Vocal Qualities in Female Singing

Michelle Evans

A thesis submitted for the degree of
Doctor of Philosophy
Department of Electronics
University of York

September 1995

Abstract

Different techniques are required to produce different vocal qualities in singing. The demands placed on modern singers as they seek to take advantage of the full range of vocal influences available potentially creates vocal strain. Good vocal health stemming from the appropriate vocal technique is an essential requirement for longevity of the voice regardless of singing style. The objective scientific analysis of professional singers' vocal qualities and vocal techniques is a good starting point for the understanding of efficient and healthy modes of singing production.

It is hypothesised that standard two-channel speech analysis and speech synthesis techniques are appropriate for modelling two perceptually very different qualities in the female singing voice; "classical" opera quality, and the non-classical "belting" quality, which is heard in much popular and ethnic music today.

This thesis details the results of an experiment comparing the vocal qualities of female opera singers with West End musical singers who are trained to "belt".

Standard two-channel speech analysis normally involves recording in stereo, vocal fold vibration by means of a single-channel electrolaryngograph on one channel, and the acoustic output from a microphone on the other. In this study, a multi-channel recording set-up has been used. This is to employ larynx height measurement, a relatively new addition to voice analysis techniques, which requires a two-channel electrolaryngograph. It has shown that the additional use of the larynx height measurement is an improvement over the standard two-channel speech technique, since it appears to be an important parameter in singing production.

The important quantifiable parameters which relate to the perceptual differences between the two qualities has been shown to be the closed quotient measure, the spectral envelope, and vibrato. These have been used as input parameters to drive a speech synthesizer in order to resynthesize the singing qualities. A perceptual test has shown that the robustness of the models derived from the results are adequate.

In conclusion, it has been shown that it is possible to use standard two-channel speech analysis techniques and speech synthesis techniques for perceptually differentiating between female opera and belting qualities. However, in terms of understanding singing production, these techniques benefit considerably from additional analysis equipment more suited to the extra features of singing production, and a synthesizer which is specifically designed for singing synthesis.

Contents

1 Research Objectives and Report Structure 1

1.1 Introduction 1

1.2 Hypothesis 1

1.3 Thesis Content and Structure 2

2 Voice and Hearing Systems 3

2.1 Introduction 3

2.2 The Human Vocal System 3

2.2.1 The Subglottal System 3

2.2.1.1 The Lungs 5

2.2.1.2 The Subglottal System in Respiration 5

2.2.1.3 Respiration in Phonation 6

2.2.1.4 Lung Volumes in Phonation 7

2.2.1.5 Subglottal Pressure in Phonation 8

2.2.1.6 Subglottal Pressure and Airflow 9

2.2.1.7 Muscular Combinations in Phonation 11

2.2.2 The Larynx 11

2.2.2.1 The Components of the Larynx 13

2.2.2.2 Vocal Fold Vibration 18

2.2.3 The Supraglottal System 19

2.2.4 The Acoustics of the Vocal Tract 20

2.2.4.1 Speech Production 20

2.2.4.2 Vowels 20

2.2.4.3 The Source-Filter Theory 21

2.2.4.4 The Tube Resonator Model 22

2.3 Human Hearing System 24

2.3.1 Hearing Physiology 24

2.3.1.1 The Outer Ear 25

2.3.1.2 The Middle Ear 25

2.3.1.3 The Inner Ear 27

2.3.1.4 Basilar Membrane Movement 27

2.3.2 Hearing Perception and Psychoacoustics 30
2 2.3. Critical Bands 31
2.3.4 Pitch 31
2.3.5 Loudness 34
2.3.5.1 The Threshold of Audibility 34
2.3.5.2 Loudness Levels and Equal Loudness Contours 35
2.3.5.3 The Musical Dynamic Scale 36
2.3.5.4 Loudness Masking of Complex Tones 37
2.3.6 Timbre 38
2.3.6.1 Steady State Spectral Components 38
2.3.6.2 Transients 39
2.3.6.3 The Amplitude Envelope 40
2.3.6.4 Timbre Changes with Frequency 41
2.3.6.5 Timbre Changes with Intensity 41
2.3.6.6 Timbre and Singing 42

3 Standard Speech Analysis Techniques 43
3.1 Introduction 43
3.2 Voice Source Parameter Analysis 43
3.2.1 Electrolaryngography 43
3. 3 Acoustic Signal Parameter Analysis 46
3.3.1 Sound Pressure Level Recording 46
3.3.2 Fast Fourier Transform Analysis 46
3.3.3 Spectrography 47
3.3.4 Inverse Filtering 49
3.3.5 Linear Predictive Coding 49

4 Vocal Qualities 52
4.1 Introduction 52
4.1.1 Voice Registers 52
53
4.2 Vocal Fold Vibration and Acoustic Quality
4.2.1 Modal Voice 53
4.2.2 Falsetto 55
4.2.3 Creak 56
4.2.4 Breathiness 57

4.2.5 Whisper 58
4.2.6 Harshness 58

4.3 Male and Female Voice Differences 60
4.3.1 Voice Source Differences 60
4.3.2 Supralaryngeal Differences 64

4.4 Opera Quality 64
4.4.1 Male Opera Quality 65
4.4.2 Female Opera Quality 67

4.5 Belting 70

4.6 Belting and Opera Quality Comparisons 71
4.6.1 Differences Between Belting and Opera 71
4.6.2 Similarities Between Belting and Opera 73
4.6.3 Larynx Height Differences 74

4.7 Conclusions 77

5 Experimental Procedure 78
5.1 Introduction 78

5.2 Recording Location 78

5.3 Subjects 78

5.4 Equipment and Experimental Method 79
5.4.1 The Two-Channel Electroglottograph 79

5.5 Recording Procedure 81

5.6 Digital Recording 81

5.7 The Speech Filing System (SFS) 81

5.8 Method for Extracting Larynx Height Data 82

6 Results 83
6.1 Introduction 83

6.2 CQ Differences Between Opera and Belting 83

6.2.1 Results 85

6.2.1.1 Average CQ of Opera Set versus Belting Set 85

6.2.1.2 Average CQ Statistics of Opera versus Belting Sets 85

6.2.1.3 Individual Average CQ Patterns for Opera Set 86

6.2.1.4 Individual Average CQ Patterns for Belting Set 87

6.2.1.5 Comparison of Opera and Belting CQ Patterns of One Singer 88

6.2.2 Discussion 88

6.2.2.1 Summary of Results 89

6.2.2.2 Discussion of Results 89

6.3 Relationships Between CQ, FO, Vibrato, and Larynx Height 93

6.3.1 Vibrato Differences Between Opera and Belting 93

6.3.2 Relationship Between CQ and Vibrato 94

6.3.3 Relationship Between Vibrato and Lx-Height 95

6.3.3.1 Discussion 95

6.4 Spectral Comparison of Opera and Belting 99

6.4.1 Comparison of Spoken Vowels 99

6.4.2 Analysis of the Opera G4/3:/ Spectra 101

6.4.3 Analysis of the Belting G4/3:/ Spectra 101

6.4.4 Analysis of the Opera E5/3:/ Spectra 101

6.4.5 Analysis of the Belting E5/3:/ Spectra 102

6.4.6 Discussion 102

6.5 Larynx Height Differences Between Opera and Belting 105

6.6 Mixed Quality 110

6.7 Conclusions 111

7 Synthesis and Perception 112

7.1 Introduction 112

7.2 Speech Synthesis 112

7.2.1 Synthesis By Analysis 112

7.2.2 Formant Synthesis 113

7.2.2.1 Vocal Tract Modelling for Formant Synthesizers 114

7.2.3 The KLSYN88 Synthesizer 115
7.2.3.1 Voicing Source Models 115
7.2.3.2 Vocal Tract Models 117

7.3 Perceptual Tests 119
7.3.1 Introduction 119
7.3.1.1 Categorical Perception 120
7.3.1.2 The Effect of Context 121
7.3.2 Perceptual Test 121
7.3.3 Results 124

7.4 Discussion and Conclusions 125

8 Conclusions and Future Research 126
8.1 Conclusions 126

8.2 Future Research 127

References 128

Appendices 141
Appendix [A] G4opera.spk Syntheisizer Algorithm 141
Appendix [B] G4belt.spk Synthesizer Algorithm (changes) 142
Appendix [C] E5opera.spk Syntheiszer Algorithm (changes) 142
Appendix [D] E5belt.spk Synthesizer Algorithm (changes) 143
Appendix [E] Answers to Perceptual Test Questionnaire 143

List of Figures

2.1 The three physiologic components of human speech production (from Lieberman & Blumstein, 1988). 4

2.2 Front view of the major structures of the pulmonary system (from Hixon et al., 1987). 4

2.3 Spirogram illustration of lung volumes and lung capacities (from Hixon et al., 1987). 5

2.4 Subglottic pressure in a tenor who sang a chromatic scale between the pitches of E3 and E4 (about 165-330 Hz) in soft, middle, and loud phonation (open circles, squares, and filled circles, respectively) (from Cleveland & Sundberg, 1983). 8

2.5 The pressure in the esophagus (P_{oes}), approximately corresponding to subglottic pressure and phonation frequency (f_0) in a professional baritone singer performing a coloratura passage (from Sundberg, 1987). 9

2.6 Recording of airflow, subglottic pressure, and sound level (curves marked A, P, and L) in a professional singer performing an ascending scale (upper graph) and a descending glissando (lower graph) (from Sundberg, 1987). 10

2.7 Simultaneous recordings of sound level (SPL), subglottic pressure, and phonation frequency in a professional singer singing an ascending major triad followed by a descending dominant-seventh chord with each tone beginning with a /p/. (from Sundberg, 1987). 10

2.8 Simultaneous recordings of sound level (SPL) pressure in the esophagus and phonation frequency in a professional singer singing a series of ascending and descending octave intervals (from Sundberg, 1987). 11

2.9 Schematic diagram of the action and location of the muscles of the hyoid complex (from Laver, 1980). 12

2.10 Schematic diagram of the principal laryngeal cartilages (from Laver, 1980). 13

2.11 Front and back views of the larynx (from Borden & Harris, 1984). 13

2.12 Frontal section of the larynx (from Borden & Harris, 1984). 14

2.13 The larynx from a superior view, showing the relationships among the thyroid, cricoid, and arytenoid cartilages, and the thyroarytenoid muscle (from Borden & Harris, 1984). 14

2.14 Schematic diagram of the location of the laryngeal muscles connecting the cricoid cartilage to the thyroid cartilage, and related organs (from Laver, 1980). 15

2.15 Schematic diagram of the action and location of the laryngeal muscles connecting the arytenoid cartilages to each other and to the cricoid cartilage, and related organs (from Laver, 1980). 16

2.16 Schematic presentation of the structure of the human vocal fold (from Sawashima & Hirose, 1983). 17
2.17 The acoustic production of a voiced sound (from Rossiter, 1993). 19
2.18 Idealized diagram of the source-filter model for speech production. 21
2.19 The sinusoidal relationship of air pressure at the first three formant frequencies (after Lieberman, 1977). 23
2.20 Classic F1-F2 chart in which a vowel is represented acoustically by its F1 and F2 frequencies (from Kent & Read, 1992). 24
2.21 Anatomical sketch of the human ear (from Campbell & Greated, 1987). 25
2.22 Schematic diagram of the human ear (from Campbell & Greated, 1987). 25
2.23 Changes in pressure and velocity between eardrum and oval window, due to reduction in area and lever action of the ossicles (from Campbell & Greated, 1987). 26
2.24 Cross-section of cochlea (from Campbell & Greated, 1987). 27
2.25 Arrival of a pressure pulse at the oval window (from Campbell & Greated, 1987). 28
2.26 Successive cross-sections of the basilar membrane showing the progress of a travelling wave. (from Campbell & Greated, 1987). 28
2.27 Amplitude envelope of basilar membrane vibrations when hearing a pure tone at different frequencies (from Campbell & Greated, 1987). 29
2.28 Illustration of a 'place theory', in which frequencies are distinguished by the positions of the corresponding amplitude envelope peaks on the basilar membrane of the inner ear (from Campbell & Greated, 1987). 29
2.29 (a)-(e) Electrical pulses on five different nerve fibres activated by the pure tone whose vibration curve is shown in (g). The sum of the signals on all five fibres is shown in (f) (from Campbell & Greated, 1987). 30
2.30 The smallest pitch change which can just be detected in a pure tone by the average listener (from Campbell & Greated, 1987). 32
2.31 The dominant harmonic in the perception of the pitch of a complex musical tone (from Campbell & Greated, 1987). 33
2.32 Critical bandwidth measurements (from Campbell & Greated, 1987). 34
2.33 The area of audible tones (from Gerber, 1974). 35
2.34 Contours of equal loudness for sine waves (from Campbell & Greated, 1987). 36
2.35 Basilar membrane amplitude envelopes (from Campbell & Greated, 1987). 37
2.36 Tristimulus diagrams representing the timbre of musical instruments (from Campbell & Greated, 1987). 39
2.37 Formants of the vowel sounds (a) "ee"; (b) "oo" (from Campbell & Greated, 1987). 41
2.38 Average source spectra from two alto and two tenor singers singing identical pitches (after Ågren & Sundberg, 1978). 42

3.1 A typical Lx waveform with open and closed phases (after Evans & Howard, 1993). 44
3.2 A few idealized cycles of a modal laryngograph output waveform (Lx) to illustrate the main phases in each cycle (from Abberton et al., 1989). 44
3.3 CQ plotted against time for a female speaking the word "bard". 45
3.4 CQ (%) against pitch scattergrams for the vowel /u:/ sung in belt and opera quality by a female (from Evans & Howard, 1993). 45
3.5 Typical configuration of the FFT analyzer (from AND AD-3523 manual). 46
3.6 A narrow band speech spectrogram (from Curtis & Schultz, 1986). 47
3.7 A wide band speech spectrogram (from Curtis & Schultz, 1986). 48
3.8 Average spectra for the vowel /a:/ sung at pitch E4 and pitch E5 in opera quality by a soprano. 48
3.9 Block diagram of a predictive coding system (from Ainsworth, 1976). 50
3.10 An LPC estimation of formant frequency and bandwidth of the word "bard" spoken by a female. 50
4.1 Geometric relationship between three laryngeal parameters (from Laver, 1980). 55
4.2 Average source spectra for three trained male singers pronouncing the vowel /α:/ on the same fundamental frequency in modal register (triangles) and falsetto register (circles). The spectra are represented by a curve showing how the spectrum contour deviates from the standard slope of -12 dB/octave. (from Sundberg, 1987). 56
4.3 Lx waveforms of the simple phonation types (from Laver, 1980). 57
4.4 Lx waveforms of compound phonation types: whispery voice and breathy voice (from Laver, 1980). 58
4.5 Spectrograms of steady-state vowels with six phonatory settings (from Laver, 1980). 59
4.6 Male-female comparisons of dimensions of the larynx. (a) Sagittal view of thyroid cartilage and (b) horizontal section showing difference in membranous length (from Titze, 1989). 60
4.7 Scaling of the glottis in terms of length L, amplitude of vibration A, and separation of the vocal processes W (from Titze, 1989). 62
4.8 Differences in medial surface contour and corresponding glottographic waveforms. (a) femalelike with linear convergence and (b) male-like with medial surface bulging (from Titze, 1989). 63
4.9 Spectrum envelopes for the vowel /u:/ as sung and spoken by a male professional opera singer (from Sundberg, 1987). 65
4.10 Sound level of the singer's formant in a baritone singing a chromatic scale on the vowel /ae:/ in soft (filled circles), middle (squares), and loud phonation (open circles) (from Cleveland & Sundberg, 1983). 66

4.11 Left: Tracings of frontal X-ray pictures of a male singer singing and speaking the same vowel. Right: Contours shown in frontal X-ray pictures of the deep pharynx when a subject deliberately raised and lowered his larynx (from Sundberg 1987). 67

4.12 The four lowest formant frequencies (f1, F2, F3, and F4) in the vowels indicated used by a professional soprano singing at various pitches (from Sundberg, 1987). 68

4.13 Mid-sagittal contours of the tongue body: dashed, solid, and chain-dashed curves pertain to the vowels /i:/, /ɑ:/, and /u:/. The upper left family of contours are from spoken vowels; the others were sung at the phonation frequencies indicated (from Sundberg, 1987). 69

4.14 A comparison of the closed quotients of opera and belting in a female singer (after Estill, 1988). 72

4.15 Comparison of average spectra for opera and belting at 5 frequencies (after Estill, 1988). 72

4.16 Spectrograms sung in "classical" mode (on the left) and "belt" mode (on the right) on the same pitch (after Schutte & Miller, 1993). 74

4.17 Vertical larynx height position observed in a professional soprano (from Sundberg, 1987). 75

4.18 A spectral comparison of the Chinese singing voice and the Western Operatic voice (from Wang, 1983). 76

5.1 A diagram of the connections between the microphone and electroglottograph output channels and the input channels of the 8-track ADAT recorder. 80

6.1 Average CQ patterns for the opera and belt sets (from final column of table 6.1). 85

6.2 Statistically significant differences in CQ between the opera and belting sets. 85

6.3 Individual CQ patterns for the opera set. 86

6.4 CQ medians for the belting sample. 87

6.5 Belting CQ patterns which display a dip at C5. 87

6.6 Belting CQ patterns which do not display a dip at C5 88

6.7 Comparison of opera and belting CQ patterns of singer CM. 88

6.8 Diagrams of the Lx waveform with CQ, F0, and larynx height (LH) analysis results from the production of the word "bard" on E4 and E5 for each singer (3 pages). 98

6.9 Average spectra for the spoken /3:/ vowel for each singer. 100

6.10 Average spectra for sung G4/3:/ tokens. 103

6.11 Average spectra for sung E5/3:/ tokens. 104

6.12 Larynx height comparisons for the four opera singers and five West End musical singers comparing the spoken word "bard" with the sung word on pitches E4 and E5 (4 pages). 106

6.13 Average spectral comparison of vowel-pitch token from exercises with similar from a song passage in same singer. 110

7.1 Cascade and parallel configurations of digital resonators (after Klatt, 1988). 114
7.2 The parameter listing for the KLSYN88 synthesizer (from Klatt, 1988). 116
7.3 Block diagram of the KLSYN88 formant synthesizer (from Klatt, 1988). 119
7.4 Average spectra of the synthesized sung vowels (right column) derived from the real sung vowels (left column). 122
7.5 Perceptual test results arranged according to the musical experience of each judge. 124
7.6 Average incorrect judgements (%) for both sung and synthesized tones. 124

List of Tables

2.1 Various speech sounds and their origin (from Borden & Harris,1984). 20
2.2 Rough correspondence between intensity and musical dynamic level for a single isolated 1000 Hz tone (from Campbell & Greated, 1987). 36
2.3 Some verbal scales used to rate timbre (after Bismarck) (from Campbell & Greated, 1987).
6.1 CQ values for sung vowels at different pitches. 84
6.2 Average CQ reults for spoken /3:/, /a:/, and /i:/ for each singer. 99

Terminology

For the purposes of this thesis:

1. "Bright" is defined as the acoustic quality associated with a tone where the spectrum is usually characterized by high partial energy in the region above 3 kHz.

2. "Dull" is defined as the acoustic quality associated with a tone where the spectrum is usually characterized by few partials above the fundamental with a steep energy decrease.

Chapter 1

Research Objectives and Report Structure

1.1 Introduction

Singing is a science as well as an artform. Understanding and tuning the physiological mechanisms that govern singing production to the extent that the appropriate physiological gestures become automatic, allows the singer to concentrate on the interpretative and communicative aspects of the performance. Modern technology has the potential of allowing singing students of any ability the chance to develop their vocal skills by providing objective technical feedback of their own voices and those of others. Modelling different vocal qualities using the appropriate speech and singing technology can create a body of data which students can draw from in order to clarify their vocal aims, thus speeding up their singing progress.

The main objective of this research is to model two different vocal qualities exhibited by female singers: the traditional, well documented "classical" opera quality (e.g., Bel Canto); and the relatively unexplored "belting" quality, a distinctively brassy quality which is heard in rock, gospel, and Broadway singing. The integrity of the models will be then be demonstrated by synthesis and perceptual tests.

1.2 Hypothesis

It is hypothesised that standard speech analysis and synthesis techniques are appropriate for sufficiently discriminating between, and therefore, modelling opera quality and belting quality in the female singing voice. The speech analysis techniques encompass non-invasive voice source and acoustic measurements.

1.3 Thesis Content and Structure

Chapter 2

provides a brief description of the human vocal and hearing systems.

Chapter 3

presents a description of standard non-invasive speech analysis techniques.

Chapter 4

discusses previous literature on vocal qualities and modes of production.

Chapter 5

describes the experimental and analysis methods used in this investigation.

Chapter 6

details the results of the experiments.

Chapter 7

is divided into two sections. The first section outlines current synthesis procedures and describes the main synthesizer used in this study. The second section describes the results of a perceptual test.

Chapter 8

provides conclusions to the above work and also includes a discussion of any future work which may be undertaken in order to further understanding of this subject area.

The appendices consist of the synthesizer algorithms for each of the synthesized tones used in the perceptual experiment, plus the results of a short questionnaire given to the judges of the perceptual test.

Chapter 2

Voice and Hearing Systems

2.1 Introduction

This chapter reviews the voice and hearing literature which contributes to the understanding of this research. The chapter is divided into two sections. The first part overviews the physiology of the vocal system. The second section concentrates on the hearing system, incorporating both physiology and perception.

2.2 The Human Vocal System

Human vocal communication utilises the same basic apparatus as that used to sustain life. The human vocal system comprises of three systems:

- the subglottal system
- the larynx, and
- the supralaryngeal vocal tract.

Catford (1977) describes the vocal system as being a pneumatic device made up of two bellows (the lungs), tubes and valves. Connected to the lungs is a large tube (the trachea), with a moveable piston with a vertical sliding motion (the larynx) sitting on top. The larynx acts as a valve. The space between the vocal folds is called the glottis. The structures situated above the larynx are collectively known as the supraglottal or supralaryngeal tract. This consists of three chambers (the pharynx, oral cavity and the nasal cavity), and other valves, such as the velum, the tongue, and the lips. A simplistic representation of the human vocal system is shown in figure 2.1.

2.2.1 The Subglottal System

Although this study is concerned with the analysis of the voice source and supraglottal vocal tract only, a description of the breathing mechanisms in singing is included here since the subglottal system drives the voice and can, under certain circumstances, substantially determine the vocal output.

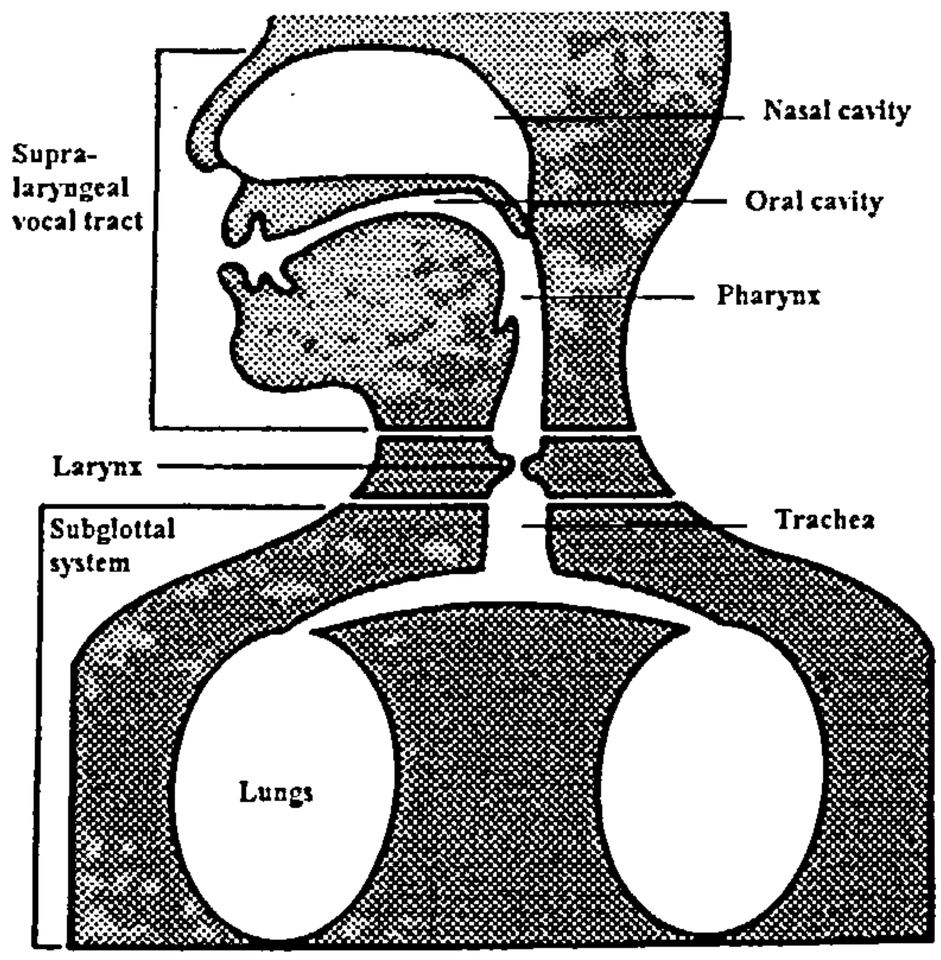


Figure 2.1. The three physiologic components of human speech production (from Lieberman & Blumstein, 1988).

The subglottal system, also known as the pulmonary system, is shown in figure 2.2. It consists of the trachea, bronchi, alveoli and lungs with their associated muscles. The trachea is joined below to the larynx and forms a tube of 18 cartilages enclosed by the trachealis muscle. The bottom of the trachea stems into two smaller tubes called bronchi. Each of these inserts into one of the two lungs where it branches further into bronchioles and ducts ending in a great number of minute air sacs called alveoli. The subglottal air ways are contained in the upper of two cavities making up the torso. This upper cavity is called the thorax (chest) and is a barrel-shaped bone and cartilage cage containing the pulmonary system, respiratory airways and the heart. The lower cavity is called the abdomen. It contains mainly the digestive system and other organs such as the kidneys, liver and reproductive organs. The thorax and abdomen are separated by a dome-shaped muscular structure called the diaphragm.

