean(db);                         %% DF2
stdDF2=std(db);

%%%%%%%
% Limit the spectral range & calculate the same ratios
```

```
% from 1k-f_cut Hz.
%%%%%%%

 size(M)
  f0=fsam/min(diff(pp));
   x=size(M,1);
  M14=M(1000/f0:x,:);
  dB_M14=20*log10(M14);

HAPAP14 = mean(abs(diff(M14'))./mean(M14');     % As per APAP but for all harmonics.
SHH14   = mean(abs(diff(dB_M14')));             %"   " SH  "  "  "  "

THAPAP14 = mean(HAPAP);                         % Aver over all freqs.
sTHA14= std(HAPAP);
TSHH14   = mean(SHH);                       % "   "   "   " .
sTSHH14  = std(SHH);

%%%%%%%%%
% Signal to Noise Ratios
%%%%%%%%%

      for i=1:(size(M14,2)-np)/sft
       Ms14=M14(:,1+i*sft:np+i*sft);
       As14 = mean(Ms14');
       AVs14=As14';
       H14 = sum(sum(Ms14.^2));
       plot(AVs14');
       Ms_AV14=[];
       for j=1:size(Ms14,2)
       Ms_AV14 = [Ms_AV14 AVs14];
       end

       N14 = sum( sum( (Ms14-Ms_AV14).^2) );

       Hn14(i)=10*log10(H14/N14);
      end
       HN14=mean(Hn14);
```

```
stdHN14=std(Hn14);                    %% HN14

A14 = mean(M14');
AV14=A14';

Hnseg14=0;
for i=1:size(M14,2)
Hnseg14(i)=10*log10(sum(M14(:,i).^2)/sum((M14(:,i)-AV14).^2));
end                                   %% HNseg14
HNseg14=mean(Hnseg14);
stdHNs14=std(Hnseg14);

for i=1:(size(dB_M14,2)-np)/sft
 dB_Ms14=dB_M14(:,1+i*sft:np+i*sft);
 dBAs14 = mean(dB_Ms14');             %% HNgeo
 dBAVs14=dBAs14';
 dBMAVs14=[];
 for j=1:size(dB_Ms14,2)
 dBMAVs14 = [dBMAVs14 dBAVs14];
end

 Hngeo14(i) = 10*log10((mean(mean(abs((dB_Ms14-dBMAVs14))))));


 dBHs14 = sum(sum(dB_Ms14.^2));
 dBNs14 = sum( sum( (dB_Ms14-dBMAVs14).^2) );

 dBHn14(i)=10*log10(dBHs14/dBNs14);          %% dBHN
end

HNgeo14=mean(Hngeo14);
stdHNg14=std(Hngeo14);
dBHN14=mean(dBHn14);
stdBHN14=std(dBHn14);


%%%%%%%%%
```

```
% Display the above:
%%%%%%%%%

fprintf('APAP=%.3f\t',APAP);
fprintf('SHR=%.3f\t',SHR);
fprintf('THAPAP=%.3f\t',THAPAP);
fprintf('TSHH=%.3f\t',TSHH);
fprintf('HN=%.3f\t',HN);
fprintf('HNseg=%.3f\n',HNseg);
fprintf('HNgeo=%.3f\t',HNgeo);
fprintf('dBHN=%.3f\t',dBHN);
fprintf('DF1=%.3f\t',DF1);
fprintf('DF2=%.3f\n',DF2);

fprintf('stdAPAP=%.3f\t',stdAPAP);
fprintf('stdSHR=%.3f\t',stdSHR);
fprintf('sTHAPAP=%.3f\t',sTHAPAP);
fprintf('sTSHH=%.3f\t',sTSHH);
fprintf('stdHN=%.3f\t',stdHN);
fprintf('stdHNseg=%.3f\n',stdHNseg);
fprintf('stdHNgeo=%.3f\t',stdHNgeo);
fprintf('stddBHN=%.3f\t',stddBHN);
fprintf('stdDF1=%.3f\t',stdDF1);
fprintf('stdDF2=%.3f\n',stdDF2);