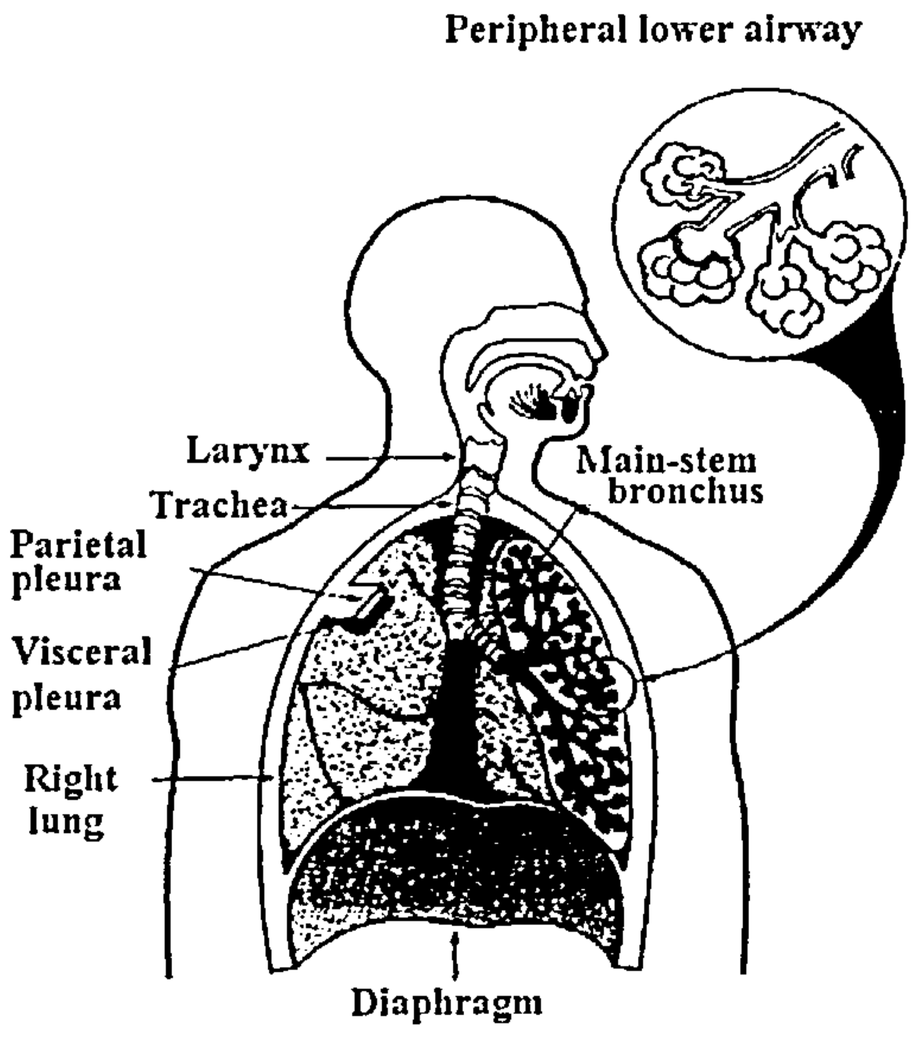


Figure 2.2. Front view of the major structures of the pulmonary system (from Hixon et al., 1987).

2.2.1.1 The Lungs

The lungs can be thought of as balloons with a large capacity to expand or collapse. They are prevented from doing so by being in close relationship with the thoracic cage. The lungs sit on top of the diaphragm and are encased in a double-walled airtight sac called the pleural cavity. The inner membrane (visceral pleura) lines the outside surface of the lungs. The outer membrane (parietal pleura) covers the inside of the thoracic cage. This pleural linkage of the lungs and thorax is essential in respiration since the lungs and thorax move as a "lungs-thorax unit" (Hixon, 1987). Due to pleural linkage, the air pressure within the lungs is very sensitive to forces applied to it by the thorax, the diaphragm and the abdomen.

2.2.1.2 The Subglottal System in Respiration

The lungs are the main organs of respiration. They act as a pump in order to transfer air to and from the alveoli where gaseous exchange occurs, maintain constant blood gas pressures within the body's cells, and are made up of mainly elastic fibres which enable them to change shape and size.

Respiration in breathing is accomplished automatically using a variety of homeostatic physiological mechanisms to drive the lungs at the required rate (Proctor, 1980). In certain types of speech and singing production, some of these mechanisms are consciously controlled to sustain phonation and to help generate different qualities of phonation.

Some simple mechanical principles govern respiration. These involve air pressures, forces, volumes, capacities, air flow, and the mechanical motion of individual structures within the respiratory system. A full overview of these principles is not provided here, since excellent descriptions are to be found in Hixon et al. (1987), and Proctor (1980). However, figure 2.3 illustrates some of the main terms used in respiratory function.

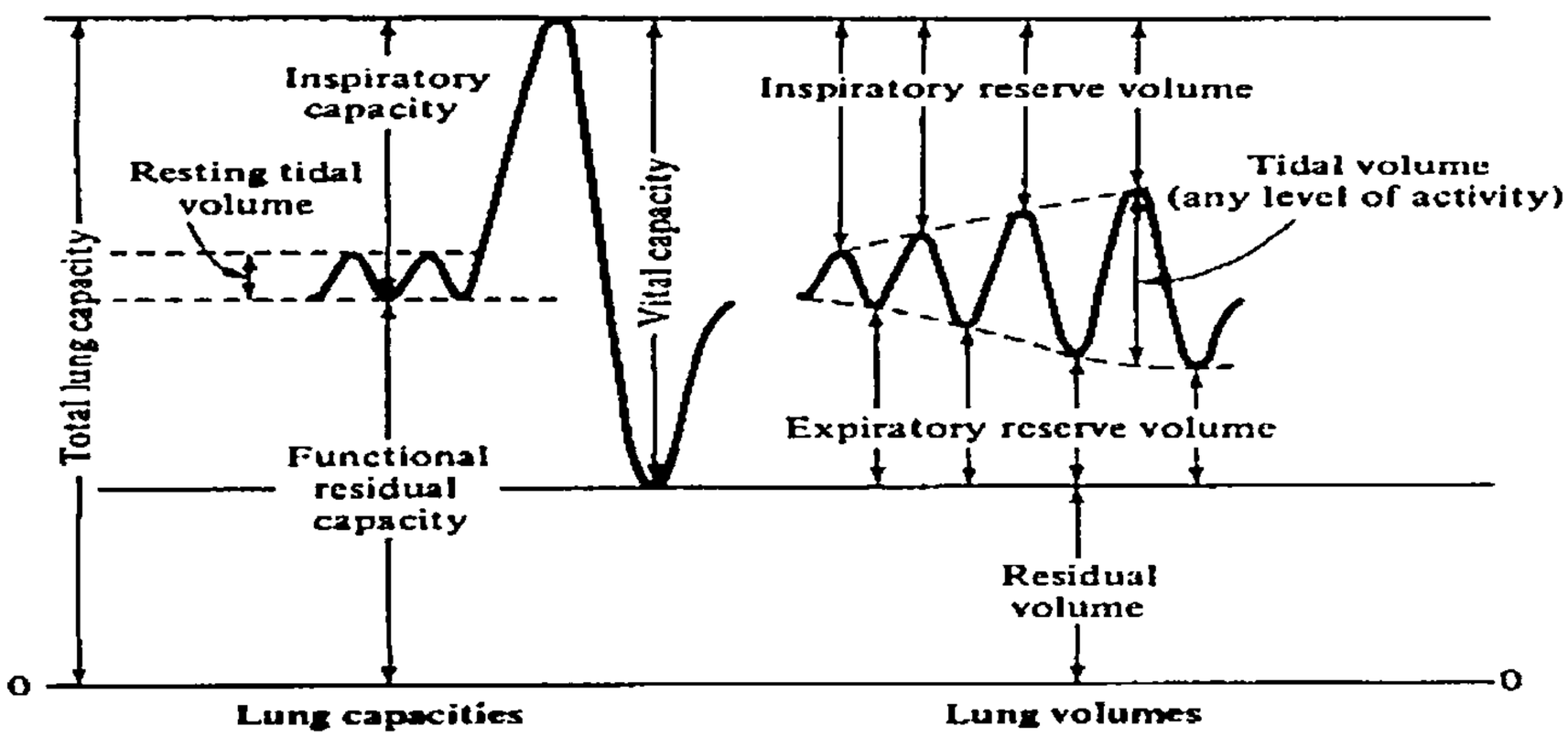


Figure 2.3. Spirogram illustration of lung volumes and capacities (from Hixon et al., 1987).

The basic principles underlying respiration are that: air movement occurs from regions of higher pressure to areas of lower pressure; and air flow velocity depends on the pressure difference between the areas concerned.

By muscular and passive recoil forces the volume of the lungs and consequently the pressure within the lungs, can be varied to achieve the appropriate air pressure difference between atmospheric pressure and within the alveoli of the lungs for air flow to and from the lungs.

Two stages of air flow occur in respiration: inspiration, when air flows into the lungs; and expiration, when air flows out of the lungs. Expiration is an important process especially in singing.

In inspiration, alveolar pressure must be lower than atmospheric pressure in order to create a pressure gradient favouring inward flow. At the resting expiratory level, which is with the airways open (i.e., with an open glottis) and the respiratory system in a neutral position, alveolar pressure equals atmospheric pressure. In order to decrease alveolar pressure required for inspiration, the lungs must be enlarged which will expand the air within them leading to a decrease in alveolar pressure. This is achieved by enlarging the size of the thorax using muscular forces which leads to a corresponding movement of the lungs due to pleural linkage.

At the end of inspiration alveolar pressure is equal to atmospheric pressure. In order for expiration to occur, alveolar pressure must be greater than atmospheric pressure. This happens by squeezing the lungs-thorax unit, which increases the alveolar pressure thus enabling air to flow outward. Both active (muscular) and passive (non-muscular) forces are used in expiration. In quiet breathing, expiration above the resting level is generally termed passive, even though some muscular energy is exerted.

2.2.1.3 Respiration in Phonation

If the breath is held, "changes in the contours of the thorax and belly must be equal and opposite" (Proctor, 1980). That is, as the thorax gets smaller the diaphragm is pushed downward and the belly is pushed outward. Sustained phonation requires :

1. a relatively constant average pressure
2. a relatively constant average airflow
3. a resistance of the upper airway (the approximation of the vocal folds) which is referred to as glottal resistance
4. both passive and active forces in regulating air pressure (Hixon, 1987).

The muscular activity required to maintain a constant subglottal pressure in phonation is dependent on lung volume and the amount of alveolar and relaxation pressures (Hixon et al., 1987). With the use of muscular pressure the subglottal pressure can be varied. When the lungs are filled with air the lung volume is large and high subglottal pressure is generated. The main muscles involved in actively regulating subglottal pressure are:

1. The inspiratory muscles: diaphragm and external intercostals
2. The expiratory muscles: the internal intercostals, external oblique, and rectus abdominis
3. The latissimus dorsi which is inspiratory or expiratory.

Sustained phonation is initiated at near the total lung capacity and is preceded by a deep inspiration, involving contraction of the muscles of inspiration; the diaphragm, external intercostals and accessory muscles. The vocal folds are positioned close together at this point.

2.2.1.4 Lung Volumes in Phonation

We normally inhale and exhale about .5 litres every 5 seconds in normal breathing. Airflow, the amount of air escaping the lungs per unit time, is consequently very low, about .1 litre per second. In normal breathing, lung volume is varied only slightly at a level just above the functional residual capacity (FRC). The FRC is the amount of air that is held in the lungs at resting expiratory level. However, in phonation, a number of factors defined by Proctor (1980) determine the lung volume at which phonation is initiated. These include:

1. inspiring sufficiently in between phrases so that there is no break in the flow of the song or speech
2. inspiring sufficiently enough to complete the phrase
3. controlling subglottal pressure in order to produce the required sound intensity
4. controlling airflow in order to sustain the desired tone quality for the phrase.

In speech, phonation is mostly initiated at about 50% of the vital capacity (VC) (Proctor, 1980). The VC is the maximum volume of air that can be expelled from the lungs after a maximum inspiration and corresponds to the amount of air used for breathing and phonation.

Sundberg (1987) suggests this is because we are taking advantage of the passive exhalatory forces in establishing the subglottal pressures required for normal speech. In normal and loud reading these passive forces are also used since people tend to take a breath when lung volumes are close to FRC. Normal breathing is characterized by its regularity of rate and rhythm, whereas in conversational speech rate and rhythm are irregular (Hixon, 1987).

The demands on lung volumes are even greater in singing than in speech. Whereas in speech a breath normally takes about every 5 seconds, in singing, phrases commonly extend over 10 seconds. There are less opportunities to take a breath, and they must often be taken during very brief pauses. Consequently, higher lung volumes are used in singing than in speaking. Long phrases are initiated at very high lung volumes, almost 100% of the VC. Also, a trained singer can use nearly all his/her vital capacity, well below the FRC. A trained singer uses a greater portion of her/his total lung capacity. Gould (1977) has demonstrated that singers have a vital capacity about 20% larger than non-singers.

2.2.1.5 Subglottal Pressure in Phonation

Subglottal pressure is raised when the abdominal muscles contract after inspiration. It is dependent on the amount of contraction and glottal resistance to airflow. Different singers use highly varying subglottal pressures, and voice category and type of voice (or singing technique) are important factors (Sundberg, 1987).

Loudness is related to subglottal pressure. In order to change loudness, subglottal pressure must be changed appropriately. Figure 2.4 illustrates that pressure is raised with increased loudness for a tenor singing tones of a chromatic scale in piano, mezzoforte and forte. It also shows that pressure increases with rising phonation frequency. This seems to be true for most singers in the upper parts of their phonation frequency ranges.

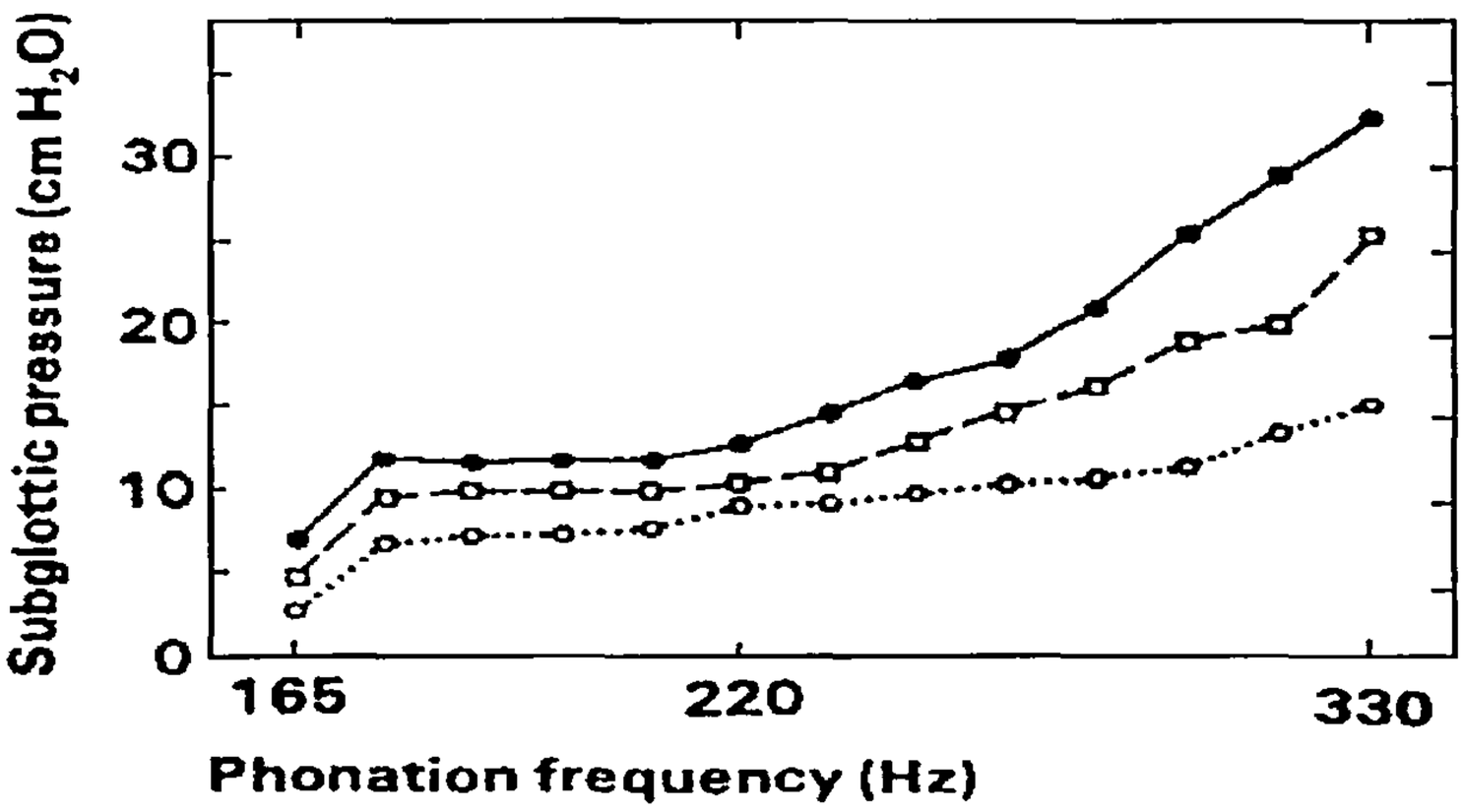


Figure 2.4. Subglottic pressure in a tenor who sang a chromatic scale between the pitches of E3 and E4 (about 165-330 Hz) in soft, middle, and loud phonation (open circles, squares, and filled circles, respectively). The pressure is increased for increasing loudness, but it is also raised with rising phonation frequency (from Cleveland & Sundberg, 1983).

There must be a delicate balance of inspiratory and expiratory muscle effort with the elastic recoil forces related to lung volume at any given instant in order to produce a tone of a desired loudness (Proctor, 1980). In order to sing a soft tone decreasing inspiratory effort and increasing expiratory effort is required, whereas for a loud tone, expiratory effort must be delicately initiated at the moment of attack but should then gradually increase in order to sustain this loudness near the residual volume (RV, which is the volume of air that always remains in the lungs after maximum expiration, and remains even after death).

In singing subglottal pressure varies rapidly adapting to changes in loudness and phonation frequency demanded by the music. This is illustrated in figure 2.5 which shows that during a coloratura passage a singer's subglottal pressure changes in synchrony with phonation frequency.

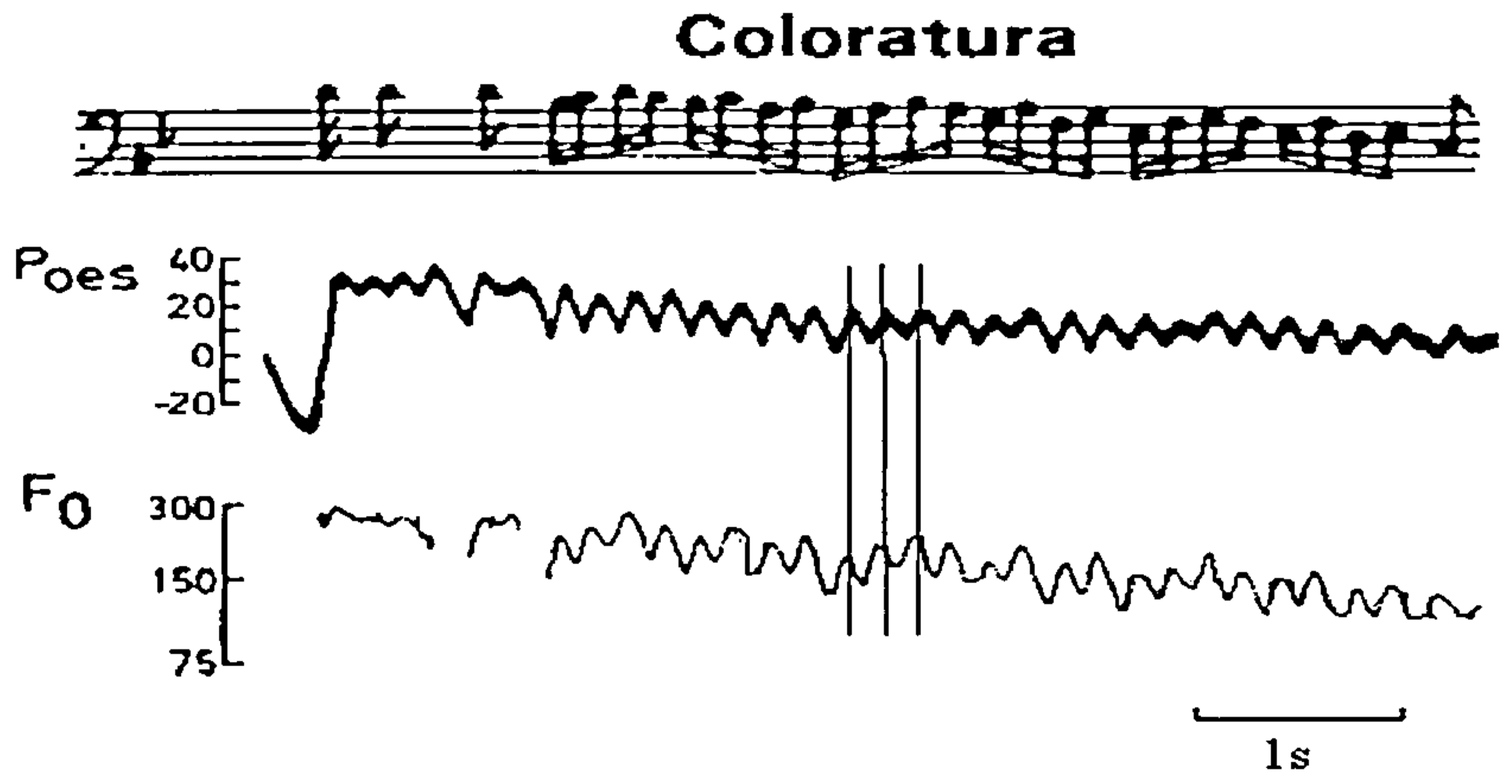


Figure 2.5. The pressure in the esophagus (P_{oes}), approximately corresponding to subglottic pressure and phonation frequency (F_0) in a professional baritone singer performing a coloratura passage. Both pressure and frequency increase and decrease once for each tone, or approximately 6 times per second (from Sundberg, 1987)

2.2.1.6 Subglottal Pressure and Airflow

Airflow increases with an increase in subglottal pressure if glottal resistance (the resistance against airflow through the glottis) is constant. Glottal resistance can be varied greatly depending on the degree of adduction of the vocal folds. It is possible to produce similar tones with widely varying airflows. Assuming an appropriate adduction, within the opera tradition, the less air consumption, the better the singer. Constant glottal leakage of air due to the vocal folds not contacting properly results in a high consumption of air and is a sign of poor voice technique. Constant glottal leakage is characterised by a "breathy" vocal quality (Laver, 1980). Conversely, one can greatly lower airflow by tensing the vocal folds together which raises the subglottal pressure high enough to be able to overcome the glottal resistance. This leads to a very high subglottal pressure and is characterised by a "pressed" or strained vocal quality (Laver, 1980).

A doubling of subglottal pressure increases sound level of phonation by 9 dB (Sundberg, 1987). Subglottal pressure also slightly increases phonation frequency.

Rubin et al. (1967) have found that airflow tends to increase when both phonation frequency and loudness are increased simultaneously. However, higher tones do not necessarily consume more air than lower tones since they have shown that trained singers can perform a rising glissando with constant loudness with no increase in airflow.

Airflow, then, does not necessarily depend on phonation frequency since it has been shown that even though doubling the fundamental frequency doubles the frequency of the glottal opening, it actually halves the time the glottis is open in each cycle so air consumption is the same for higher tones as for lower tones.

For non-singers airflow does tend to increase with phonation frequency in the falsetto register, though for trained singers there seems to be no dependency of airflow on either loudness or phonation frequency. This is demonstrated in figure 2.6, which shows that for a professional singer performing a rising scale and descending glissando, airflow remains constant whilst subglottal pressure rises with pitch.

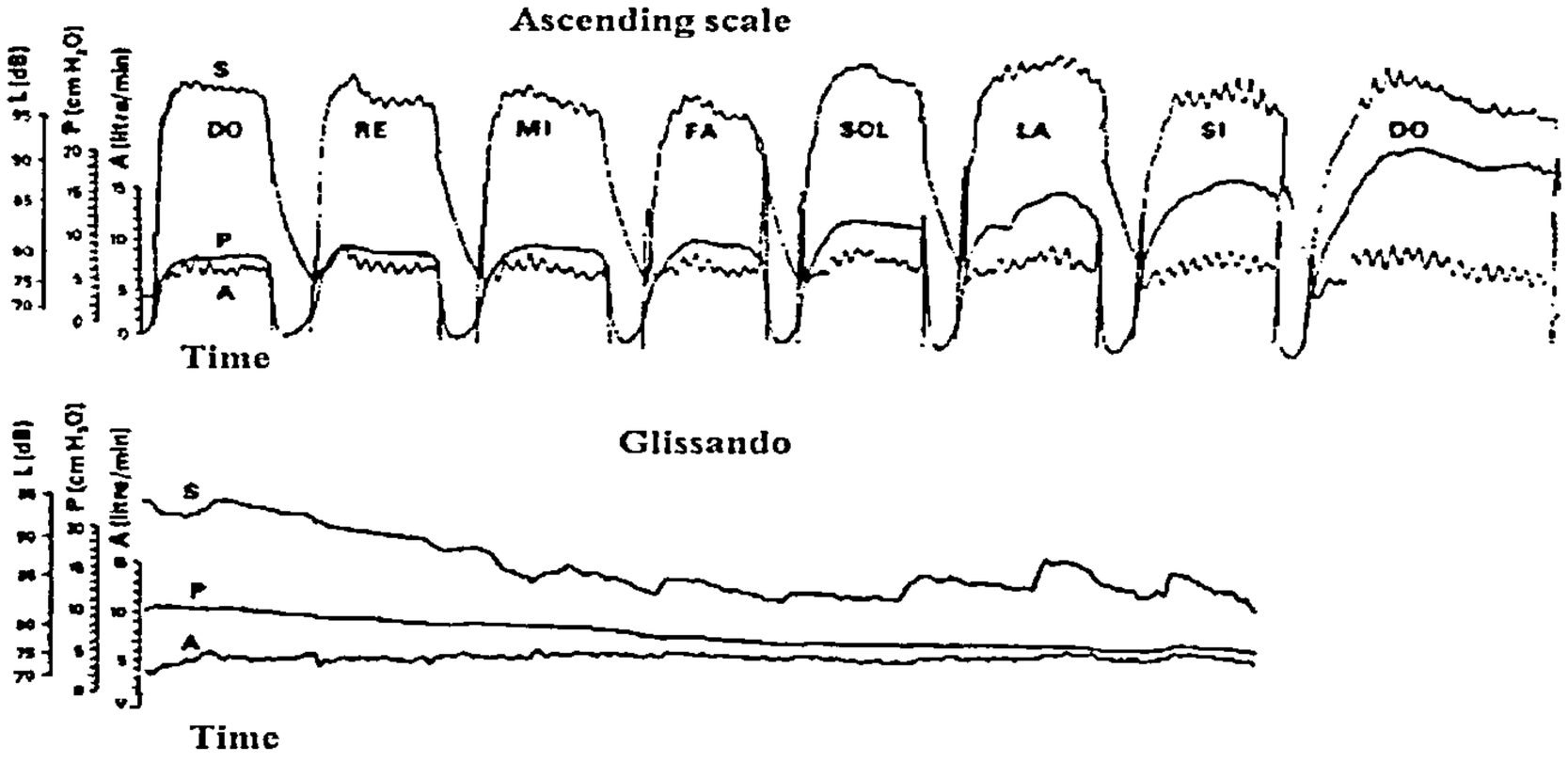


Figure 2.6. Recording of airflow, subglottic pressure, and sound level (curves marked A, P, and L) in a professional singer performing an ascending scale (upper graph) and a descending glissando (lower graph). Airflow is kept essentially constant, while subglottic pressure rises with pitch (from Sundberg, 1987).

A demonstration of the very rapid changes in subglottal pressure needed in quickly changing pitch is given in figure 2.7 for a singer singing an ascending major triad and a dominant-seventh chord with each tone beginning with a /p/.

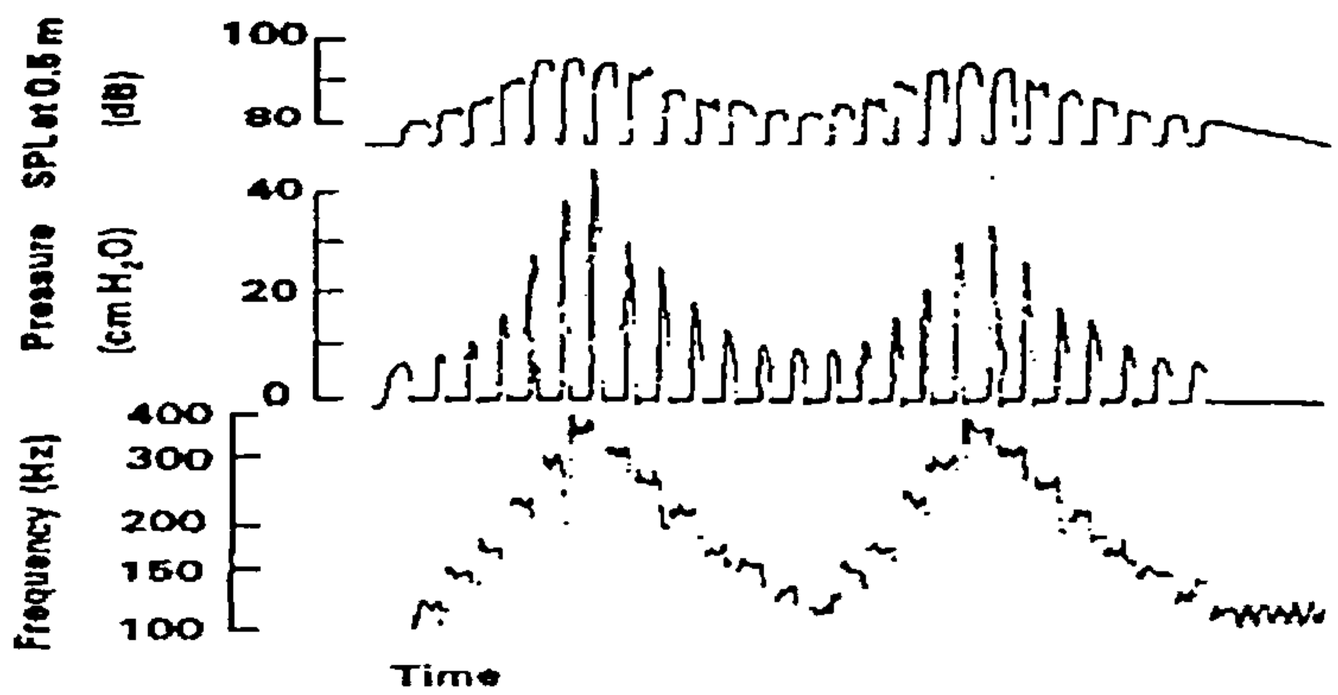


Figure 2.7. Simultaneous recordings of sound level (SPL), subglottic pressure, and phonation frequency in a professional singer singing an ascending major triad followed by a descending dominant-7th chord with each tone beginning with a /p/. Higher tones are sung with higher pressure. Pressure regulates loudness, and the musically most stressed note is the one that follows the highest pitch. This stressed note is given the highest pressure (from Sundberg, 1987).

This can also be seen in figure 2.8, where subglottal pressure is substantially raised for a pitch in the upper part of the singer's range than for a lower one for a professional singer singing a series of ascending and descending octave intervals (Sundberg, 1987).

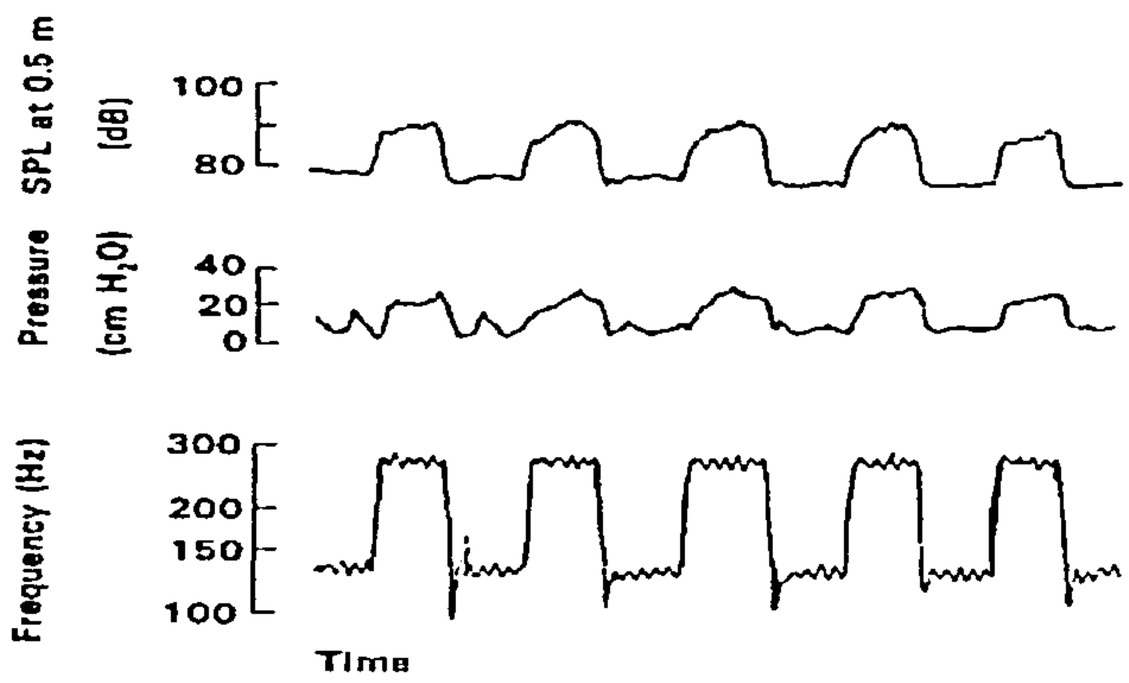


Figure 2.8. Simultaneous recordings of sound level (SPL) pressure in the esophagus and phonation frequency in a professional singer singing a series of ascending and descending octave intervals. As higher tones are sung with higher pressure, the pressure has to be changed in accordance with pitch (from Sundberg, 1987).

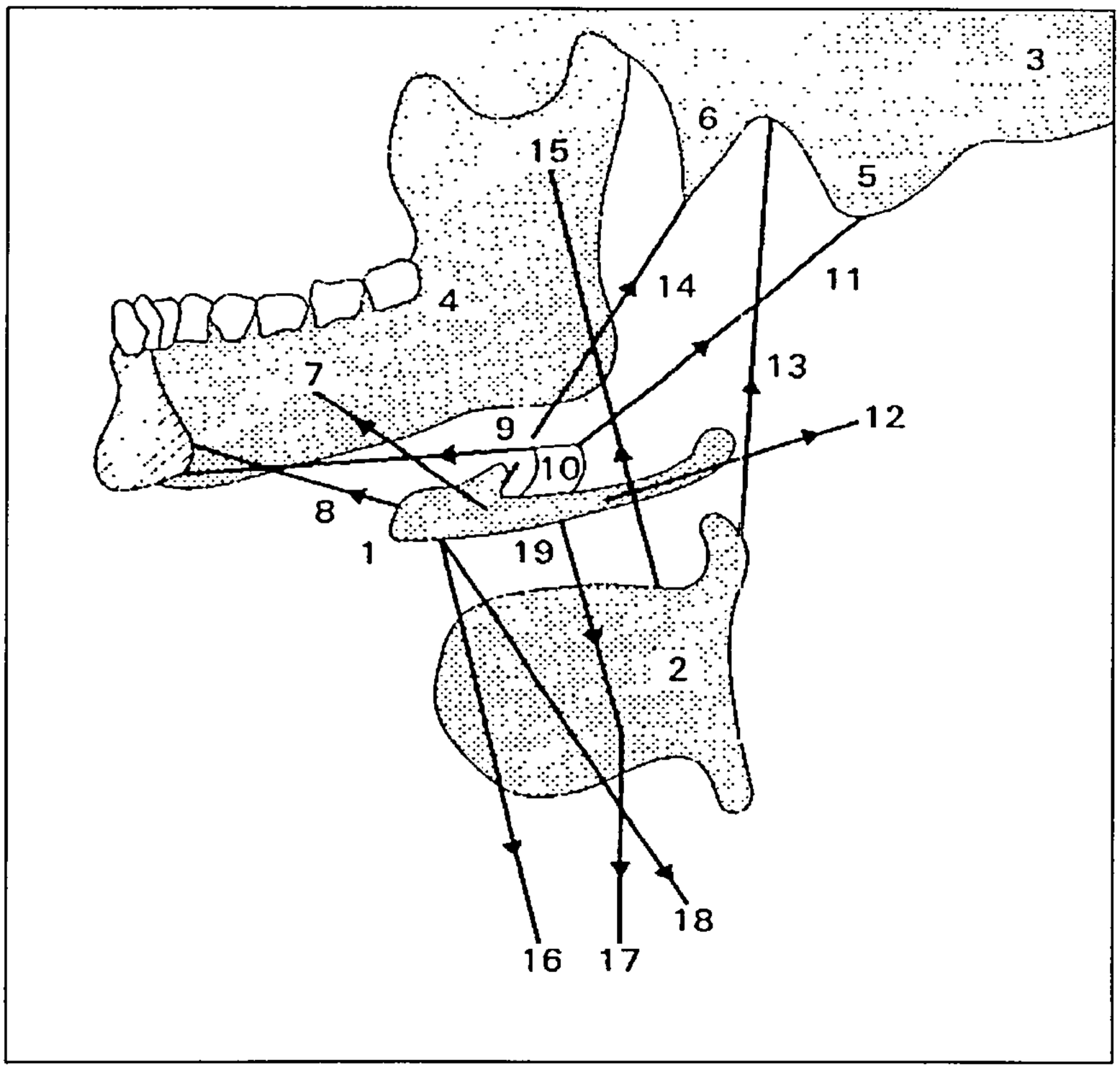
2.2.1.7 Muscular Combinations in Phonation

Different combinations of muscular activity can be used to accomplish the desired subglottal pressure. However, some combinations are better than others.

Singers use different positionings of the abdominal wall. Some singers sing with their abdominal wall pulled in (belly-in) whilst others sing with an expanded abdominal wall (belly-out). With the belly-in strategy, the contraction of the abdominal wall pushes the diaphragm into the rib cage, whilst with the belly-out method the diaphragm is flatter. Experimental findings comparing these different breathing techniques and diaphragm and abdominal activity during speech and singing are to be found in Proctor (1980); Leanderson et al., (1987); and Watson & Hixon (1983).

2.2.2 The Larynx

The larynx is suspended from the hyoid bone, a small horseshoe-shaped bone just underneath the jaw. The muscles of the hyoid bone form what Laver (1980) describes as a "triple sling system", shown schematically in figure 2.9. Vertical movement and support of the larynx is accomplished by the infrahyoid and suprahyoid muscles. These belong to a group of muscles known as the extrinsic laryngeal muscles. The fine balancing of the tensions of these muscular hyoid slings also provide precise articulatory control of the jaw.

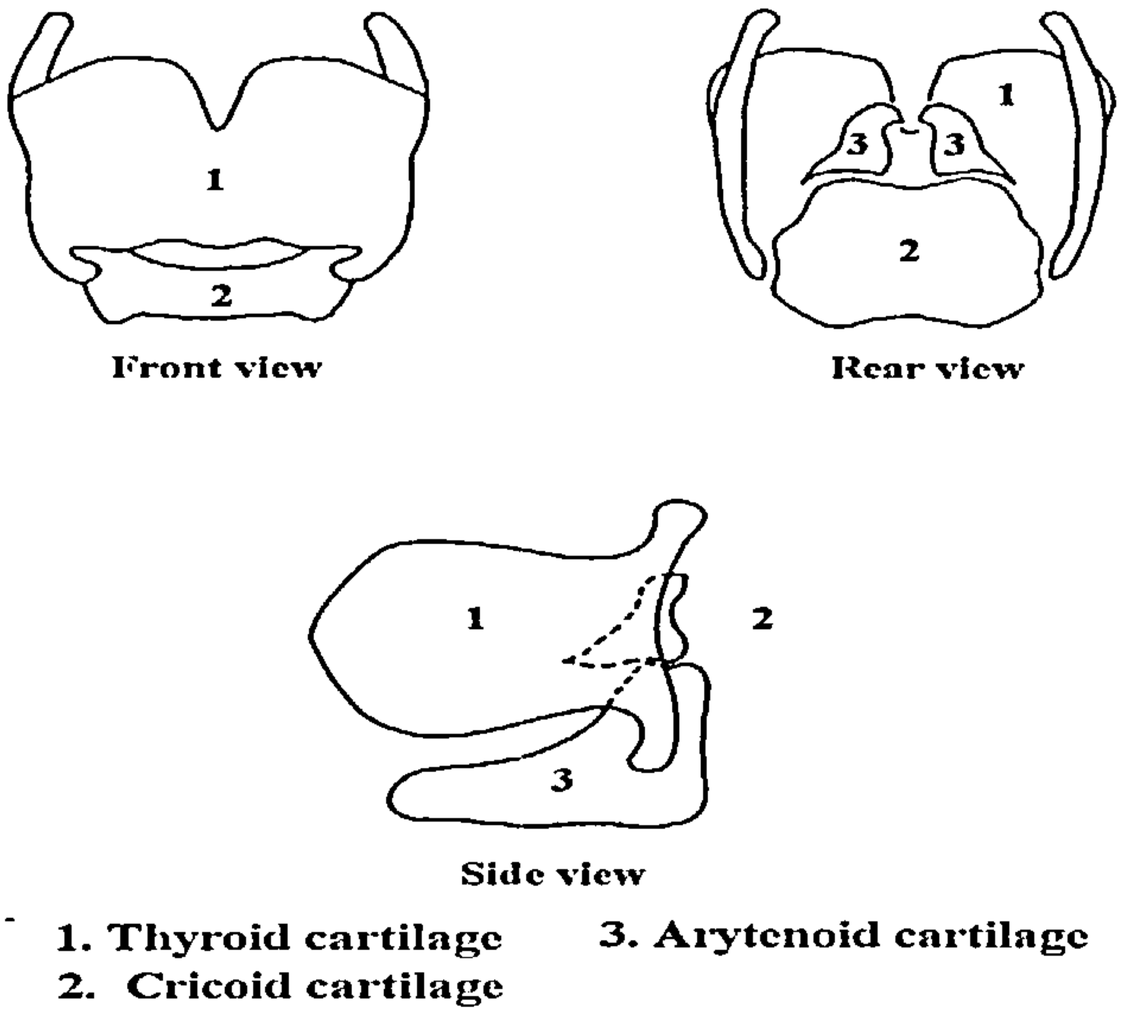


1. Hyoid bone
2. Thyroid cartilage
3. Skull
4. Internal surface of lower jaw
5. Mastoid process
6. Styloid process
7. Mylohyoid muscle
8. Geniohyoid muscle
9. Anterior belly of the digastricus muscle
10. Fascial sling
11. Posterior belly of the digastricus muscle
12. Middle pharyngeal constrictor muscle
13. Stylopharyngeus muscle
14. Stylohyoid muscle
15. Palatopharyngeus muscle
16. Sternohyoid muscle
17. Sternothyroid muscle
18. Omohyoid muscle
19. Thyrohyoid muscle

Figure 2.9. Schematic diagram of the action and location of the muscles of the hyoid complex (from Laver, 1980).

The laryngeal framework consists of five separate cartilages: the epiglottis, thyroid, cricoid and two arytenoid cartilages. They are connected by ligaments and muscles, and the total framework is covered by mucous membrane. The larynx is situated in front of the lower pharynx which leads to the esophagus and stomach. The main role of the epiglottis, a leaf shaped cartilage, is to cover the entrance to the larynx during swallowing. This allows food and liquids to travel past the larynx into the esophagus.

The other three types of cartilage take part in phonation. The relative positions of the thyroid, cricoid, and arytenoid cartilages to each other are shown in figure 2.10. Various views of the larynx showing the relationship between its components are presented in figures 2.11, 2.12, and 2.13.



1. Thyroid cartilage
2. Cricoid cartilage
3. Arytenoid cartilage

Figure 2.10. Schematic diagram of the principal laryngeal cartilages (from Laver, 1980).

2.2.2.1 The Components of the Larynx

The following section provides a summary of the main component functions of the larynx.

The Thyroid Cartilage

The thyroid cartilage is large and shield shaped. Together with the cricoid it forms a protecting structure for the larynx. It consists of two side plates fused anteriorly under a central V-shaped notch. The two plates form a more acute angle in males than in females, forming what is known as the "Adam's Apple". The two plates enclose the arytenoids and are widely separated at the back. They also extend above and below forming two superior horns which project toward the hyoid bone above via the thyrohyoid muscle and ligament, and two inferior horns which articulate with the cricoid below via the cricothyroid muscle. The inner surface of the fused anterior angle of the thyroid also forms the points at which the true vocal folds and false folds are anteriorly attached.

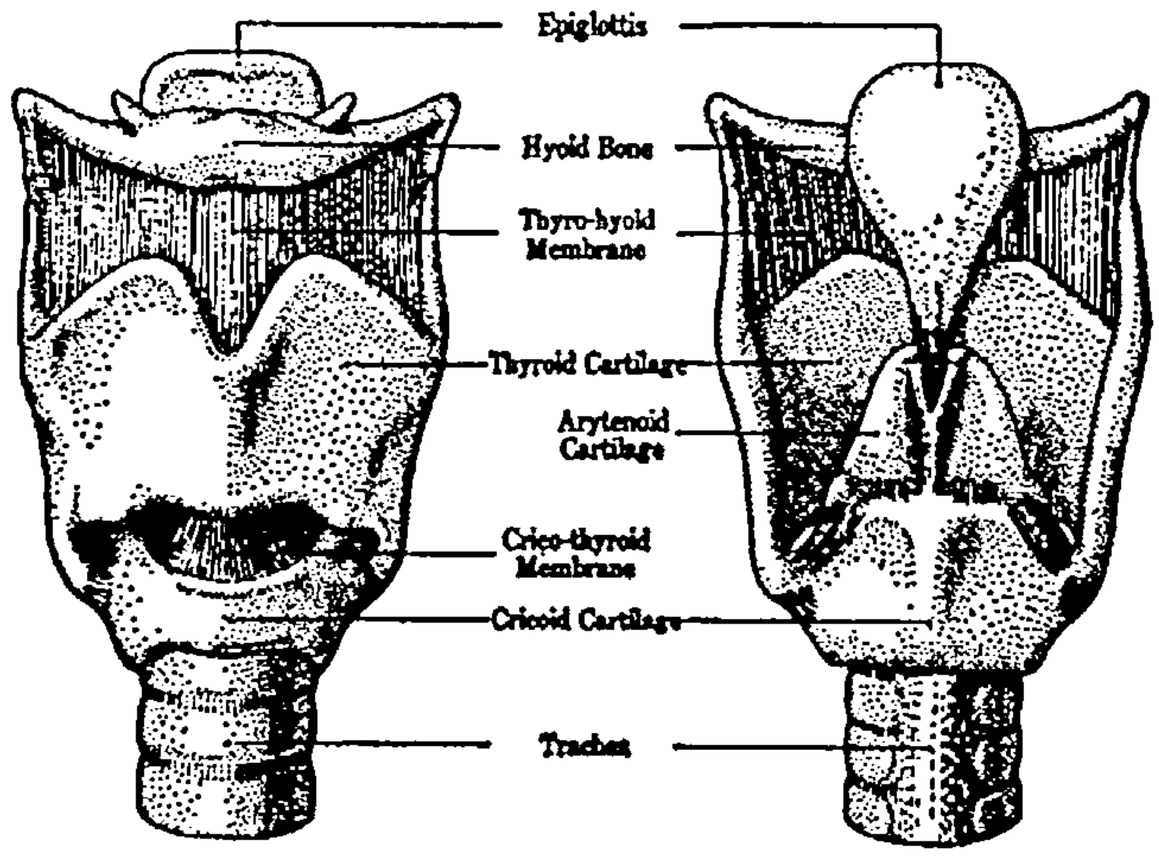


Figure 2.11. Front and back views of the larynx (from Borden & Harris, 1984).

The Cricoid Cartilage

The cricoid cartilage forms the top ring of the trachea. However, unlike the other horseshoe-shaped tracheal rings, the cricoid forms a complete ring. The posterior part of the ring forms a distinctive large plate like the signet of a ring. The cricoid articulates with the thyroid and the arytenoid cartilages.

The Arytenoid Cartilages

The arytenoids are small pyramid-shaped cartilages which sit on top of and articulate with the top of the back plate of the cricoid. They can be moved very quickly and precisely, and can rotate horizontally, vertically and slide from side to side thereby separating or bringing together the posterior ends of the vocal folds. The small projections at the base of each arytenoid are the vocal processes. The posterior ends of the true vocal folds are attached to these vocal processes. The apex of each arytenoid is attached to the posterior ends of the false vocal folds (the ventricular folds). The normal position for breathing is with the arytenoids apart. This results in a triangular gap between the vocal folds, called the glottis. When the arytenoids are closely approximated, the vocal folds close together and act as a shut valve.

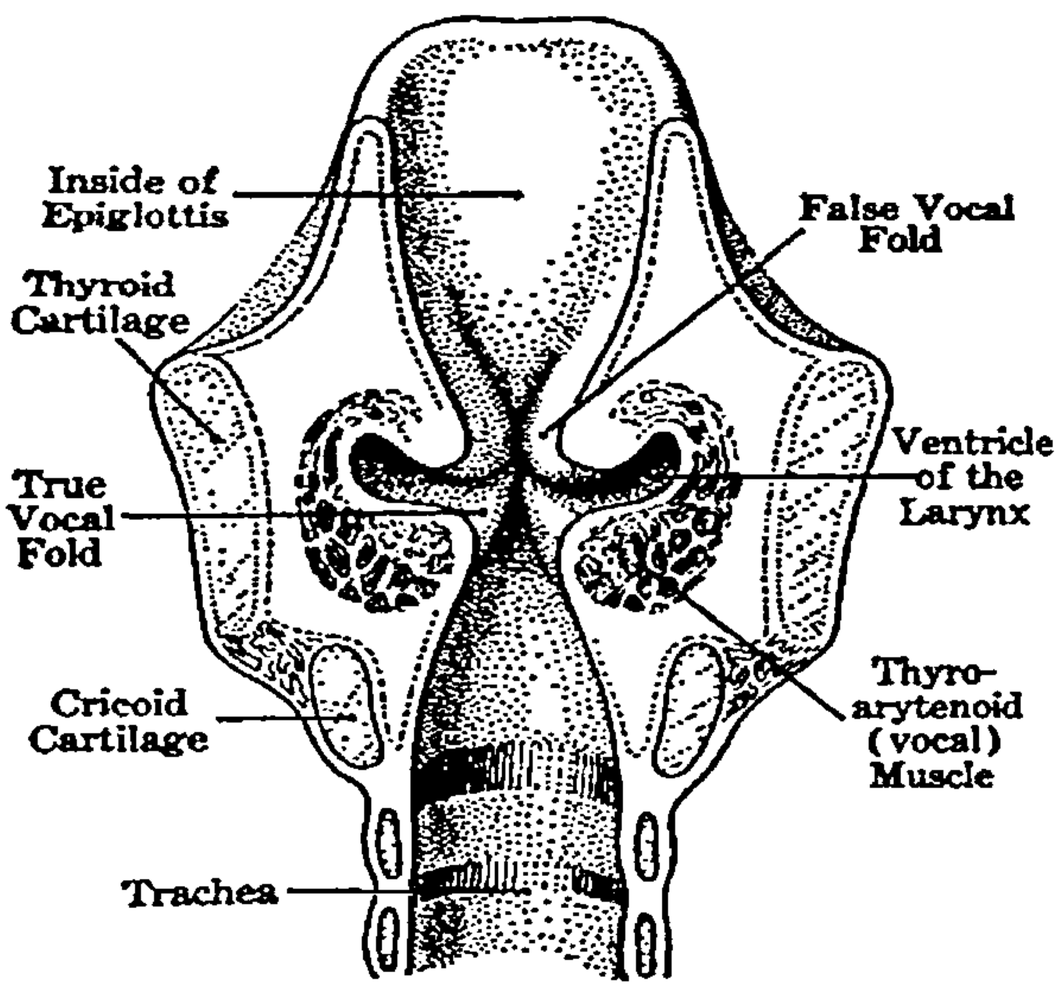


Figure 2.12. Frontal section of the larynx. Notice the constrictions formed by the ventricular folds and the 'true vocal folds' below (Borden & Harris, 1984).