fprintf('THAPAP14=%.3f\t',THAPAP14);
fprintf('sTHA14=%.3f\t',sTHA14);
fprintf('TSHH14=%.3f\t',TSHH14);
fprintf('sTSHH14=%.3f\t',sTSHH14);
fprintf('HN14=%.3f\t',HN14);
fprintf('stdHN14=%.3f\n',stdHN14);
fprintf('HNseg14=%.3f\t',HNseg14);
fprintf('stdHNs14=%.3f\t',stdHNs14);
fprintf('HNgeo14=%.3f\t',HNgeo14);
fprintf('stdHNg14=%.3f\n',stdHNg14);
fprintf('dBHN14=%.3f\t',dBHN14);
fprintf('stdBHN14=%.3f\n',stdBHN14);
```

A.3 Long Term Average Spectrum Analysis

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Program:   LTAS3 - long term average spectra
%
% Date:      Tues. 04-02-'97
%
% By:        Peter Murphy
%
% call:      [averspec] = ltas3(sp_data,lseg,hop,pad);
%
% Note:      This is a copy of ltas1 altered to obtain the
%            difference between spectra.
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

 function [longspec,a,b]=ltas3(sp_data,lseg,hop,pad);
 f_cut = 3800;

% Long term average spectra-estimates the gross spectral features for a sample
% of speech.

   [sect]=mousie(sp_data);
     nsegs = (length(sect)-lseg)/hop;
    clf
for i=1:(length(sect)-lseg)/hop
 specamp = specampl(sect(1+(i-1)*hop:lseg+(i-1)*hop),pad);
 M(1:pad/2+1,i) = specamp'.^2;
 dB_M(1:pad/2+1,i) = 10*log10(specamp'.^2);
end

waterfall(M');
title('Linear spectra');
xlabel('freq');
ylabel('ampl');
```

```
pause
waterfall(dB_M');
pause
DM = diff(M');
DdB_M = diff(dB_M');
waterfall(DM);
pause
waterfall(DdB_M);
pause

avspec = mean(M');
plotspec(avspec,pad);
pause
plotdB(mean(dB_M'),pad);
title('dB spectrum of LTAS');
a=diff(avspec,2);
pause
plot(diff(diff((10*log10(avspec)))));
b=diff(diff((10*log10(avspec))));
pause
plot(10*log10(avspec));
pause
plot(mean(DM));
pause
```

A.4 Cepstral Analysis

```
%%%%%%%%%%
%
% Program: cpphnr.m To obtain the hnr for a speech sample using cepstral
%        analysis. A comb liftered cepstrum is ffted to give a noise
%        estimate.
%
% Name:   Peter Murphy
%
% Date:   Thurs. 27/02/'97
%
% call:   [RP,CRAPP,CRAPP1,CN,dBCN] = cpphnr(sp,hop);
%        lseg = 2048; hop = 256.
%%%%%%%%%%

function [RP,CRAPP,CRAPP1,CN,dBCN] = cpphnr(sp,hop);
lseg=2048;
bw = 10;
cw = 10;
[sect] = sonastar(sp,5000);

  for i = 1:(length(sect)-lseg)/hop

[logh, c_sect] = cceps3(sect(1+i*hop:lseg+i*hop));
                                % Returns real cepstrum.
 C(:,i) = c_sect';
 O(:,i) = logh';

  end


% plot(((Hf)));                 % Graphing outputs.
% title('cepstral peaks vs no. of hops');     % Cepstral peaks & Cepstrum.
% xlabel('no. hops');
```

```
% ylabel('amplitude (linear)');
% pause;

% plot(20*log10(c_sect));          % pause;
% dBc_sect = 20*log10(c_sect);
% title('cepstrum ampl. dB vs quefrency');
% xlabel('quefrency (ms)');
% ylabel('amplitude (dB)');
% pause;
% clf

                          % 10k/25:10k/142 freq range.
[Y,I]=max(C((25:142),:));          % Obtain the 1st rahmonic index
                          % for each cepstrum.
  I=I+24;

I=29;


for i=1:size(C,2)
 for j=1:size(C,1)/max(I)/2

 [a,b]=max(C(j*I(i)-bw:j*I(i)+bw,i));

  ii(j)=b+j*I(i)-bw-1;

 end
 RP(:,i) = ii';                  % Rahmonic peaks RP
end                              % Each column contains the rah.
                          % peaks for that cepstrum.

for i=1:size(RP,2)               % Mean Rahmonics amplitude (dB)
CRAPP1=C(RP(:,i),i);
 end
CRAPP1 = mean(sum(CRAPP1);

C1=C;
```

```
for j=1:size(C,2)
 for i=1:size(RP,1)                    % Comb liftering
   C1(RP(i)-0.5*cw:RP(i)+0.5*cw,j)=zeros(size(RP(i)-0.5*cw:RP(i)+0.5*cw))';
    C1((1:40),j)=zeros(size(1:40))';
    C1((2048-40:2048),j)=zeros(size(1:41))';
 end
end

CN=sum(sum(C1));
dBCN=10*log10(CN);
fprintf('CRAPP = %6.3f\t', CRAPP);
fprintf('CRAPP1 = %6.3f\t', CRAPP1);
fprintf('CN    = %6.3f\t',CN);
fprintf('dBCN  = %6.3f\n',dBCN);
```