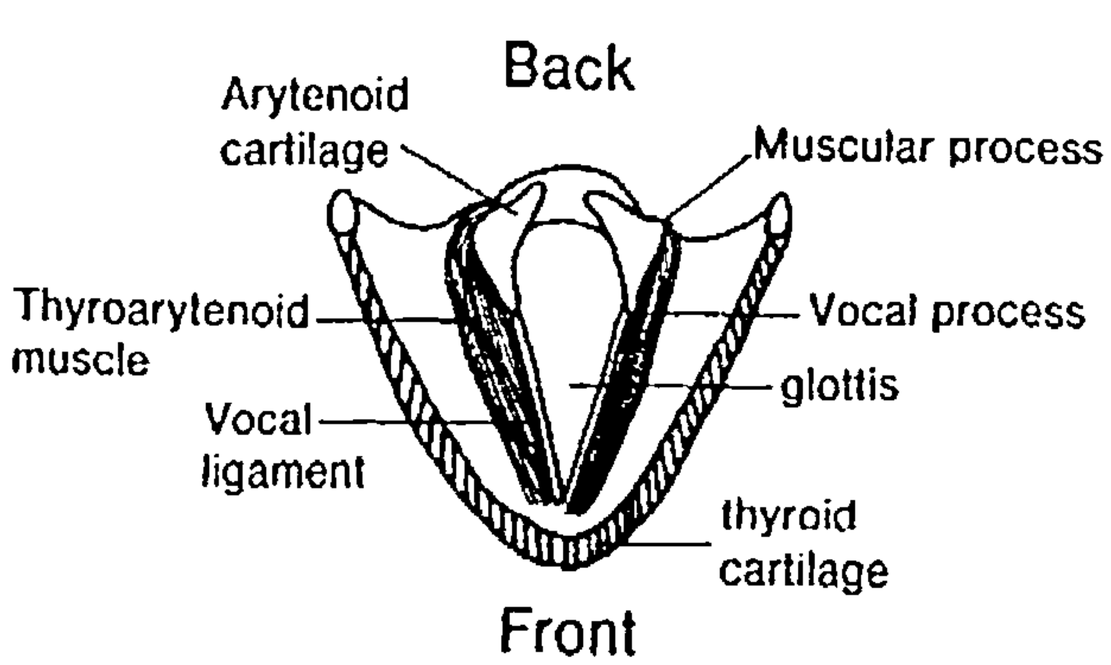


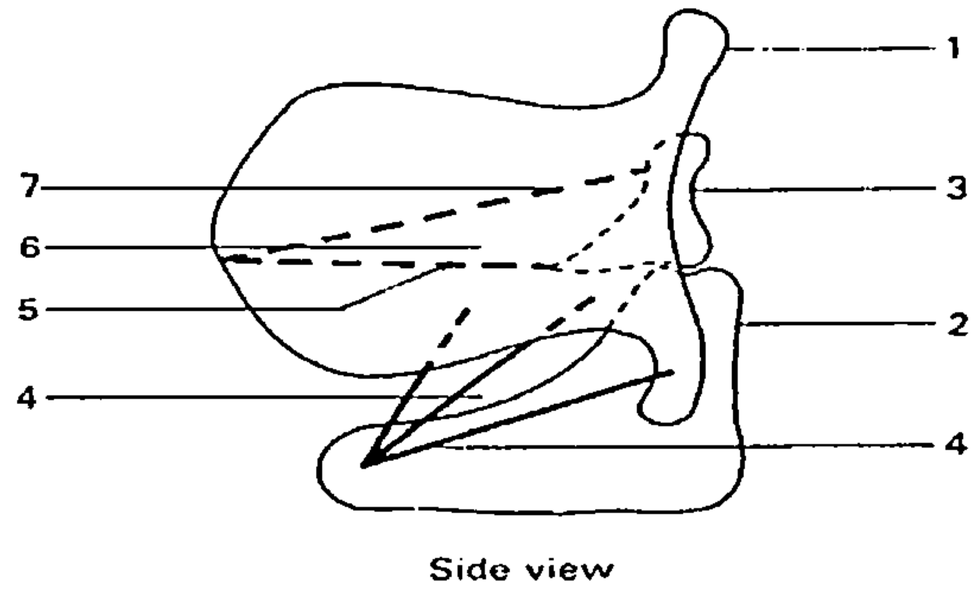
Figure 2.13. The larynx from a superior view, showing the relationships among the thyroid, cricoid, and arytenoid cartilages, and the thyroarytenoid muscle (from Borden & Harris, 1984).

The Laryngeal Musculature

The main laryngeal musculature can be divided into two groups (Laver, 1980):

1. those that change phonation frequency by changing the position of the cricoid relative to the thyroid, shown schematically in figure 2.14

2. those that control abduction and adduction by changing the position of the arytenoids relative to the cricoid, shown schematically in figure 2.15.



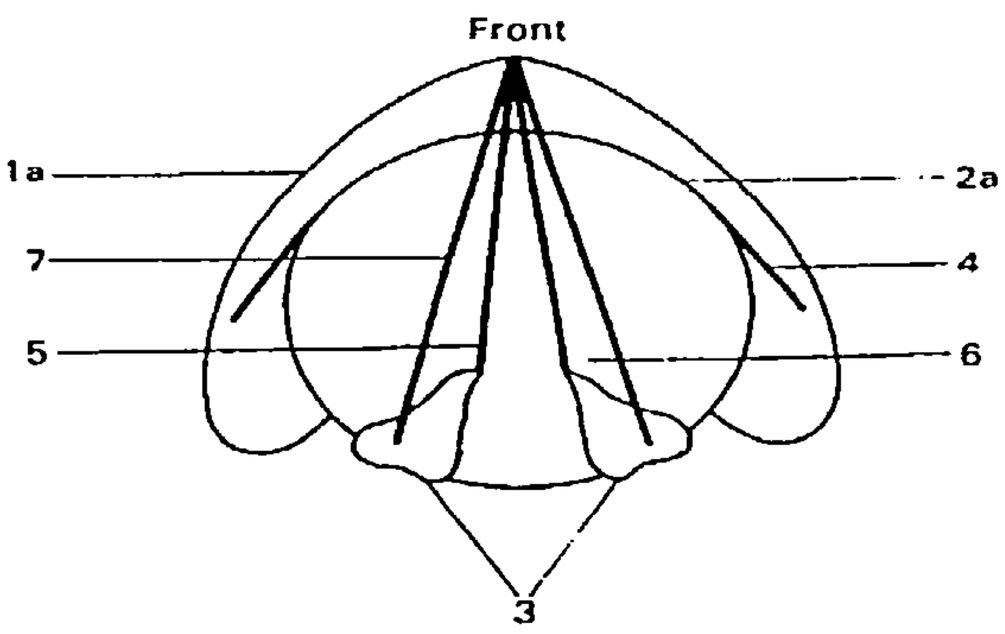
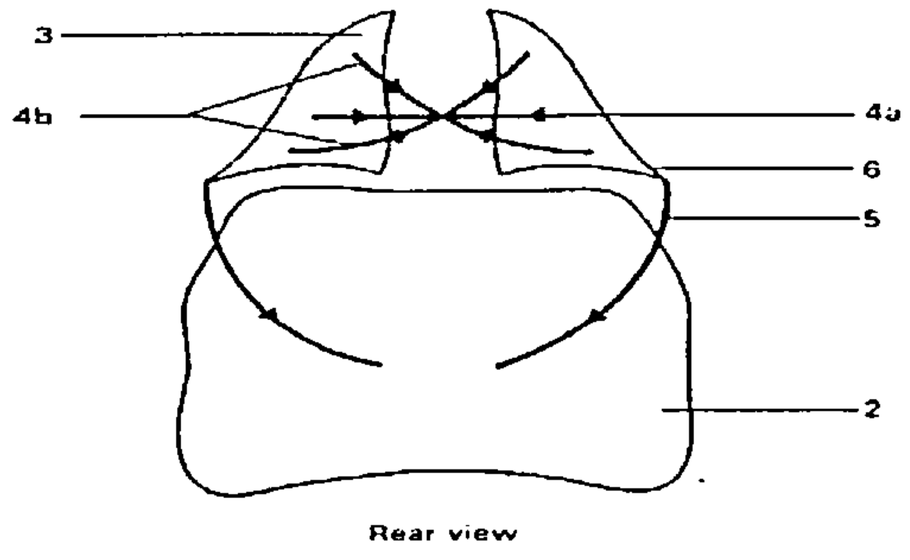


Figure 2.14. Schematic diagram of the location of the laryngeal muscles connecting the cricoid cartilage to the thyroid cartilage, and related organs (from Laver, 1980).

1. Thyroid cartilage
1a. External edge of thyroid
2. Cricoid cartilage
2a. External edge of cricoid
3. Arytenoid cartilages
4. Cricothyroid muscle
5. Glottal border of true vocal fold
6. Ventricle of Morgagni
7. Inner border of ventricular or false vocal fold

The Muscles Determining Phonation Frequency

The muscles which change phonation frequency are the cricothyroid (CT) muscles and the thyroarytenoid (TA) muscles. The paired cricothyroid muscles connect the thyroid to the cricoid and are the main muscles for determining phonation frequency by stretching the vocal folds. Contraction of the CT muscles enlarges the distance between the thyroid and the cricoid cartilages which lengthens and tenses the vocal folds (Borden and Harris, 1984). This causes an increase in phonation frequency and also introduces fluctuations in phonatory quality due to small variations in the vocal fold movement (Laver, 1980).



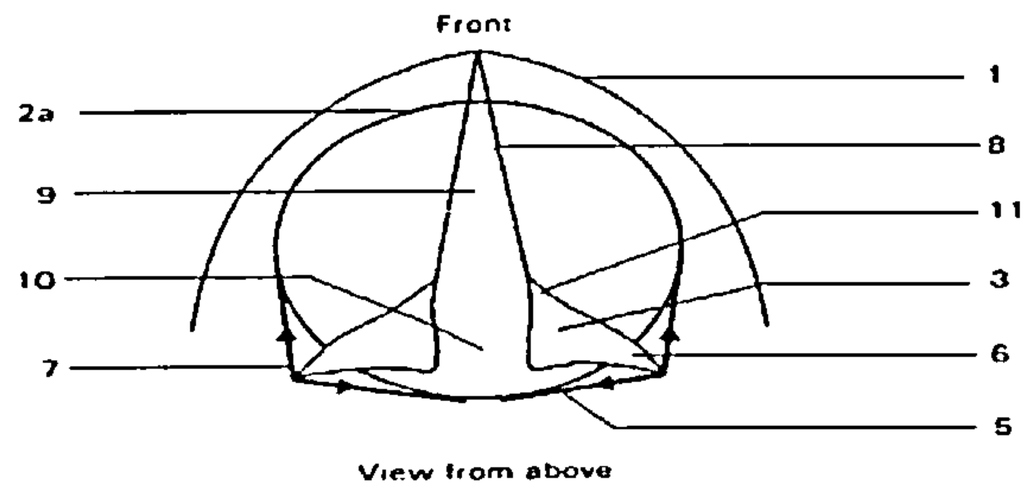


Figure 2.15. Schematic diagram of the action and location of the laryngeal muscles connecting the arytenoid cartilages to each other and to the cricoid cartilage, and related organs (from Laver, 1980).

1. Thyroid cartilage
2. Cricoid cartilage
2a. External edge of cricoid
3. Arytenoid cartilage
4. Interarytenoid muscles
4a. Transverse arytenoid muscle
4b. Oblique arytenoid muscle
5. Posterior cricoarytenoid muscle
6. Muscular process of arytenoid
7. Lateral cricoarytenoid muscle
8. Vocal ligament
9. Ligamental glottis'
10. Cartilaginous glottis
11. Vocal process of arytenoid

The paired thyroarytenoid muscles form the true folds and the ventricular folds. The TA muscles are divided into an upper portion and a lower portion by a small cavity, known as the laryngeal ventricle (or sinuses of Morgagni).

The lower portion of each TA muscle is called the vocalis muscle. Each vocalis muscle stretches from the vocal process of each arytenoid to the inner surface of the fused anterior angle of the thyroid. The vocal folds consist of the vocalis muscles, which are connected to the vocal ligaments which form the innermost glottal edges of the vocal folds, and mucous membrane. Contraction of the vocalis muscle creates a longitudinal tension of the vocal folds resulting in a shortening in their length.

The upper portion of each TA muscle is connected to the upper part of the arytenoid and are called the ventricular folds (or false folds). The ventricular folds comprise of a few muscle fibres covered in a thick mucous tissue. Contraction of the TA muscles pulls the arytenoids anteriorly, tilting them towards the thyroid.

The Muscles determining Abduction and Adduction

This second category of muscles positions the vocal folds by moving the arytenoid relative to the cricoid thus adducting (closing) or abducting (opening) the glottis. These muscles arise from the muscular process, the larger extension at the base of each arytenoid.

Abduction of the vocal folds is required in inhalation and also in producing voiceless consonants in speech. It is achieved by contraction of the posterior cricoarytenoid muscles (PCA). They originate on the posterior surface of the back wall of the cricoid and insert into the top of the muscular process of each arytenoid. Upon contraction they rotate the arytenoids by pulling the muscular processes downwards and backwards. This causes the vocal processes to move outwards, and consequently, the vocal folds separate at the back in a V-shape. This is the natural resting position of the vocal folds.

In order to produce voiced sounds the vocal folds must be set into vibration by bringing them together. This process is called adduction and is achieved by the interarytenoid muscles (IA), and the lateral cricoarytenoid muscles (LCA).

The IA muscles consist of the transverse muscle and the paired oblique arytenoid muscles which cross over it diagonally in both directions. They tilt the top of the arytenoids closer together with the vocal processes rotated inwards. This brings the vocal folds together along their length.

The LCA muscles are attached to the muscular processes of the arytenoid at one end, and the top outer surface of the cricoid at the other. Their contraction rocks the muscular process anteriorly and downwards which adducts the vocal folds. The LCA muscles directly oppose the action of the PCA muscles.

The Structure of the Vocal Folds

The human vocal fold consists of three tissues:

1. vocal ligaments
2. vocalis muscle
3. mucous membrane.

A three layer structural model of the human vocal fold based on tissue examination is presented schematically in figure 2.16. Each layer has a different structural property, and hence, has a different mode of vibration. The transition portion can be considered as part of the body. The three layer model can then be simplified further to a two layer cover-body model (Sawashima and Hirose, 1983).

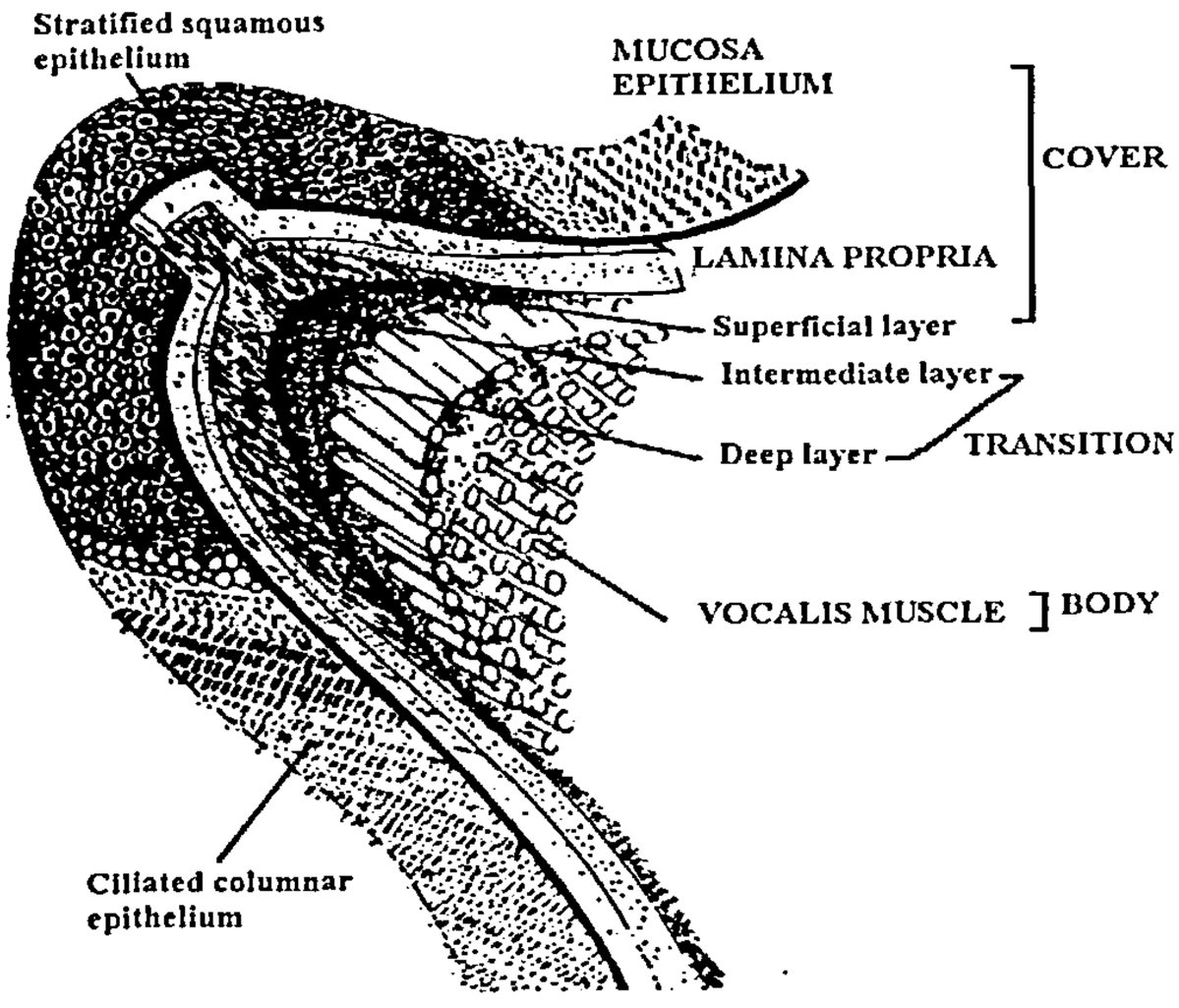


Figure 2.16. Schematic presentation of the structure of the human vocal fold (from Sawashima & Hirose, 1983).

The pattern of vibration differs vertically and longitudinally along the length of the vocal fold. This is due to the vocal fold having a posterior cartilaginous portion where the vocal process of the arytenoid to which the vocal ligament and vocalis muscle is attached, protrudes about one third along the length of the vocal fold. Consequently this cartilaginous portion is stiffer than the anterior flexible membranous portion. Borden and Harris (1984) describe the relaxed vocal folds as being thick which "open and close in an undulating manner, the mucous membrane moving somewhat independently like flabby skin on a waving arm."

The main basic features of laryngeal adjustments are (Sawashima & Hirose, 1983):

"1. abduction-adduction of the vocal folds
2. constriction of the false folds and other supraglottic laryngeal structures
3. changes in the length and thickness of the vocal fold
4. up and down movements of the larynx." (Sawashima & Hirose, 1983).

2.2.2.2 Vocal Fold Vibration

The two aerodynamic forces which produce vibration of the vocal folds are the subglottal air pressure applied to the lower part of the folds, forcing them open, and the negative pressure which occurs as air passes between the folds, due to the Bernoulli effect. These positive and negative pressures set the vocal folds into vibration due to the elasticity of the folds.

The vocal folds are set into vibration when the airstream from the lungs is forced past. This results in the rapid opening and closing of the glottis, the air passage between the vocal folds, which chops up the airstream into tiny pulses.

Each vocal fold closure results in an acoustic excitation/pressure pulse set up at the glottis which is transmitted via the vocal tract. A series of such pulses, produced by periodic vocal fold closures result in a buzz-like voice source. The tensions within the vocal folds and arising from the positioning of the arytenoids varies the mode of vibration, the vibration frequency, and the spectral components of the voice-source waveform.

Voicing relies on three principles. One is that the air pressure below the folds (subglottal air pressure) must exceed the supraglottal air pressure in order to force them apart. The second is the Bernoulli effect. The vocal folds are quickly sucked together as the air pressure drops against the edges of the vocal folds due to air being forced past the glottis. This accounts for the closing phase of each vibratory cycle. The third principle is that the vocal folds are elastic which allows them to be blown open and also permits them to recoil (the elastic recoil force). The above account is known as the traditional myo-elastic theory of vocal fold vibration. In normal male chest voice the vocal folds peel apart slowly from the bottom upwards in a wave-like motion. There is a vertical phase difference as the bottom part closes while the top part of the vocal folds opens (Titze, 1989).

2.2.3 The Supraglottal System

Also known as the supralaryngeal vocal tract, the supraglottal system consists of the various air passages from the glottis to the lips. Figure 2.17a shows the relative positions of the structures of the vocal tract.

It includes the pharynx, the mouth and nasal cavities, and various articulators such as the tongue, teeth, the velum, and the lips. The vocal tract can be modelled as a tube with a variable cross-sectional area. Articulatory movement results in the creation of various cavities, each with their own resonance mode, called formants, within the tube according to which articulator is moved and the degree of movement. The vocal tract filters the voiced sound generated by vocal fold vibration, the voice source, shown in figure 2.17b.

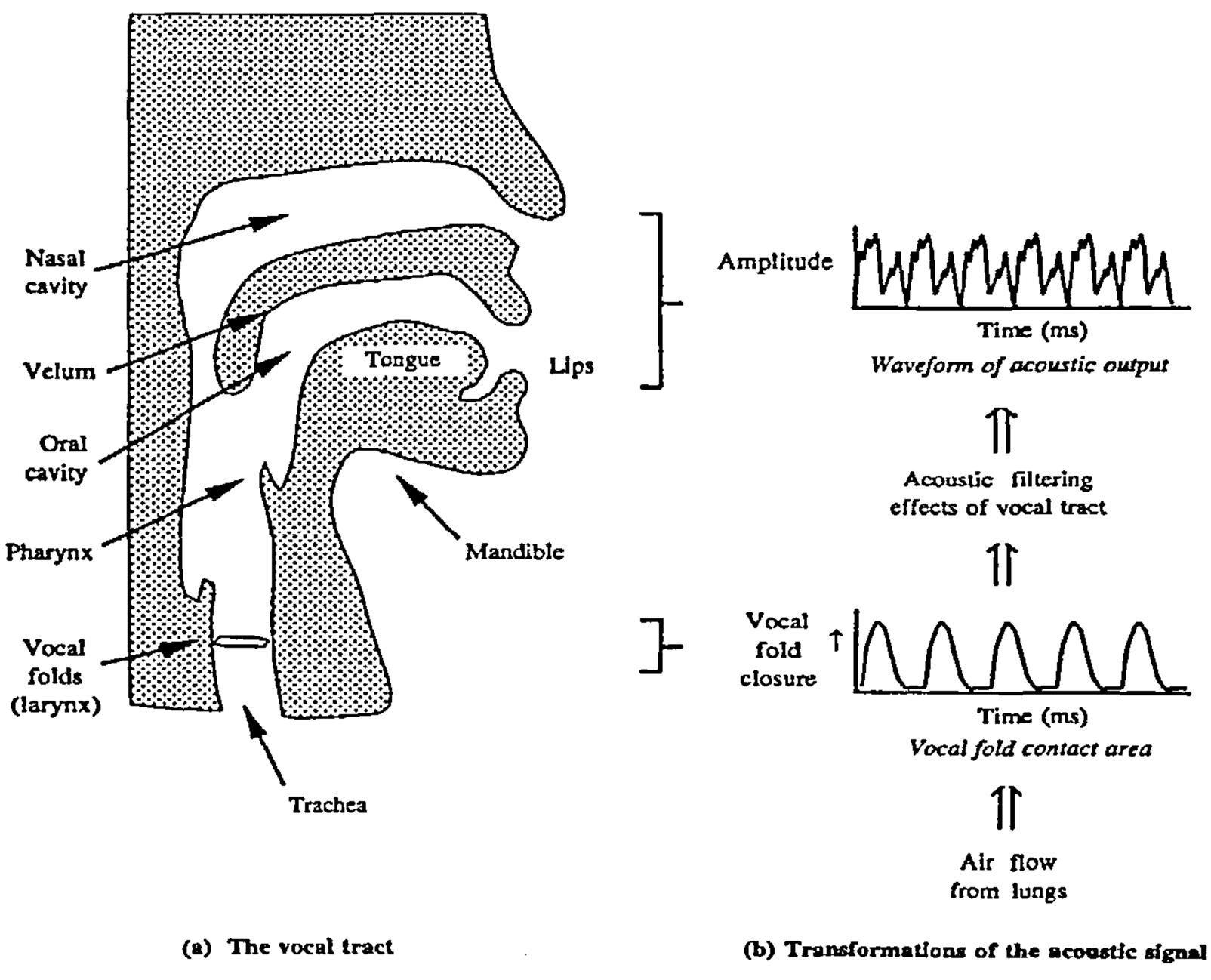


Figure 2.17. The acoustic production of a voiced sound (from Rossiter, 1993, personal correspondence).

The vocal tract not only acts as a variable resonator, but also serves as a sound source for unvoiced sounds (aperiodic sounds) and combined voiced and unvoiced sounds. Table 2.1 presents some examples of these.

Articulation is responsible for creating separate vowel sounds. The lower the formant frequency, the more that frequency depends on articulatory factors. The most important formants for determining vowel quality are the first and the second formants, which have ranges of about 250-900 Hz and 800-2200 Hz respectively. These differences appear to be similar across languages. In speech, nonarticulatory factors such as pharynx length and larynx tube size tend to be responsible for

formant frequencies the higher in frequency they are. For example, the fourth formant frequency is highly dependent on the dimensions of the larynx tube, which is independent of vowel articulation.

Speech sound sources

Source	Resonator	Sound	Manner	Examples
Vocal tract	Vocal tract	Periodic	Vowels	/i/ /u/
			Diphthongs	/ai/ /ou/
			Semivowels	/w/ /y/
			Nasals	/m/ /ŋ/
Vocal tract	Vocal tract	Aperiodic	Stops	/p/ /k/
			Fricatives	/s/ /f/
			Affricate	/tʃ/
Vocal folds and vocal tract	Vocal tract	Mixed periodic and aperiodic	Voiced stops	/b/ /g/
			Voiced fricatives	/z/ /v/
			Voiced Affricate	/dʒ/

Table 2.1. Various speech sounds and their origin (from Borden & Harris, 1984).

2.2.4 The Acoustics of the Vocal Tract

This section describes the acoustic production of spoken vowels in relation to the source-filter theory (Fant, 1960), and the modelling of the vocal tract as a tube resonator.

2.2.4.1 Speech Production

A speech sound can be recorded in terms of a sound pressure waveform in the time domain or a sound pressure spectrum in the frequency domain. The sound pressure spectrum, or "spectral envelope" consists of the transfer function of the vocal tract plus source and radiation characteristics. Resonances within the vocal tract are shown as poles in the transfer function, and anti-resonances are shown as zeros. A pole or zero is characterized by its bandwidth and its centre frequency.

A non-nasal vowel is the simplest model to begin with since it can be characterized by just poles. Many other sound classes can be described in terms of modifications to the vowel theory described below. The following passages are based primarily on Kent & Read (1992). This book also contains a review of these other sound classes which are not described here.

2.2.4.2 Vowels

Vowels are sounds which are spectrally shaped according to the resonance properties of the vocal tract. They are termed voiced when excited by vocal fold vibration. The resonances of the vocal tract are termed "formants". The vocal tract can be thought of as a frequency-variable filter, since altering the shape of the vocal tract will result in formant frequency shifts, heard as a timbral modification in vowel quality.

2.2.4.3 The Source-Filter Theory

In speech literature, the voicing source and the vocal tract have traditionally been modelled as being independent. Vibration frequency and vibration amplitude of the vocal folds does not appear to appreciably affect the resonance characteristics of the vocal tract. This apparent independancy of source and vocal tract is the basis of an important theory, the source-filter theory, which states that the output energy is a product of the source energy and the resonator response (Fant, 1960).

Vibrating vocal folds produce a sound spectrum, called the laryngeal, or voice source spectrum, where the energy lies at integer multiples of the fundamental frequency, typically falling at a rate of approximately 12 dB per octave with increasing frequency for speech.

The vocal tract resonator has an infinite number of formants associated with it. A formant is categorised by its centre frequency (or formant frequency) and its bandwidth (the width of the energy band, or the frequency range within the band). All the formants together give the transfer function of the vocal tract.

As sound energy escapes the mouth it undergoes a sound radiation. The filtering effect termed the "radiation characteristic" behaves as a high-pass filter, dampening more low frequency energy than high frequency energy, and is modelled with a +6 dB per octave slope. The resulting drop in energy of the ouput speech signal when the 12 dB per octave roll off in the laryngeal spectrum is combined with the 6 dB per octave increase in the radiation characteristic is at a rate of -6 dB per octave. This is shown in figure 2.18.

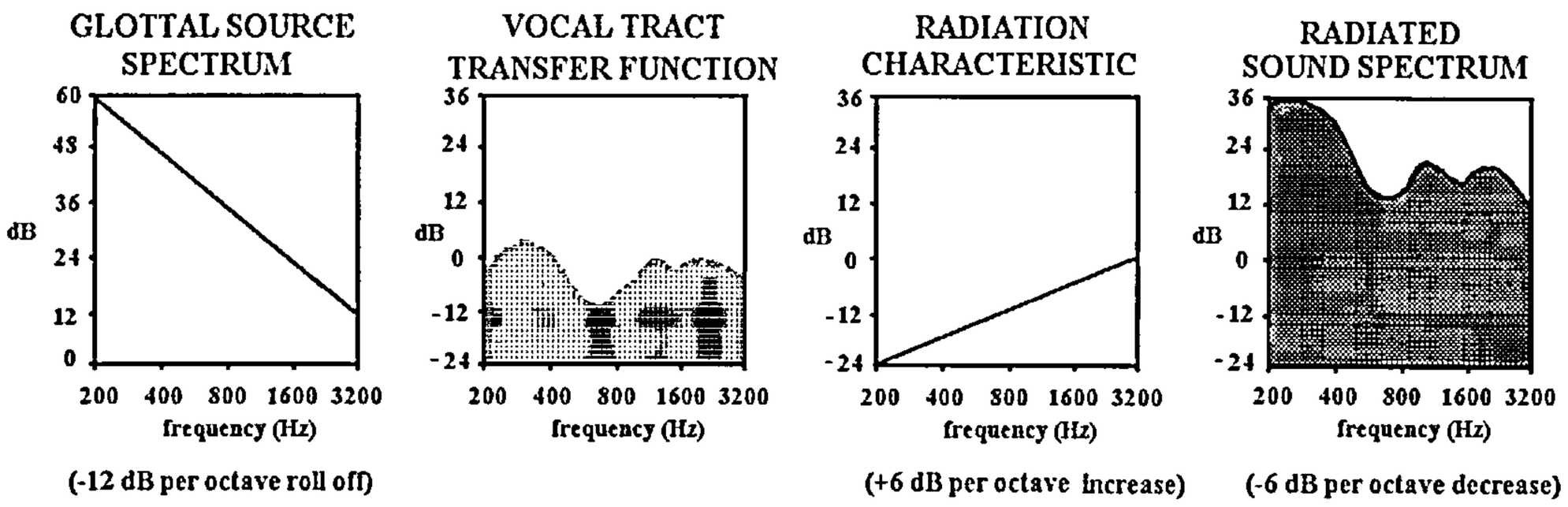


Figure 2.18. Idealized diagram of the source-filter model for speech production.

The source-filter theory of vowel production states that "the radiated sound pressure waveform of speech is the product of the laryngeal spectrum, the vocal tract transfer function, and the radiation characteristic" (Kent & Read, 1992). This is summarised in the following equation as:

$$P(f) = U(f).\ T(f).\ R(f)$$

P(f) is the radiated sound pressure spectrum of speech

P is pressure and (f) indicates it is a function of frequency

U(f) represents the laryngeal source spectrum as volume velocity

T(f) is the transfer function.

R(f) is the radiation characteristic.

U(f) and R(f) can be taken as being constant. Different vowels can then be described in terms of just the transfer function and the radiated spectrum (Kent & Read, 1992).

2.4.4.4 The Tube Resonator Model

The resonance frequencies of a tube resonator are determined by two factors; its length and its cross-sectional area as a function of its length. A single uniform tube can produce two types of resonance depending on whether the ends of the tube are the same or are different. If the ends are both open or both closed, the first resonance occurs for a tone with a wavelength double the tube length, and all the higher resonance frequencies are integer multiples of the first resonance.

The relaxed vocal tract can be modelled as a tube which has one open end and one closed end. For this model, the first resonance frequency has a quarter wavelength (i.e., occurring for a tone with a wavelength four times the tube length) and higher resonances are odd multiples of the first resonance. This relationship can be expressed using the odd-quarter wavelength formula:

$$\mathbf{Fn = (2n - 1)\ c/4L}$$

n is an integer

c is the speed of sound (34400 cm/sec),

L is the length of tube.

The equation shows that a sound with a wavelength 4 times (or odd multiples of 4) the length of the tube will resonate with the maximum amplitude. The higher resonances are then located at c/4L, 3c/4L, 5c/4L, 7c/4L, etc. For example, if a tube has a length (L) of 17.5 cm, the average size of a relaxed male vocal tract, the first resonance frequency (F1) can be calculated according to:

F1 = c/4L = 35,000 cm/s / (4*17.5 cm) = 500 1/s, or 500 Hz

From this, F2 = 1500 Hz, F3 = 2500 Hz, F4 = 3500 Hz, F5 = 4500 Hz, etc.; that is, the formants for a relaxed male vocal tract fall at about 1 kHz intervals. Formants are identified by their number which increases with frequency.

The relaxed vocal tract relates to the articulation of the vowel known in phonetics as a schwa. The air pressure relationship between the first three formants of a schwa is shown in figure 2.19.

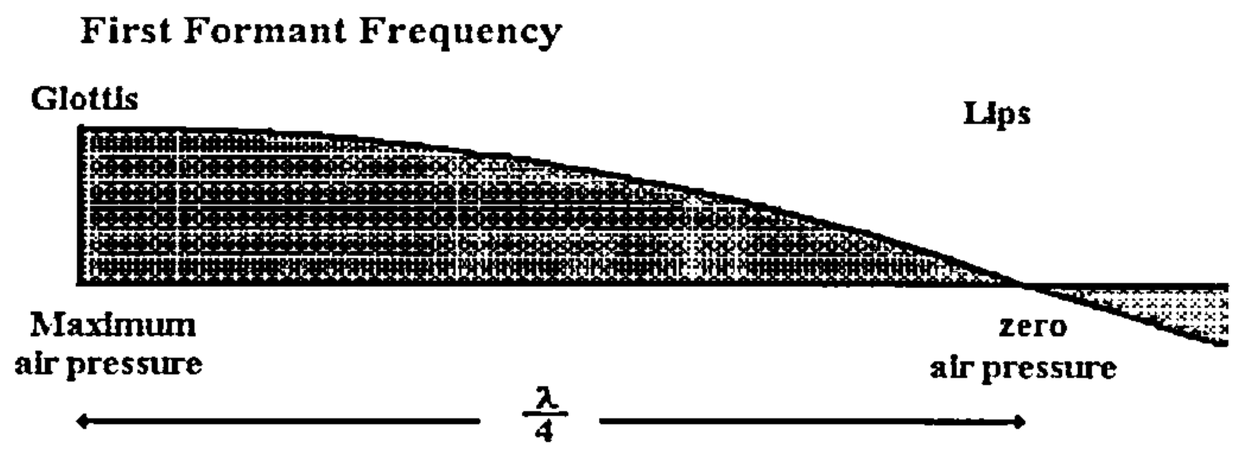


Figure 2.19. The sinusoidal relationship of air pressure at the first three formant frequencies (after Lieberman, 1977).

Overpressure from the lungs forces the vocal folds open and upwards. The speed of the air particles increases as they flow through the glottis, whilst the air pressure decreases. The decrease in pressure between the folds causes them to snap shut. At the instance of closure the pressure behind the folds suddenly increases again, forcing the folds open once more. Single pressure pulses are emitted which travel up and down the vocal tract. These pulses can be thought of as individual waves of compression and rarefaction.

Sound waves are changes in pressure over time. When air molecules are pushed together they spring back to their original position due to their inherent elasticity. This sets up a disturbance which is transferred through the medium as a series of alternating compression (high pressure) and rarefaction (low pressure) waves. It is this disturbance which travels as the sound wave.

When two waves meet they superimpose, resulting in an acoustic amplitude change in that location which is determined by the direction of travel of the waves and their individual amplitudes. When two waves moving in the same direction meet, their molecular movements combine to produce a wave of increased amplitude, i.e., an increase in molecular vibration. This is known as resonance. The converse holds true also; the amplitudes of two similar waves moving in opposite directions may even cancel each other out, resulting in no molecular movement. This occurs if the wavelength of the sound and the length of the tube obey a fixed ratio, and results in the formation of standing waves. In standing waves, the locations of minimum molecular vibration are called nodes, and the locations of maximum molecular vibration are called antinodes. In a tube each formant has its own characteristic standing wave pattern consisting of nodes and antinodes which are determined by the superimposition of incidence (incoming) and reflected waves.

Curvature of the tube does not appreciably alter the resonant frequencies of the model. This is why it is convenient to model the vocal tract as a two-dimensional straight tube with a variable cross-sectional area, assuming that the vocal tract approximates a cavity which is circular along its length. Each vowel will then have associated with it a different shaped tube which generates a set of formant frequencies which differ from the schwa pattern.

Combining some 2-D models of different vowels with their associated spectra reveals a relationship between tongue height and the first two formant frequencies. The frequency of F1 is inversely related to tongue height, and the frequency of F2 is related to tongue advancement, as seen graphically in figure 2.20 (Kent & Read, 1992). Lip rounding increases the length of the vocal tract which lowers the formant frequencies. In English, lip rounding occurs in some middle and back vowels such as those in who, hoe, and her, but does not occur in front vowels.

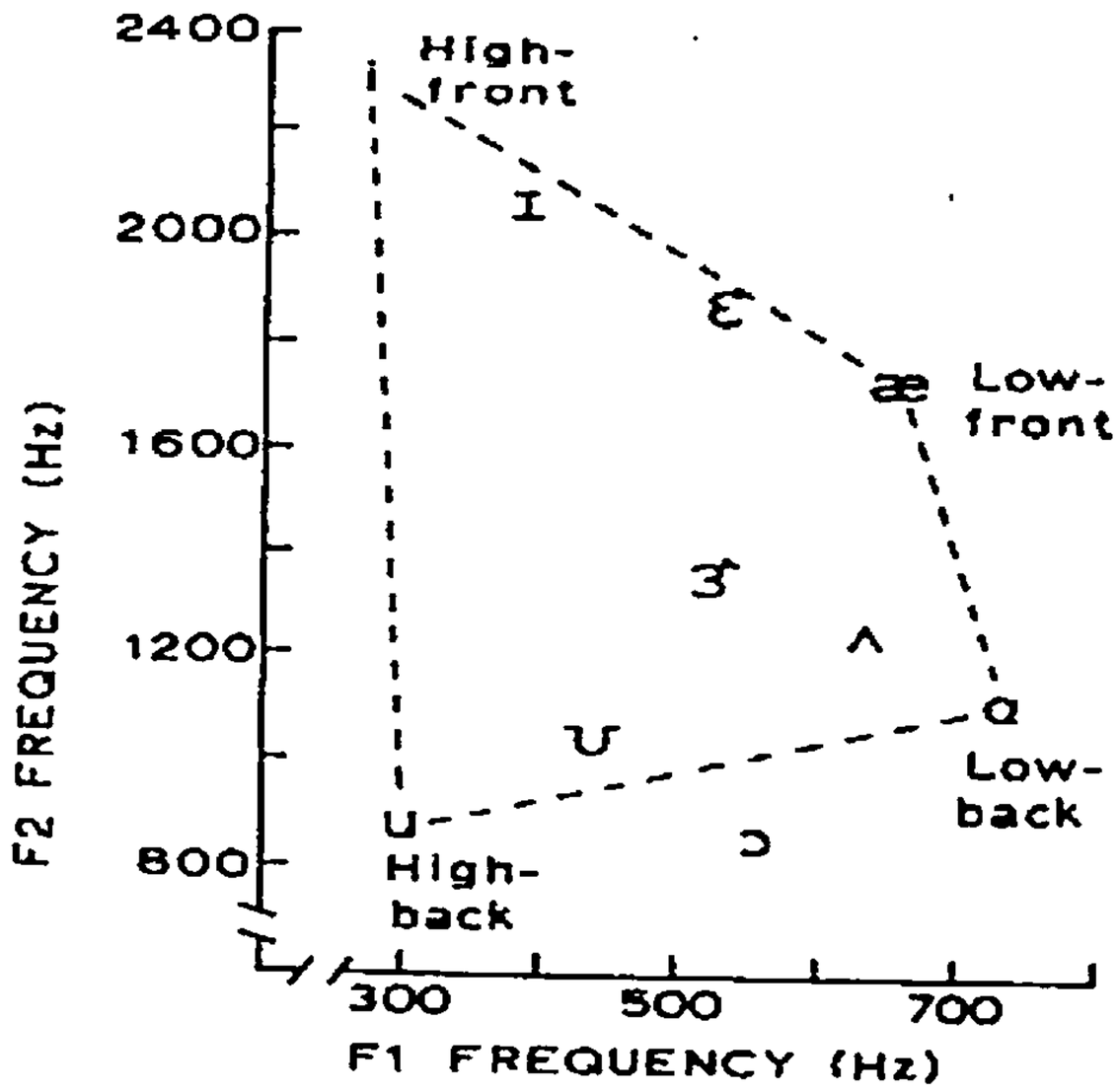


Figure 2.20. Classic F1-F2 chart in which a vowel for an average adult male is represented acoustically by its F1 and F2 frequencies. The phonemic symbols are positioned to show the F1 and F2 values for that vowel. An articulatory-acoustic relationship is suggested by the labels in the figure. (from Kent & Read, 1992).

2.3 The Human Hearing System

This section describes how many perceptual phenomena are determined by the physiological functions of the hearing system. It begins with a description of the physiology of the hearing system which is largely based on Campbell & Greated (1987).

2.3.1 Hearing Physiology

Figure 2.21 is a simplified diagram of the main features of the ear. The main task of the ear is to transform incoming air pressure fluctuations into electrical signals which the brain can then process.

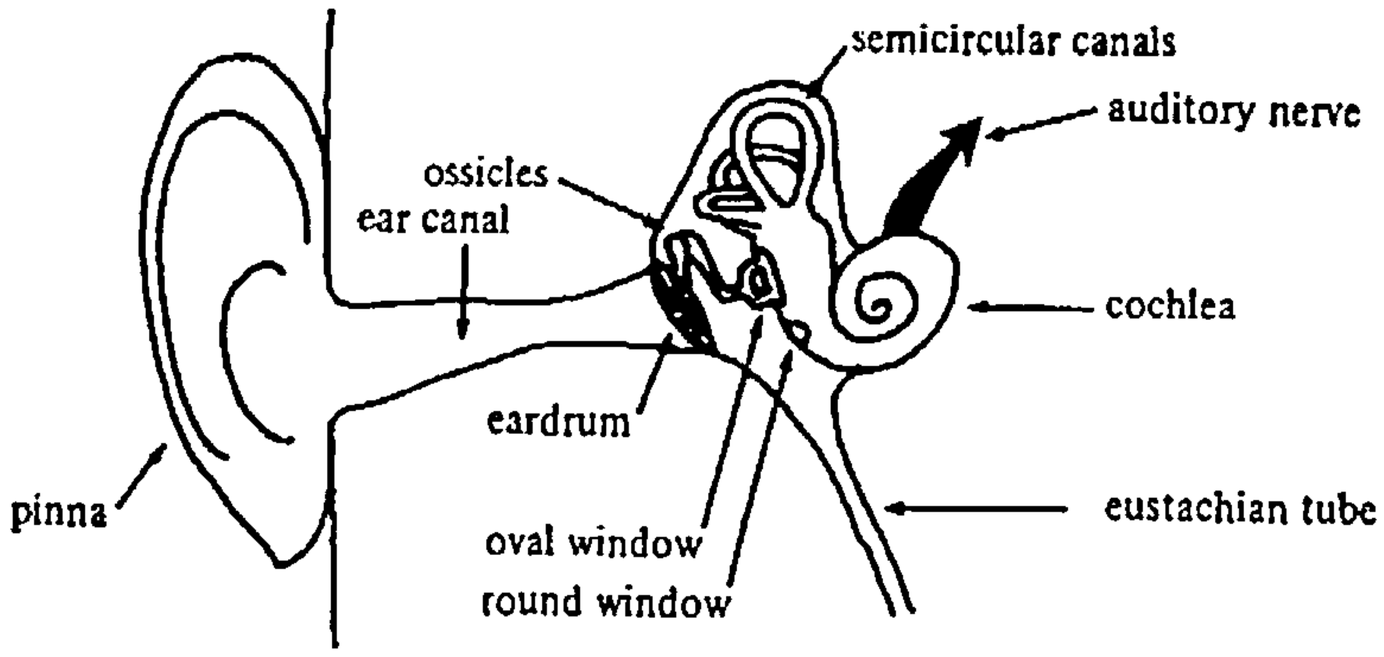


Figure 2.21. Anatomical sketch of the human ear (from Campbell & Greated, 1987).

The ear can be divided into three sections - the outer, the middle, and the inner ear. Each of these sections are shown in figure 2.22.

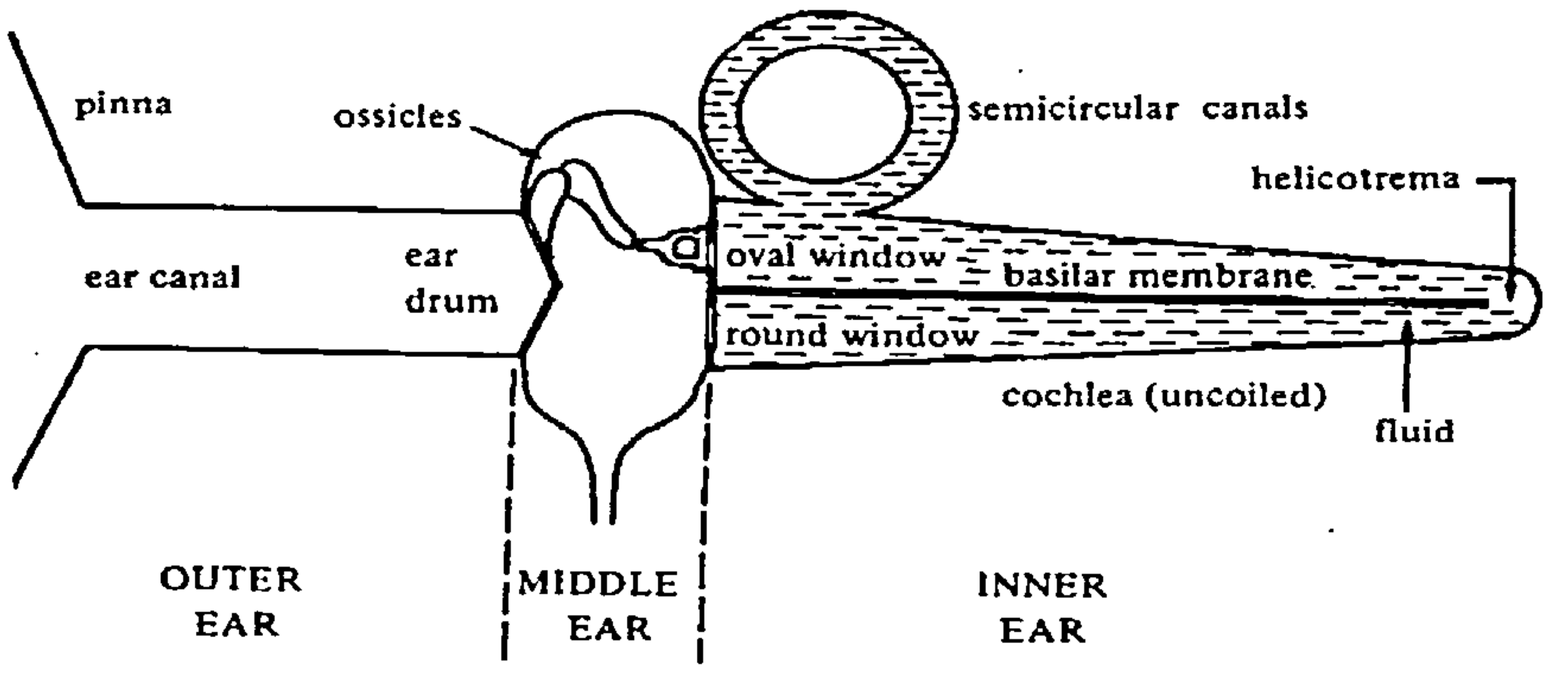


Figure 2.22. Schematic diagram of the human ear (from Campbell & Greated, 1987).

2.3.1.1 The Outer Ear

The outer ear comprises of the pinna (the prominent visible flaps, one on each side of the head), leading to the auditory canal (auditory meatus) and ending at the eardrum (tympanic membrane), a membrane at the entrance to the middle ear. The pinna's task is to funnel short-wavelength sound into the canal and towards the eardrum, and also provides some high frequency selectivity between anterior and posterior located sounds. Having one ear on either side of the head provides us with binaural hearing enabling sound localization also on the horizontal plane.

2.3.1.2 The Middle Ear

The middle ear comprises of an air-filled cavity in the skull bone, extending from the eardrum on the outside, to the oval window (fenestra ovalis) and round window (fenestra rotunda) on the inner side, which are two small holes in the bone marking the entrance to the inner ear.

The outer ear is linked to the inner ear through a system of levers which operate within this cavity. The lever system is made up of three small connecting bones known as ossicles. They are individually titled, in turn, the hammer (malleus), which is connected to the inner part of the eardrum, the anvil (incus), and the stirrup (stapes), which acts like a piston, the footplate of which moves in and out of the oval window (Rhode, 1978). Both the oval window and round window are covered with a thin membrane.

A sound wave consists of air pressure fluctuations. When these air pressure fluctuations hit the eardrum, they set it vibrating. The main function of the lever system is to transmit these vibrations to the oval window as efficiently as possible, that is, with minimal energy loss, shown in figure 2.23.

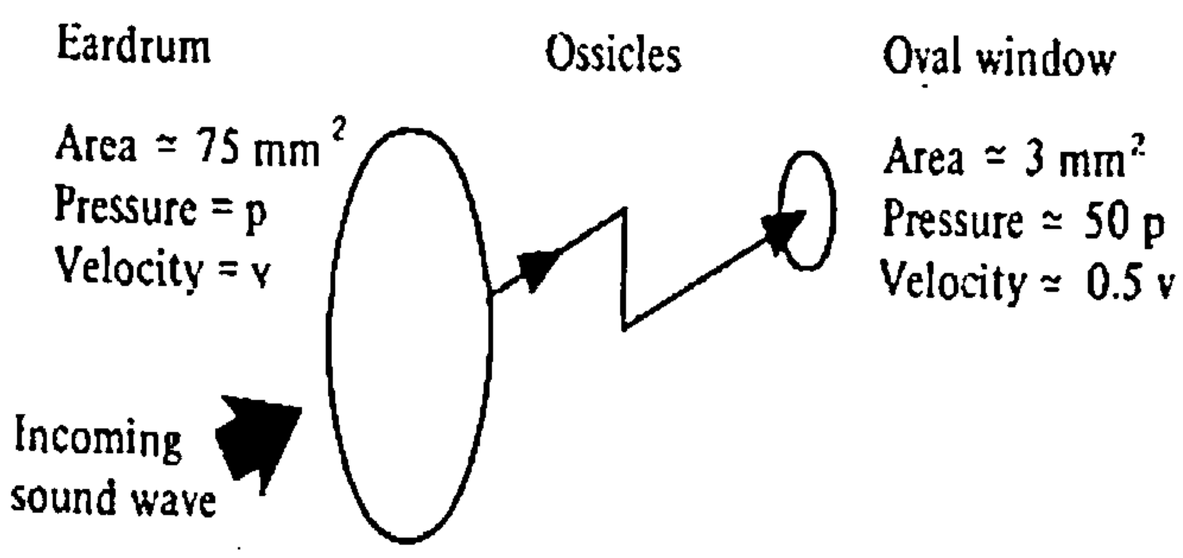


Figure 2.23. Changes in pressure and velocity between eardrum and oval window, due to reduction in area and lever action of ossicles (from Campbell & Greated, 1987).

About 50% of energy is transmitted to the inner ear, as opposed to 1% if the middle ear mechanism was absent. This is due to a massive increase in pressure across the middle ear. The pressure exerted on the oval window is approximately 50 times larger than on the eardrum due to a combination of lever action which increases pressure, the relative surface areas of the oval window and eardrum, and the difference in specific acoustic impedance of the materials making up barriers (which is the ratio of pressure amplitude to velocity amplitude). The higher the specific acoustic impedance of a material, the lower the transmission of energy. Air can be thought of as a barrier with perfect sound energy transmission. The eardrum is three times less efficient in transmitting energy than air, but is 100 times more efficient than the oval window (de Boer, 1980). Bone has an even higher specific acoustic impedance. For middle-range frequencies, approximately 50% of the sound energy is reflected back up the canal. The amount of reflected sound is much greater for low-frequency energy below 100 Hz and for high-frequency energy above 10 kHz, due to the mass and stiffness properties of the middle-ear mechanism.

The eustachian tube which runs from the middle ear to the back of the throat regulates air pressure between the eardrum, between the middle ear and the outside atmosphere.

2.3.1.3 The Inner Ear

The inner ear is a fluid-filled cavity in the skull just behind the middle ear. It consists of a complex labyrinth of passages and chambers. It performs two major functions, giving us our sense of balance as well as our sense of hearing. The semi-circular canals and the cochlea respectively are responsible for these two tasks.

The cochlea is a tube about 3.5 cm long which is coiled about 2.5 turns and resembles a snail's shell. The base of the spiral is about 2 mm in diameter, gradually tapering off at the apex.

It is the cochlea's job to convert vibrations from the middle ear into electrical signals which are then transmitted via the auditory nerve to the brain. How the cochlea actually achieves this has been subject to debate in recent times, resting on the battle between temporal and placement theories (see below). However, it is agreed that many musically important parameters of hearing can be attributed to the cochlea and that both theories can account for most, but not all, pitch perception phenomena (Moore, 1989).

A simplified cross-section of the cochlea is shown in figure 2.24. The tube is divided into three parts by two membranes. The Reissner's membrane is extremely flimsy and separates the upper gallery (scala vestibuli) from the cochlear duct (scala media). The basilar membrane is more solid and divides the cochlear duct from the lower gallery (scala tympani). It performs a particular function as a "mechanical frequency analyser" (Moore, 1989), which will be described later. The membranes run most of the length of the cochlea. The helicotrema, a hole at the apex of the spiral is the only place where the upper and lower galleries are connected.

Covering the upper surface of the basilar membrane is a blanket of tiny hair cells, collectively known as the organ of Corti. Approximately 30,000 nerve fibres carry the electrical signals from the organ of Corti to the brain.

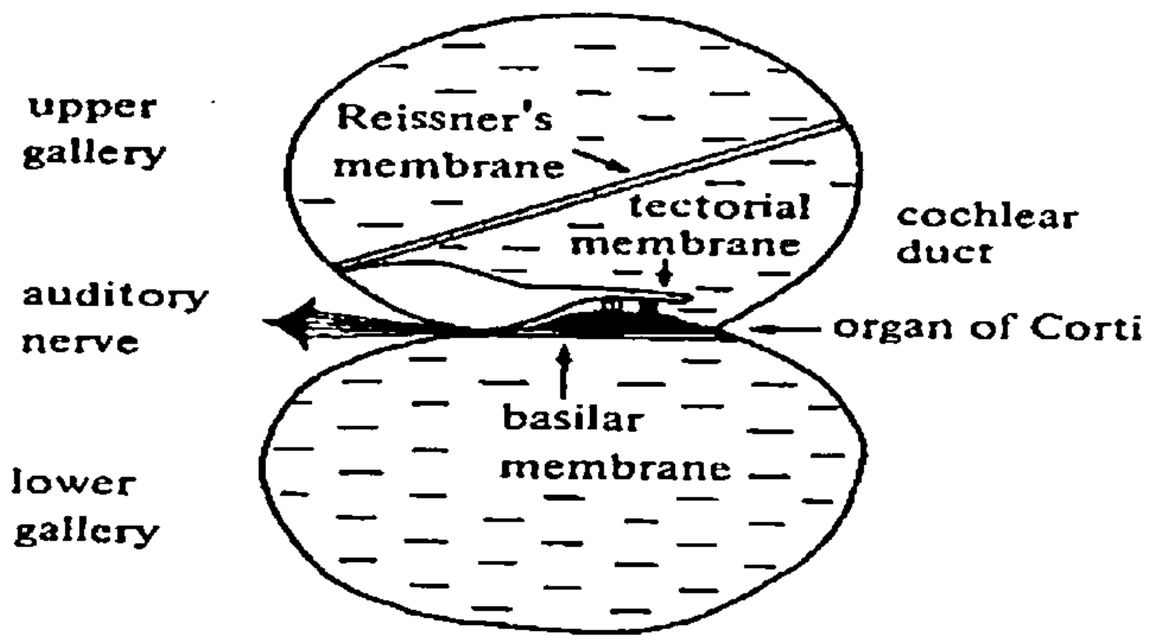


Figure 2.24. Cross-section of cochlea (from Campbell & Greated, 1987).

2.3.1.4 Basilar Membrane Movement

One can think of the behaviour of the peripheral auditory system as if it were a bank of bandpass filters with continuously overlapping centre frequencies (Fletcher, 1940). These filters are called "auditory filters". Experiments suggest that these filters have their origin on the basilar membrane, and each of these filters responds to a small range of frequencies (Fletcher, 1940; Moore, 1986).

Figure 2.25 shows the action of the basilar membrane and round window as a pressure pulse arrives at the oval window.

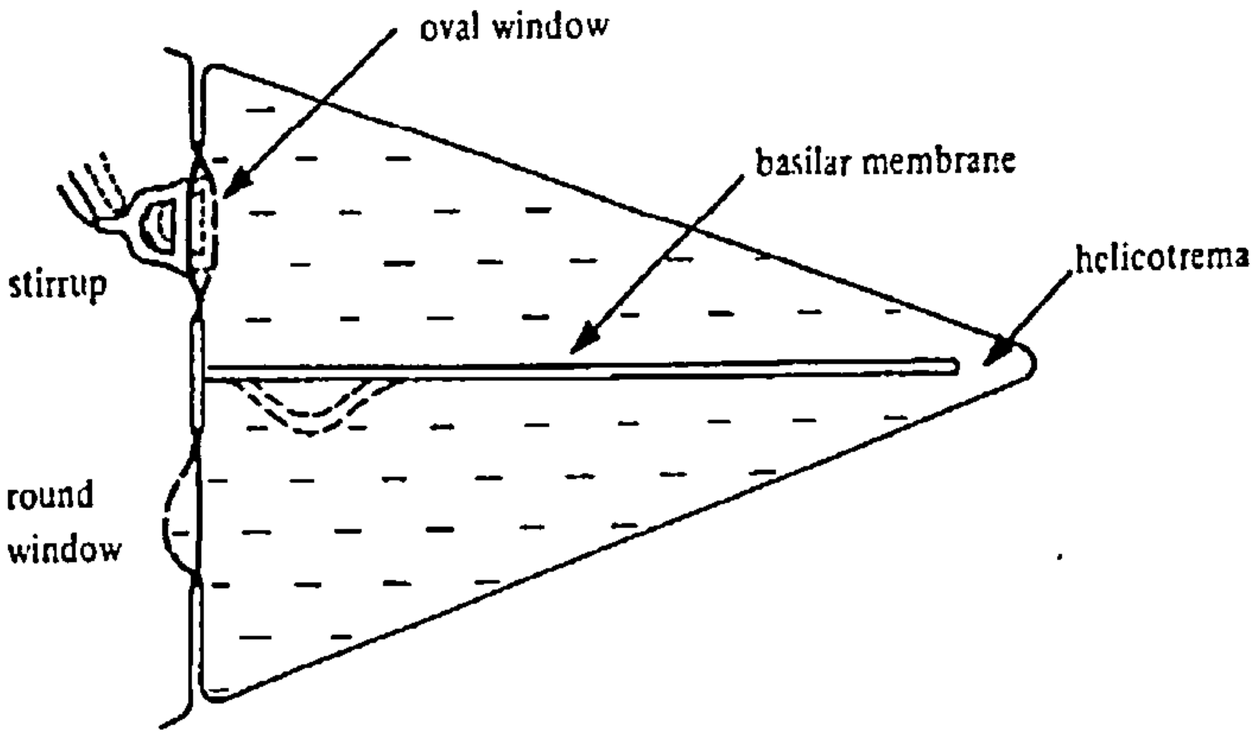


Figure 2.25. Arrival of a pressure pulse at the oval window (from Campbell & Greated, 1987).

A continuous pure tone played into the eardrum causes it to vibrate periodically with simple harmonic motion. The stirrup footplate moves rhythmically in and out of the oval window causing a series of alternately upward and downward bulges to travel up the basilar membrane from the oval window. These bulges gradually increase in size until they reach a peak amplitude at a particular place along the basilar membrane related to the frequency of the pure tone being heard. This generates impulses along the amplitude envelope, the strongest ones under the amplitude peak which are sent to the brain. The brain recognises the sound signal as having a particular frequency sensation (von Bekesy, 1960). The progress of a wave travelling along the basilar membrane is shown in figure 2.26.

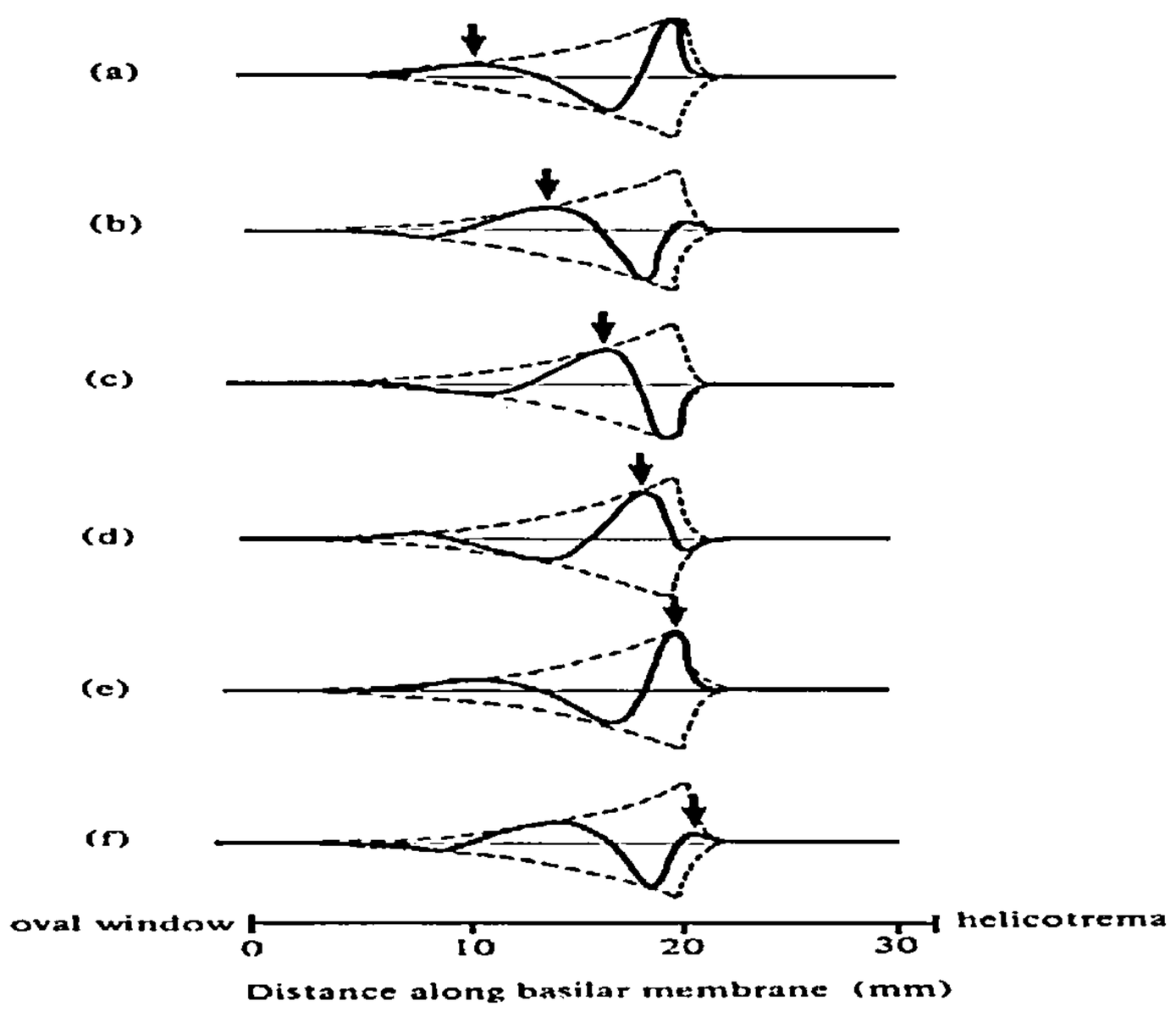


Figure 2.26. Successive cross-sections of the basilar membrane showing the progress of a travelling wave (from Campbell & Greated, 1987).

Higher frequencies travel shorter distances along the basilar membrane (see figure 2.27). This frequency response of the basilar membrane is due to its shape and stiffness. It is stiffer and narrower at the oval window end, gradually flattening and widening out towards the helicotrema end.

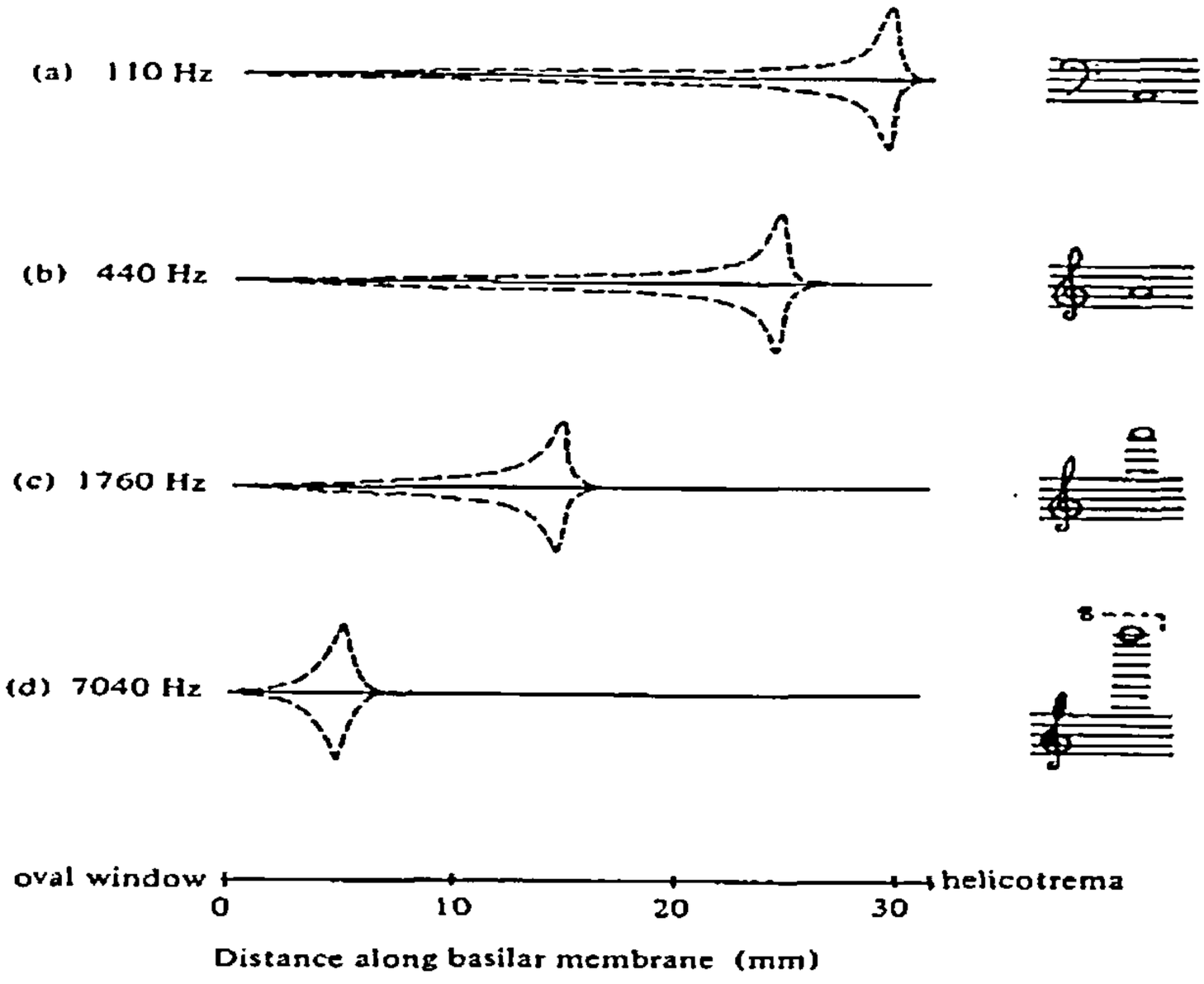


Figure 2.27. Amplitude envelope of basilar membrane vibrations when hearing a pure tone at different frequencies (from Campbell & Greated, 1987).

This theory of frequency discrimination is titled the "place theory", and is illustrated in figure 2.28. It does not, however, totally account for pitch perception in complex tones where individual harmonics are not resolved resulting in a pattern of distribution of displacement. The maximum displacement along the basilar membrane may not correspond to the pitch heard (Moore, 1989).

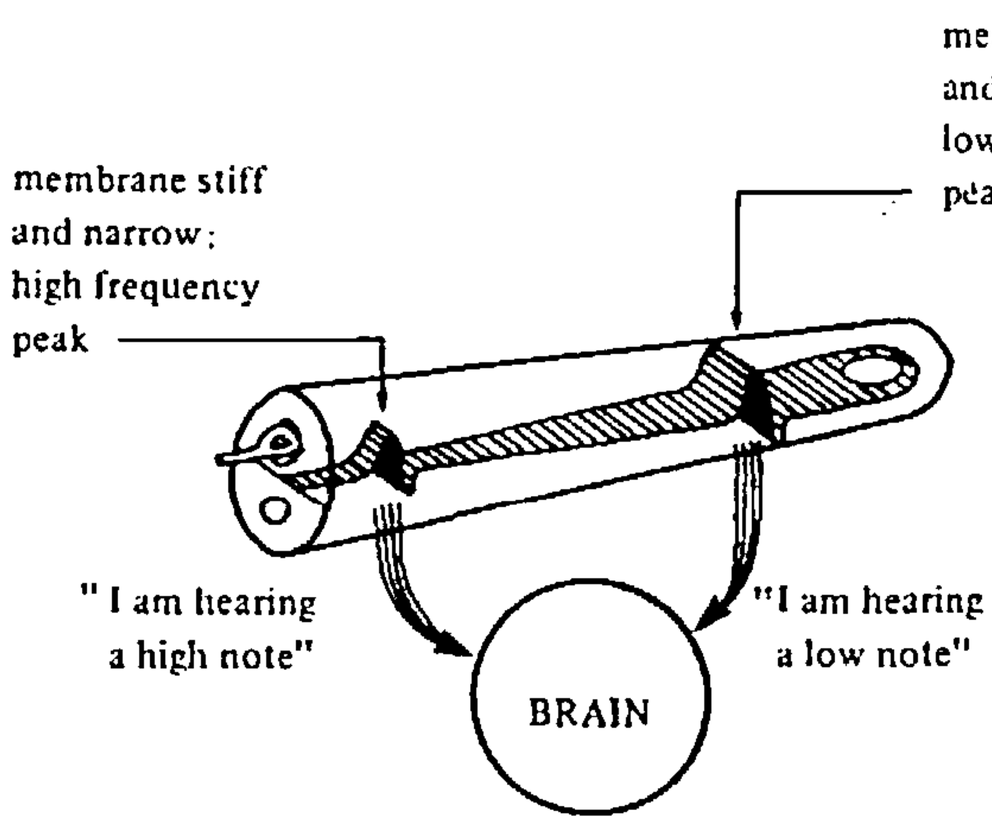


Figure 2.28. Illustration of a 'place theory', in which frequencies are distinguished by the positions of the corresponding amplitude envelope peaks on the basilar membrane of the inner ear (from Campbell & Greated, 1987).

Several studies (Rhode, 1978; Sellick et al., 1982) suggest that the amplitude peak is very sharp with a very steep cut off. This possibly explains why two tones in succession varying very slightly in frequency can be distinguished. The brain may either concentrate on the peak amplitude where the signal is strongest, or on the cut-off, where the position of the signal changes the most (Evans, 1975).

Another theory, called the "volley theory" (Wever, 1949) suggests that frequency recognition is dependent on the timing of nerve signals. Nerve fibres are collected into a large bundle. A "volley" of impulses are sent down this bundle at every peak of the vibration cycle and are combined by the brain. The pitch of the tone is determined by the relative number of fundamental periods of the vibrations. The volley theory is illustrated in figure 2.29.

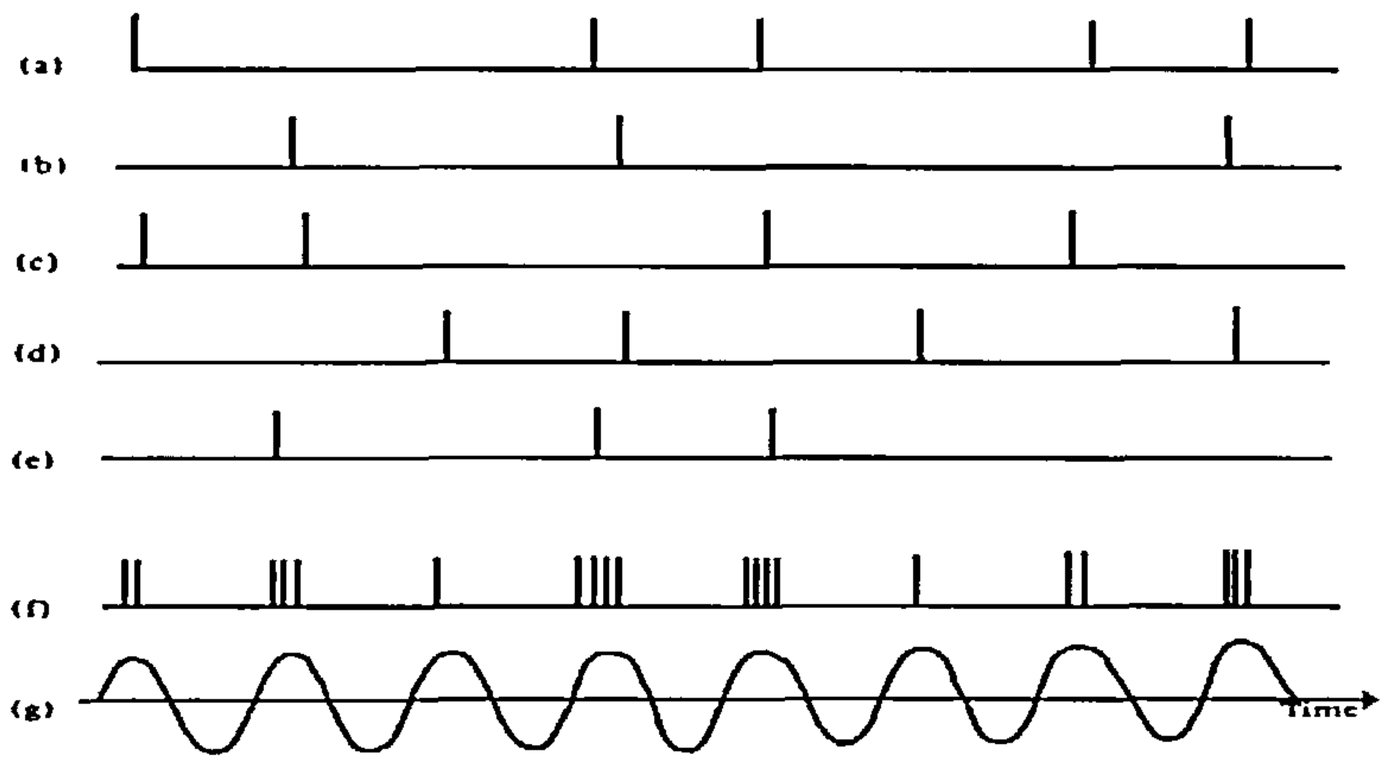


Figure 2.29. (a)-(e) Electrical pulses on five different nerve fibres activated by the pure tone whose vibration curve is shown in (g). The sum of the signals on all five fibres is shown in (f) (from Campbell & Greated, 1987).

2.3.2 Hearing Perception and Psychoacoustics

There are three main attributes of hearing perception: pitch, loudness and timbre. Previous studies have fallen into two categories: pure tone perception which can be studied with some accuracy, but which is ultimately of little consequence to real world sounds; and complex tone perception, which is more relevant to everyday hearing, but which proves difficult to quantify. In order to understand hearing perception as it relates to singing, it is useful to begin by focusing in on those aspects of fundamental hearing perception and physiology which form the basis for discrimination of complex sounds.

Musical perception probably involves rule-governed processes akin to those used in speech perception with the perceptual values of many phenomena being affected by musical training, culture and experience.

Pitch can be considered as the perceptual correlate of frequency, and loudness can be thought of as the perceptual correlate of intensity. Both pitch and loudness values can be thought of as points on a scale ranging from one extreme to another (either low to high as in pitch, or soft to loud as in loudness). These can be quantified using pitch-halving or loudness-halving tests. However, the subjective nature of these attributes makes judgement and quantification difficult, though they do provide some perceptual clues in the understanding musical sound.

2.3.3 Critical Bands

The concept of the "critical band" underlies much of hearing perception. It accounts for many of the phenomena associated with pitch, loudness and timbral discrimination and has its origins in the pattern of movement of the basilar membrane. For example, when two pure tones are heard simultaneously but are of a sufficient interval apart (e.g., an octave), two separate patterns of vibration are generated on the basilar membrane. The amplitude envelopes of these vibrations do not overlap sufficiently to excite many of the same hair cells and the two tones are heard separately with ease. However, reducing the frequency separation increases the number of hair cells which are fired by both signals. The amplitude envelopes of pure tones under a tone apart overlap considerably on the basilar membrane. When there is this strong overlap between the two tones, they are said to fall between one critical band. If the two tones fire essentially two sets of hair cells, they are said to be separated by more than one critical band (Fletcher 1940; Zwicker et al. 1957; Plomp, 1976). The critical bandwidth is about 0.9 mm on the basilar membrane for most of the audible frequency range (Moore, 1986). The critical band is a frequency band which has a varying bandwidth over frequency reflecting our auditory frequency scale, which is linear at low frequencies and logarithmic above about 500 Hz (Terhardt, 1974). For low frequencies the critical bandwidth of hearing is approximately 100 Hz, and at 500 Hz it is about 20 % of the centre frequency (Sundberg, 1987).

2.3.4 Pitch

Pitch is defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale" (American Standards Association, 1960, cited in Moore, 1989). Most psychoacoustical experiments are based on this concept.

Pitch is a subjective phenomenon. The ability to attribute a specific pitch to an acoustic signal is a fundamental property of the human auditory system. (Whitfield, 1980).

Pitch is generally thought of as the perceptual correlate of the physical property of frequency. That is, for most cases, we perceive the pitch of a periodic sound as equalling the pitch of a sinusoid

at that frequency. There are exceptions. Not only does the perceived pitch of a tone vary as a function of frequency, but it is also determined to a lesser extent by other factors such as acoustic background, and duration. Below is a summary of some results from pure tone studies that are relevant to this thesis, followed by some results from complex tone studies. Pitch perception in pure tones:

1. The audible range of frequencies varies from person to person, though generally a young person with normal hearing may hear a lower limit of between 20 to 30 Hz, and an upper limit of 15 to 18 kHz. The upper limit gradually drops as one gets older (Campbell & Greated, 1987).
2. The pitch of a pure tone varies as a function of duration. Below 1000 Hz the shortest duration that a pure tone can be perceived of as having pitch varies as a function of frequency. An audible tone will be perceived of as having pitch if it is of sufficient duration. If it is too short, a click will be heard. If it is a little longer it will be heard as a click with some timbral quality. Two or three cycles of waveform are required to perceive the pitch of a pure tone below 1000 Hz. So, longer durations are necessary for lower frequencies (e.g. 25 ms at 125 Hz, but only 10 ms above 1000 Hz). Above 1000 Hz only a constant minimum duration is required.
3. Some changes in frequency are not perceptible. There needs to be a certain amount of frequency change, called the just noticeable difference (JND), before two pure tones are heard as being different in pitch. The JND is a function of frequency. Figure 2.30 (a) shows that the ability of the ear to discriminate between pitches at low pitches is severely impaired. So, a listener would find it extremely difficult to hear any difference in pitch between a sinusoid B0 followed by a sinusoid C1. However, the listener should not have difficulty discriminating between these two pitches when using complex tones, since the information about the pitches of complex tones at low pitches is contained in the upper harmonics.

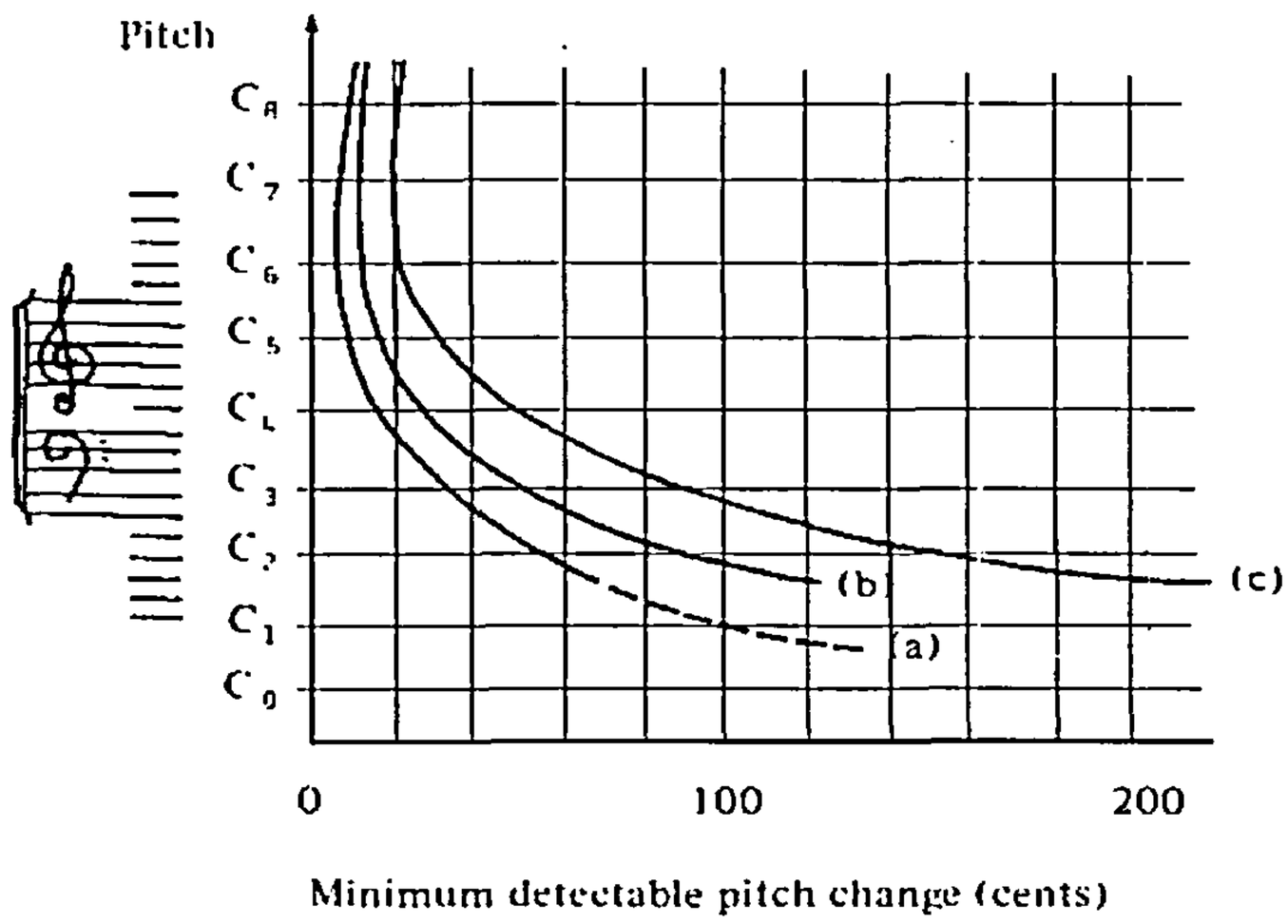


Figure 2.30. The smallest pitch change which can just be detected in a pure tone by the average listener: (a) Abrupt change in loud tone, with SL = 80 dB, (b) Steady fluctuation in loud tone, with SPL = 80 dB, (c) Steady fluctuation in quiet tone, with LL = 30 phons (from Campbell & Greated, 1987).

Pitch Perception in Complex Tones:

It appears that most pitch phenomena are dependent on spectral frequency clues. The main experimental discoveries are summarised below:

1. In a complex musical tone the spectral components lying in the frequency region between about 500 Hz and 2000 Hz, called the "dominance region" are primarily responsible for the pitch of that tone. (Plomp 1967). Figure. 2.31, which summarises these results shows that the 4th and 5th harmonics are responsible for determining the pitch of the complex tone for notes in the bass clef, the 2nd and 3rd harmonics for notes in the top of the treble clef, and the 1st harmonic (the fundamental) only for the upper extreme of the range.

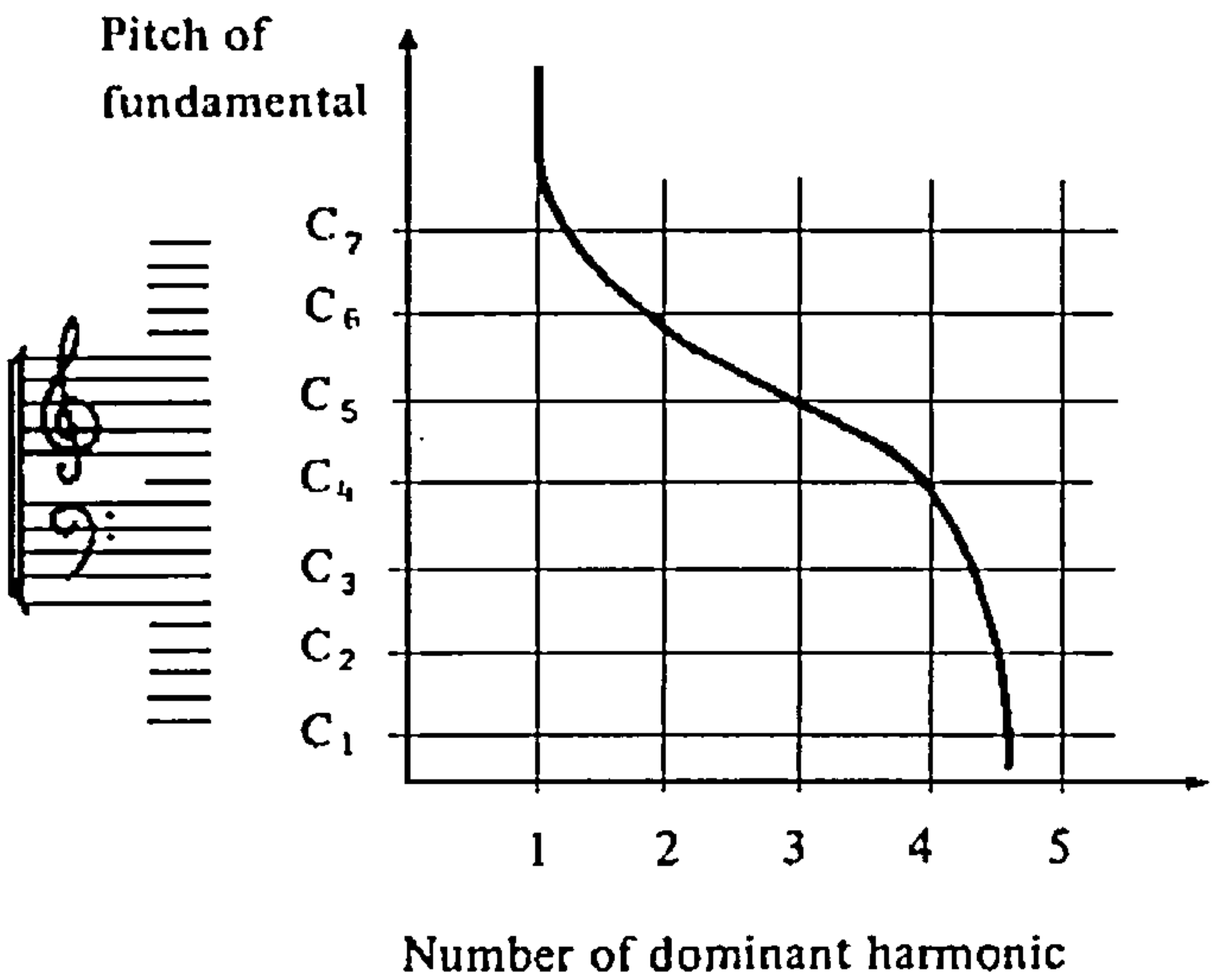


Figure 2.31. The dominant harmonic in the perception of the pitch of a complex musical tone (from Campbell & Greated, 1987).

2. Under certain circumstances the pitch of one sinusoid can be affected by the addition of another at a different frequency (Terhardt and Fastl, 1971). So, for even completely harmonic complex tones the pitch does not necessarily correlate with a sinusoid at the fundamental frequency.

3. Reducing the frequency separation between two tones lying within one critical band increases the sensation of roughness, reaching a peak at about a quarter of a critical bandwidth (Plomp, 1976).

4. Listeners can generally pick out separately the first five to eight harmonics in a complex tone, depending on the critical bandwidths of the components. Critical bandwidths increase with frequency over 500 Hz but the harmonics of a complex tone are spaced equidistantly. This means that high harmonics are separated by less than the critical bandwidth, so the components cannot be discriminated separately. Figure 2.32 shows the critical bandwidths for pure tones.

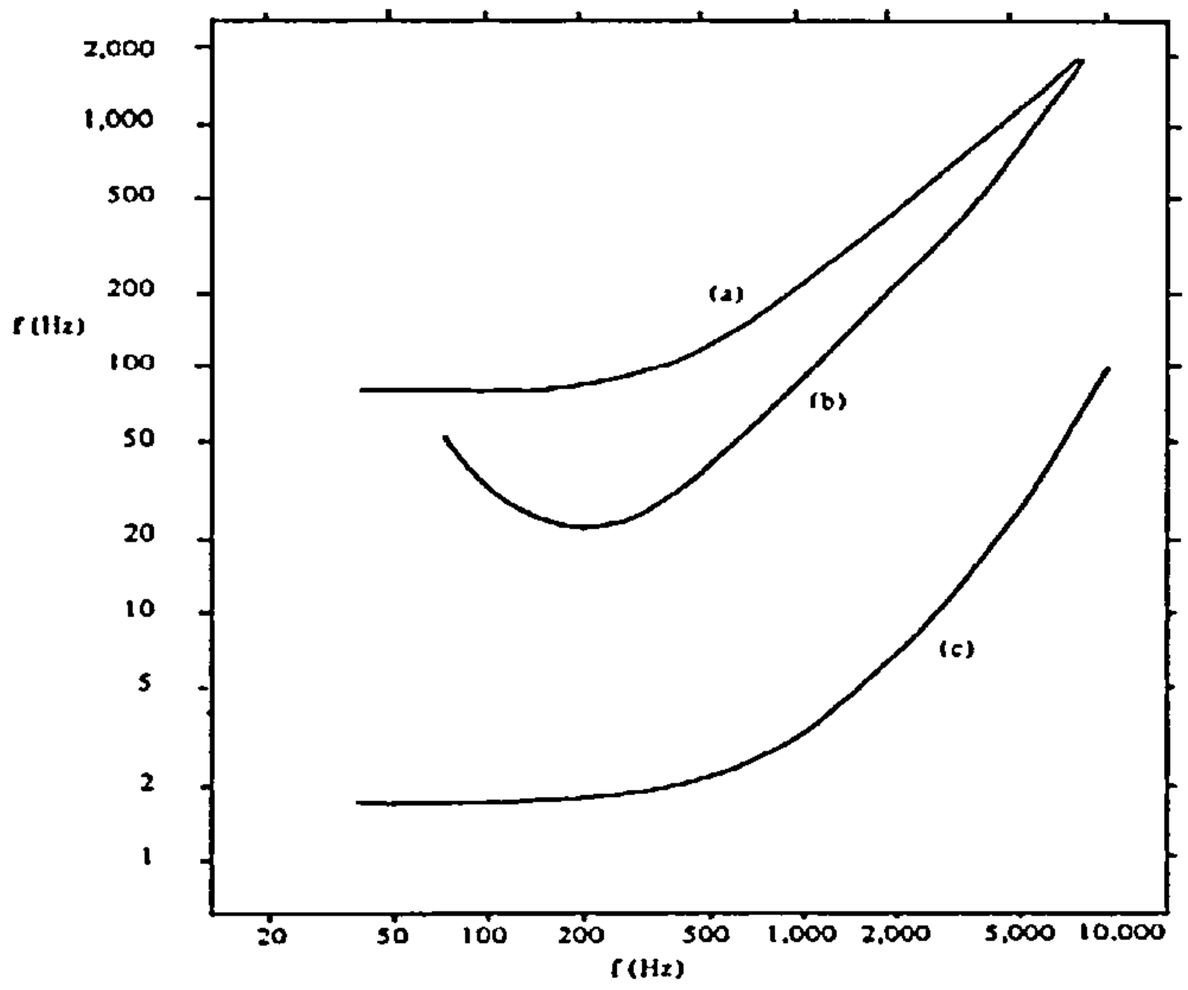


Figure 2.32. Critical bandwidth measurements: (a) Critical bandwidth, (b) Minimum frequency separation for which two simultaneous pure tones can still be distinguished, (c) Minimum detectable sudden change in the frequency of a pure tone, with SL = 80 dB (from Campbell & Greated, 1987).

2.3.5 Loudness

As pitch can be thought of as the perceptual correlate of frequency, so loudness can be thought of as the perceptual correlate of intensity. Like pitch, loudness changes in discrete steps whilst its physical correlate intensity can be continuously varying. Again, as in determining pitch, other factors influence the perception of loudness, such as frequency, spectral shape and context.

2.3.5.1 The Threshold of Audibility

Some sounds lie below the threshold of audibility, that is, they are so quiet that they cannot be heard. Likewise, some sounds are so loud that they induce a tactile sensation of tickling, or pain, along with or instead of an auditory sensation. These loudness levels lie at the upper limit of loudness. The threshold of pain does not depend on frequency.

The smallest pressure which is required for a tone to be audible in the absence of any external sounds is called the minimum audible pressure (MAP). This varies as a function of frequency. More pressure is needed to hear a low-frequency or high-frequency tone than a mid frequency tone, that is, we are less sensitive at low and high frequencies than to mid-range frequencies, reflected in the threshold curves in figure 2.33.

Von Bekesy (1960) suggests that the minimum audible pressure curve is a biological function aimed to block out the sound of natural processes of our body, such as blood flow or footsteps and is determined partly by the transmission characteristics of the middle ear (Moore, 1989). Our sensitivity to mid-range frequencies (in the range 1000-5000 Hz) can also be attributed partly to the functions of the pinna and ear canal, the outer ear increasing sound pressure at the eardrum for frequencies ranging from 1-9 kHz, with a peak at 3 kHz of 15 dB (Moore, 1989). We become incredibly desensitized to intensity discrimination and pitch at the extremes of audibility (below C0).

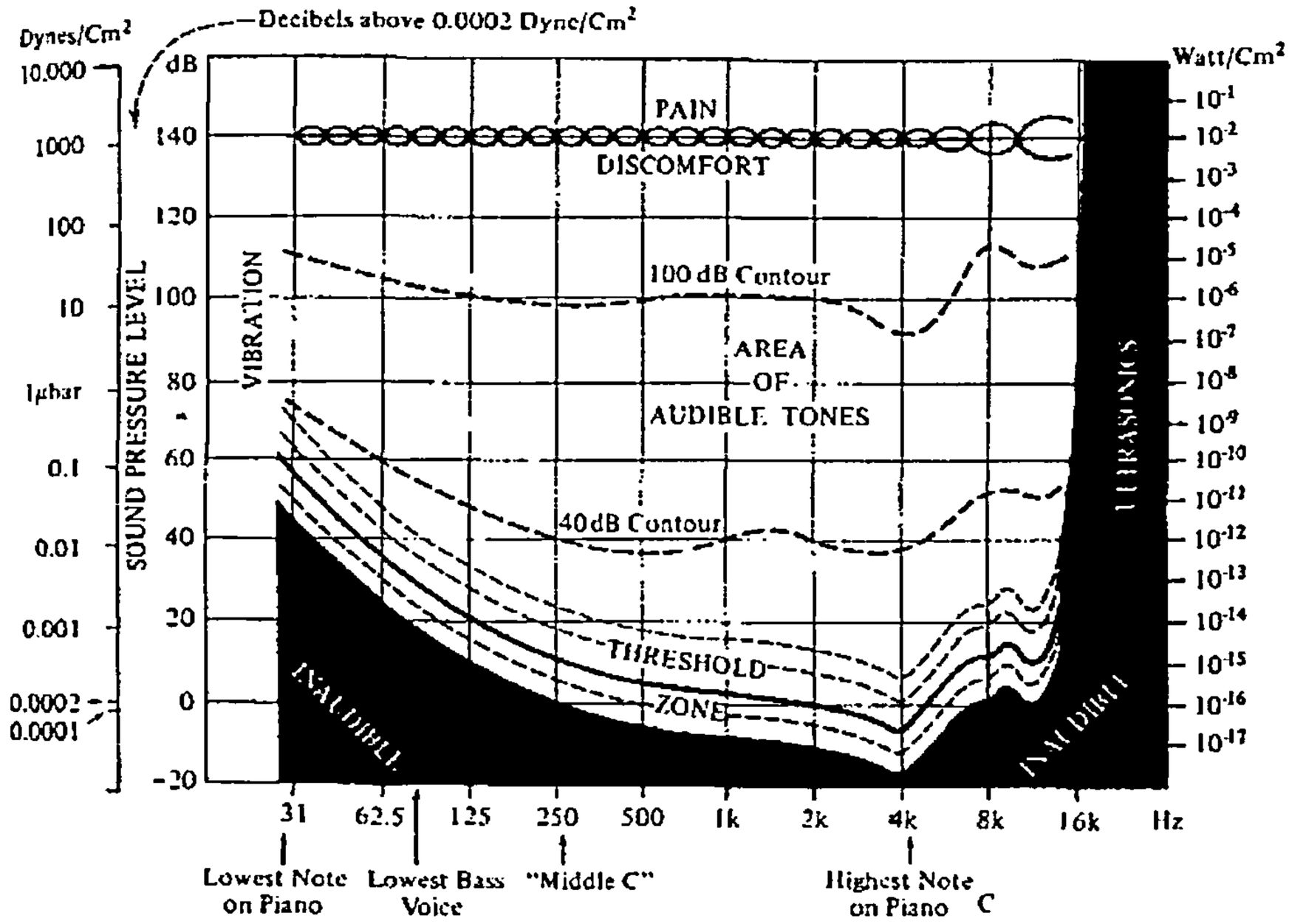


Figure 2.33. The area of audible tones (from Gerber, 1974).

The threshold of audibility of a tone depends on its duration up to 500 ms (Garner, 1947). Perceived loudness increases up to about 1 sec (Littler, 1965).

Sounds having the same intensity can differ greatly in loudness (Gerber & Bauer, 1974). This is demonstrated by using a Fletcher-Munson diagram which represents contours of equal loudness for sine waves at different frequencies. This is shown in figure 2.34. If intensity level of a tone is kept constant and then its frequency is changed, its loudness changes.

2.3.5.2 Loudness Levels and Equal Loudness Contours

Loudness levels, measured in phons, determine how intense a pure reference tone of 1000 Hz must be in order to sound equally loud to a previously heard pure test tone, or vice versa. From equal loudness contour graphs (see figure 2.34) it can be shown that the rate of growth of loudness of pure tones is dependent on frequency. Low frequencies and very high frequencies have a greater growth of loudness with increasing intensity than for mid-range frequencies.

The overall level of a complex sound determines the relative loudness of the different frequency components in that sound. Changing the overall level of a complex sound will vary its 'tonal balance' (Moore, 1989). At low sound levels, the ear is less sensitive to very low frequencies and very high frequencies in a complex sound than mid-range frequencies so these components do not add much to the total loudness of the sound. However, at high levels, we are relatively more sensitive to very low frequencies and very high frequencies at high levels; all frequencies similarly contributing to the loudness level, as shown in flat equal-loudness contours (Moore, 1989).

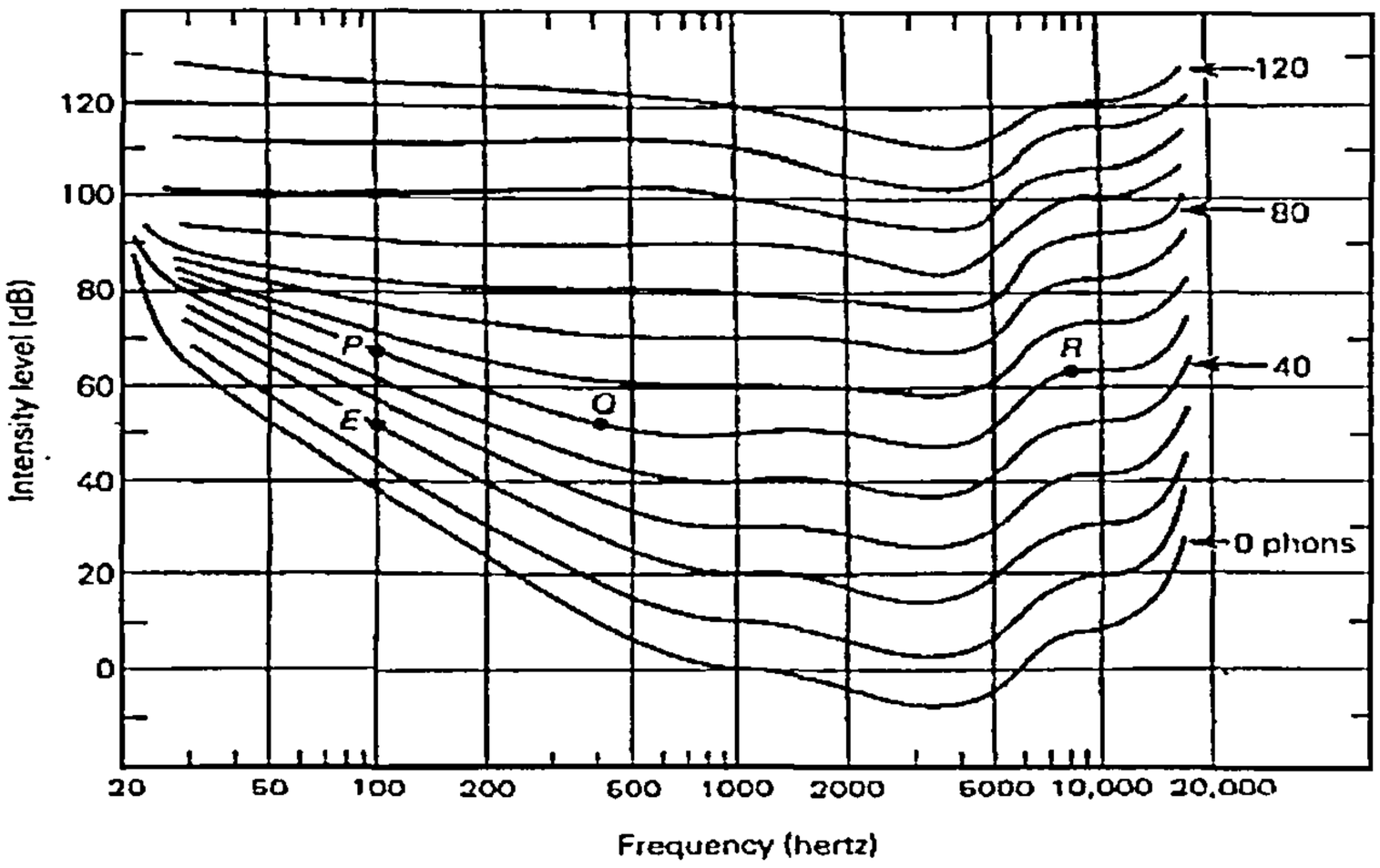


Figure 2.34. Contours of equal loudness for sine waves (Campbell & Greated, 1987).

2.3.5.3 The Musical Dynamic Scale

Decibels measure ratios, or the difference between two sounds; not absolute quantities. It is a scale of magnitude. Similar to turning the relative pitch scale into an absolute scale where A4 = 440 Hz, an absolute logarithmic intensity scale can be drawn by choosing a standard intensity reference. The mathematics behind intensity measurements are found in Campbell & Greated (1989).

In terms of music, loudness is expressed in discrete dynamic markings usually ranging from ppp to fff. For a pure 1000 Hz tone there appears to be a logarithmic relationship between intensity and musical dynamic level, shown in table 2.2.

The ear's dynamic range is large; the intensity of a fff tone is ten million times greater than for a ppp tone. The ear's dynamic range is not proportional to frequency, but contracts at the lower end. A low pitch requires a smaller change in intensity to go through the dynamic levels from ppp to fff, than a higher pitch. This can be related to the dynamic range available on the threshold curves.

Musical dynamic level	Intensity (Wm^{-2})	IL(dB re 10^{-12} Wm^{-2})
fff	10^{-2}	100
ff	10^{-3}	90
f	10^{-4}	80
mf	10^{-5}	70
mp	10^{-6}	60
p	10^{-7}	50
pp	10^{-8}	40
ppp	10^{-9}	30

Table 2.2. Rough correspondence between intensity and musical dynamic level for a single isolated 1000 Hz tone (from Campbell & Greated, 1987).

2.3.5.4 Loudness Masking of Complex Tones

The context in which one sound is heard determines the apparent loudness of that sound. This happens for different instrument combinations and for individual partials in a single complex tone.

The loudness of a complex tone is not necessarily simply the sum of the loudnesses of its partials. This is because each partial may reduce the loudness of adjacent partials. This process is known as masking. A sound is masked if its threshold of audibility is raised by the presence of another sound. A sound is considered to be totally masked if it is inaudible in the presence of another sound but audible in its absence (Zwislocki, 1978). The amount by which the threshold of audibility is raised is measured in decibels.

The chances of masking occurring is increased if the masking signal has frequency components which are the same or similar to the signal being masked (Mayer, 1894; Wegel & Lane, 1924).

Masking can be related to critical bands on the basilar membrane. Pure tones which are sufficiently close together in frequency activate the same group of neurons. This results in a complex mixture of the signals. However, if the tones are sufficiently separated in frequency, they will activate different sets of neurons, two separate signals are sent to the brain and no masking will occur.

Lower frequencies mask higher frequencies far better than the other way round. This is because the amplitude envelope drops sharply in the direction of the lower frequencies, yet fans out to the higher frequencies in the direction of the oval window. as seen in figure 2.35. A lower tone will mask an upper tone if it comes under its envelope tail. Total masking can be thought of as a distortion of the threshold of audibility curve.

It is also possible for one tone to partially mask another tone. That is, the masker does not obliterate the presence of another tone, but just reduces the loudness of it. Partial masking can be thought of as a distortion of the equal loudness contours as shown in figure 2.34.

A loud musical note may include the effects of masking over 2 octaves above its fundamental pitch. In this case each partial may contribute to the overall masking effect, raising the threshold of the tone.

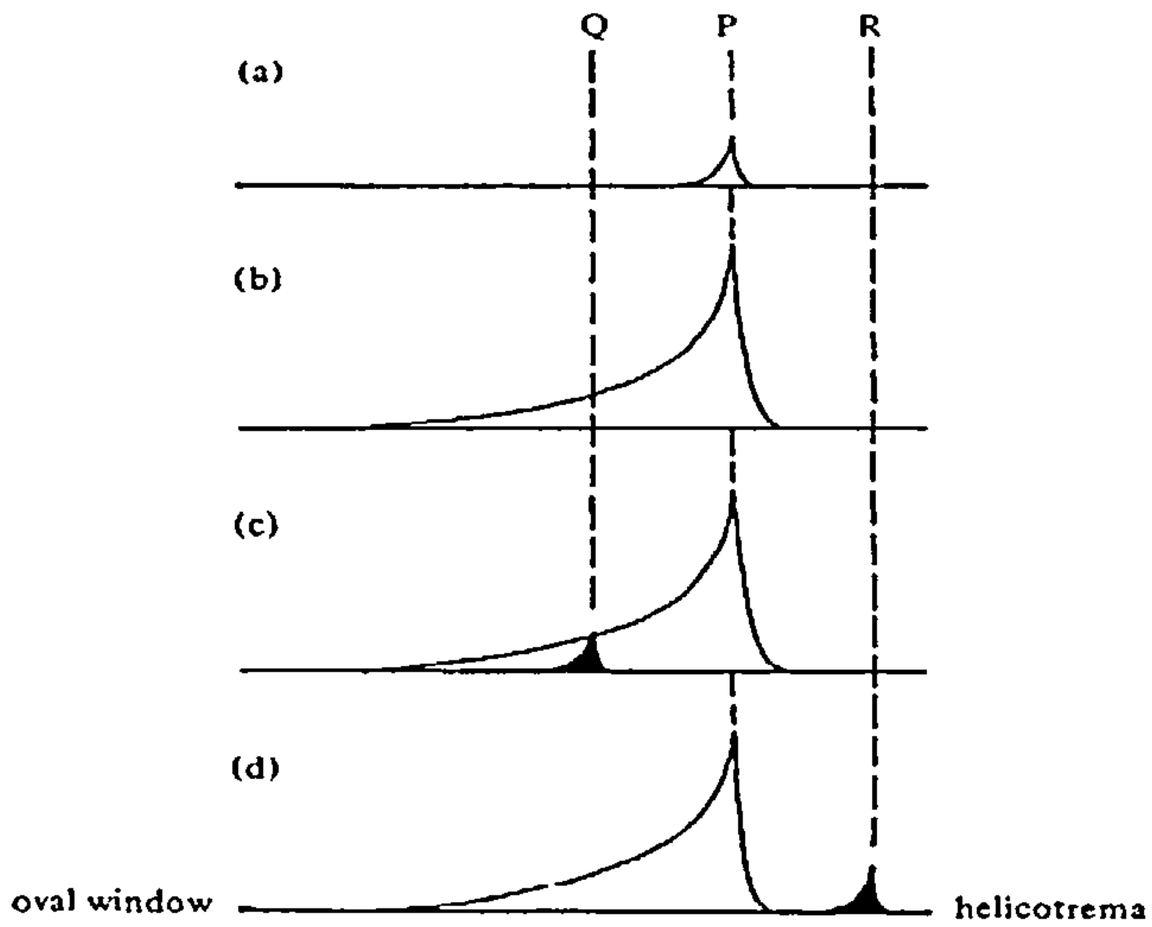


Figure 2.35. Basilar membrane amplitude envelopes corresponding to (a) quiet 1000 Hz tone; (b) loud 1000 Hz tone; (c) loud 1000 Hz tone + quiet 2000 Hz tone; (d) loud 1000 Hz tone + quiet 500 Hz tone (from Campbell & Greated, 1987).

For a moderately loud tone with partials separated by several critical bands, there is hardly any masking of partials. The total loudness will be the sum of the loudnesses of the component partials. If, however a complex tone has a small "cluster" of upper partials all lying within one critical band, each partial excites roughly the same region of nerve fibres, so the intensities can be added together. In summary, adding a second partial to one within one critical band and at equal intensity increases the loudness level by a lot less than if the second partial is outside the critical band.

2.3.6 Timbre

Timbre can be defined as "that attribute of a tone by which a listener can judge that two sounds of the same loudness and pitch are dissimilar (ANSI 1973)" (Handel, 1989). It has also been described as "the characteristic quality of an instrument which enables it to be identified" (Campbell & Greated, 1987).

Whereas pitch and loudness can be described along a one-dimensional continuum, the parameters of timbre are multidimensional and complex, changing both dependently and independently.

Several parameters contribute to the perception of timbre. These include the steady-state spectral components of the sound, the amplitude envelope of the sound, and any transients that may be present. The degree of contribution of each attribute to the overall timbral perception will depend on the nature of the instrument producing the sound and the way the sound is produced. These will be described in turn.

2.3.6.1 Steady State Spectral Components

Differences in the spectrum of a sound changes its timbre. The ear is very sensitive to differences in the spectrum of periodic waveforms, but is relatively insensitive to changes in phase. Timbre is dependent on the number and amplitude of the components in the sound. It is the combined strength of the spectral components within the critical band which determines the sensation of timbre. Harmonics whose frequencies are within 15% of each other lie within one critical band. Harmonic components up to the sixth or seventh harmonic contribute independently to timbre. At around the seventh harmonic, the components start to overlap and merge. The 28th through to the 32nd harmonics lie within one critical band.

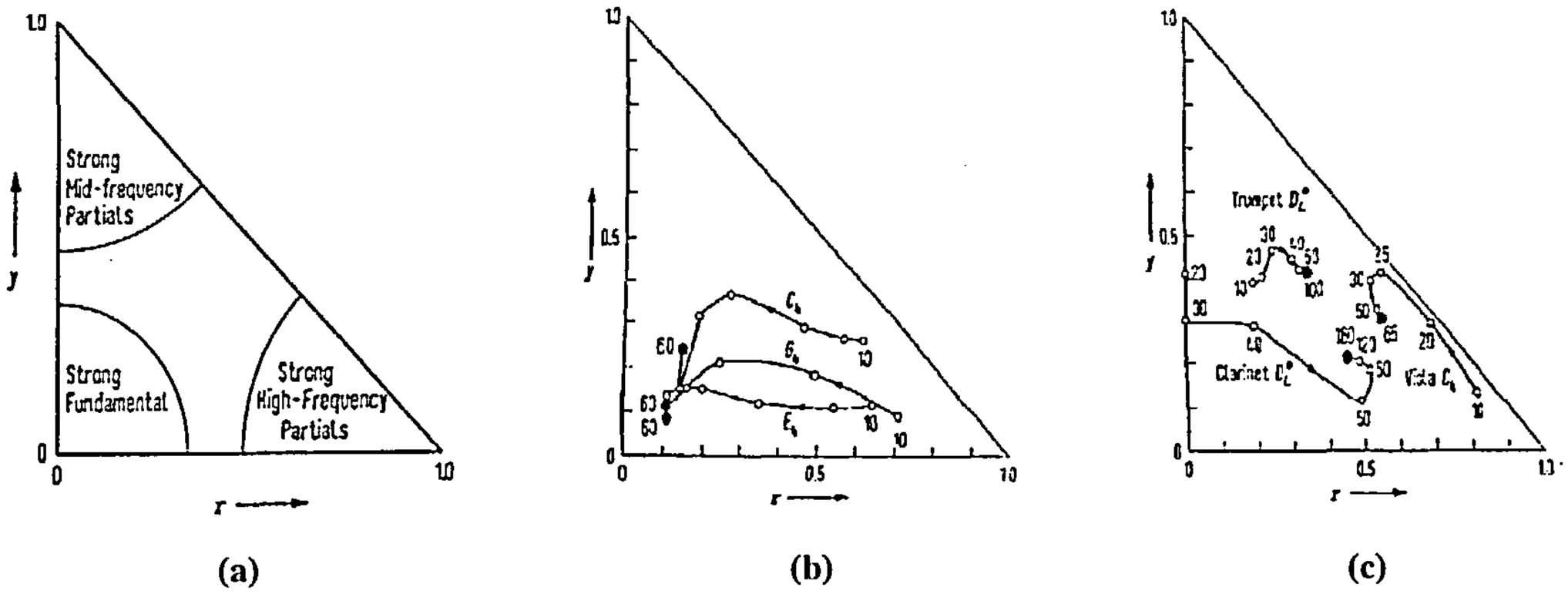
Rating the timbre of steady-state portions of a sound has proved to be very difficult due to its subjectivity. However perceptual studies have independently reached the same conclusion that most timbres can be reasonably correctly described using three scales (Bismarck 1974; Plomp 1976).

Out of these, the dull-sharp scale gives consistent results. Hall (1980) uses scales catalogued in a similar way to taste sensations where the brain only picks out just three or four major "independent intensity parameters" (Plomp, 1976) representable as numbers on polar scales, regardless of the actual number of separate pieces of information.

fine	---	coarse
reserved	---	obtrusive
dark	---	bright
dull	---	sharp
soft	---	hard
smooth	---	rough
broad	---	narrow
wide	---	tight
clean	---	dirty
solid	---	hollow
compact	---	scattered
open	---	closed

Table 2.3. Some verbal scales used to rate timbre (after Bismarck) (from Campbell & Greated, 1987).

Bismarck (1974) also uses 28 verbal scales used to rate timbre, many of which overlap. Another method uses a tristimulus diagram. This represents the timbre of any steady state sound as a point in 2 dimensions, by dividing the spectrum up into 3 parts and using three independent variables to represent the loudness (Pollard & Jansson, 1982). Tristimulus diagrams are presented in figure 2.36.



(a) (b) (c)

Figure 2.36. Tristimulus diagrams representing the timbre of musical instruments. The fraction of the intensity in partials 2,3, and 4 is plotted along the y-axis, and the fraction in higher harmonics along the x-axis.

(a) Regions where tones with strong fundamentals, midfrequency partials, and high-frequency partials would be found.

(b) Tibre of three Gedact organ pipes from 10 to 60 ms after attack.

(c) Attack transients of trumpet, clarinet, and viola (Pollard and Jansson, 1982).

2.3.6.2 Transients

Different pressure/time waveforms can generate sounds with perceptually the same timbres. These waveforms have the same overall shape whilst having different phase relationships. They can all be represented by a single harmonic spectrum. Conversely, the same perception of timbre can be obtained from different harmonic spectra. This is accounted for by the fact that the timbre of an

instrument may change for each note, and each dynamic level, and also the player will introduce "involuntary fluctuations" into the sound. Part may also be due to small differences in the relative position of the instrument to the microphone, where the pattern of directionality of independent harmonics may be different, thus preferentially favouring certain harmonics (Meyer, 1978).

Musical timbre depends critically on transient cues produced during the first 20-50 ms following the onset of the tone. Such transient cues correspond to the cues for consonants in speech. In other words, temporal patterning can be crucial in the perception of timbre.

A steady-state repeated waveform cut from the middle of a complex tone is sometimes not enough to establish the identity of the instrument being played. Even though each instrument can be broadly characterized by its own spectum, it is sometimes difficult to recognise since our recognition of instruments necessarily relies on more than just the steady-state spectrum. It appears that we are very sensitive to non-linear additional spectral bursts which accompany the steady-state tone. This can include, the attack (beginning) and the decay (ending) of a complex spectral envelope, the initial sound from an instrument before it settles down to its steady-state natural mode of vibration, or incidental noise or sound during production. Not only that, but the instrumentalist may add his/her own colour to the instrument spectrum and can vary the attack and decay, thus modifying the transients. Whole notes may possibly consist entirely of transients such as for jabbing staccato notes from wind or brass instruments.

Saldanha and Corso (1964) showed that steady-state patterns are not sufficient as a basis for all musical timbre discrimination. They showed that the clarinet, oboe, and flute were easiest to identify and the trombone, violin, cello, and bassoon were the most difficult to identify. Overall performance dropped from 47% correct to 32% correct when the onset transients of the instrumental sounds were eliminated, leaving the steady-state part of the musical instrument tone.

Onset transients ("attacks") are more important as a cue to instrument identification or spoken syllables than decay transients (Saldanha and Corso, 1964). The spectral components of a complex tone can vary in different ways during initiation: the onset may include transient noises; there may be formant shifts arising from varying intensity changes of the harmonics; and the actual duration of the transient, the overall rate of attack (the steepness in reaching an intensity peak) also serve as strong cues to instrument identification (Winckel, 1967). Initial attack times usually differ for low and high notes. The spectrum during an attack is always varying, each spectral component rises at different rates. A tristimulus diagram can also be used to chart the relative intensities of the spectral components as the transient progresses, as in figure 2.36 (Rossing, 1990).

2.3.6.3 The Amplitude Envelope

The tone onset or attack is an important feature which helps us identify the instrument being played. Some instrument tones do not have a steady state, such as the harp or piano. Their principle timbral cue is the "amplitude envelope". Timbre comprises of both steady state components and transient components. The "characteristic" of an instrument, leading to its recognition, is its

"combination of acoustic variables" (Campbell & Greated, 1987). Risset & Matthews (1969) have found that brass instruments are principally characterised by the manner in which their timbre changes with volume. Clarinet tones are characterised by strong first and third harmonics for most of the playing range, whilst flute sounds have very strong fundamentals but their harmonic series decreases in intensity with increasing pitch. It has been suggested that the listener gains a "timbral constancy" through familiarity with the instrumental groups, allowing identification of the instrument at any point through its playing range (Erickson, 1977).

2.3.6.4 Timbre Changes with Frequency

A pure sine wave (that is, all the energy in the first partial) has a timbre that changes from very dull at low frequencies (due to lack of energy above the first partial) to very bright at high frequencies (due to the frequency of the sinusoid falling within the frequency region where our hearing perception is at its most acute). Instruments characterized by strong formant ranges tend to keeps timbre uniform for about an octave (Hall, 1980). This is true for vowels, plotted on a pitch scale in figure 2.37, and also trombones (Campbell & Greated, 1987).

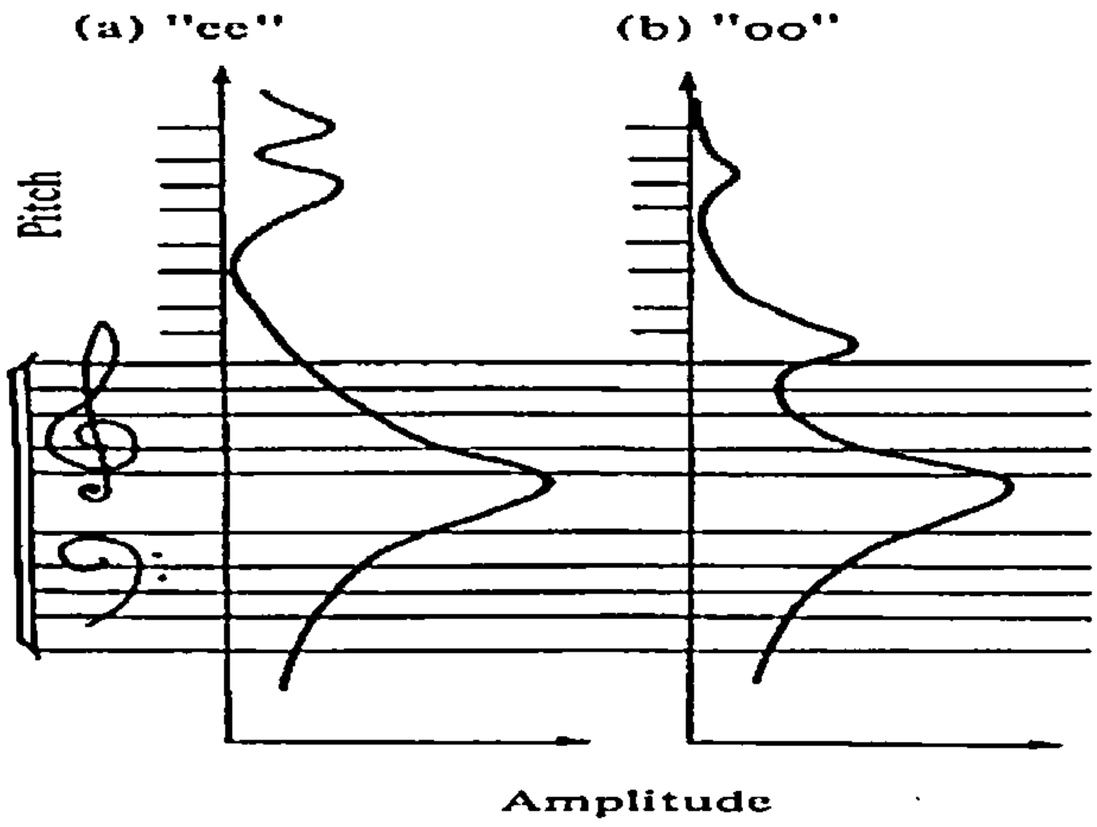


Figure 2.37. Formants of the vowel sounds (a) 'ee'; (b) 'oo' (from Campbell & Greated, 1987).

2.3.6.5 Timbre Changes with Intensity

Timbre also changes with intensity. The intensity of a given harmonic is proportional to the square of its pressure amplitude. A complex tone with strong harmonics and a fundamental frequency in the range of 100 Hz to 200 Hz has a rich full timbre when presented to the ear at levels of 80 dB to 100 dB. This is because the ear is sensitive to all harmonics at this intensity level. However, if the same tone, with the same relative distribution of partial energy is presented to the ear at levels of only 40 dB to 50 dB, the tone will sound softer and also thinner in quality. This is because the ear's sensitivity to lower partial components decreases much more than for others as intensity decreases.

Harmonic spectra do not show the relative importance of the amplitudes of the harmonics in the perceived sound. For this we need a loudness spectrum which is similar to a harmonic spectrum but takes into account that the ear is more sensitive to high frequencies. Each harmonic is converted into its corresponding loudness level or loudness reflecting the fact that a harmonic with double the amplitude has four times the intensity.

Dowling & Harwood (1986) simplify the interactions by first reducing timbre down to the two types of acoustic categories derived from speech science, namely the vowel and the consonant. "..the steady-state correlates of vowel-like timbre differences and the transient (rapidly changing) correlates of consonant-like timbre differences" apply to music perception as well as speech perception. This may be because "complex sounds such as spoken vowels are actually discriminated in terms of timbre variations rather than variations of fundamental frequency" (Gerber & Bauer, 1974).

2.3.6.6 Timbre and Singing

Timbral differences between tenors and altos can be explained by the larger size of the larynx tube and the vocal folds of males as compared with females. The larynx tube dimensions influence the 4th formant frequency. By increasing the larynx tube dimensions, the 4th formant frequency is lowered in tenors, bringing it closer to the 3rd formant, and positioning both the 3rd and 4th formants within a critical band. The increased ability of the vocal tract to transmit sound between the two formants results in an intensification of the spectral partials in this region, perceived as a harsh or rough auditory quality (Terhardt, 1974; Sundberg, 1987), shown in figure 2.38.

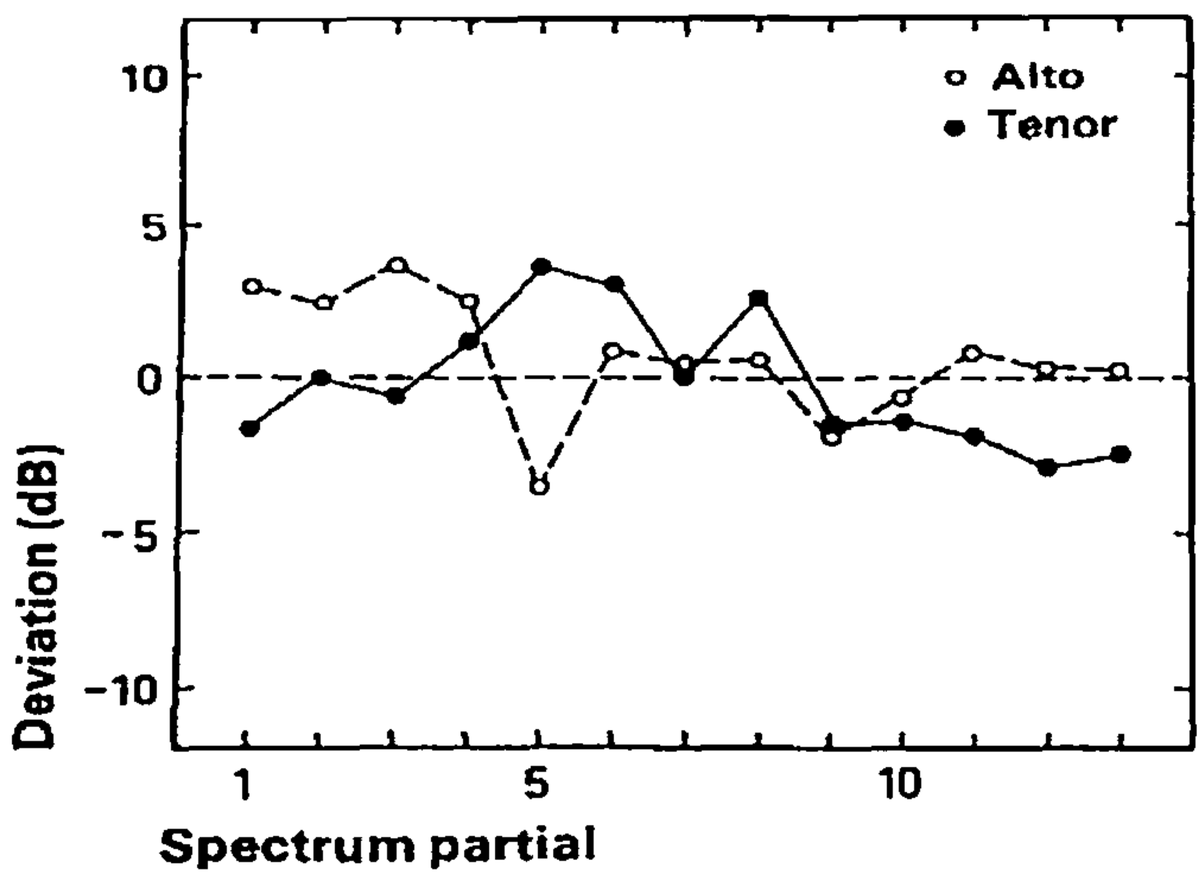


Figure 2.38. Average source spectra from two alto (open circles) and two tenor (filled circles) singers singing at identical pitches (after Ågren & Sundberg, 1978).

This is possibly not the reason for the difference in quality between opera and belting in the same singer, since larynx tube dimensions remain relatively fixed. Sundberg (1987) suggests that voice source differences also determine perceived harshness. He believes that reducing voice source spectral tilt, (i.e. increasing the amplitude of the higher spectral partials of the voice source) will also increase the roughness of the voice quality. The amplitudes of the voice source partials indicate the rate at which the glottis closes.

Chapter 3

Standard Speech Analysis Techniques

3.1 Introduction

This chapter will describe standard non-invasive speech analysis techniques which are commonly used by voice scientists and speech therapists in assessing speech qualities.

3.2 Voice Source Parameter Analysis

The voice source analysis method used in this study concentrates on laryngography techniques. Laryngography provides a means of assessing vocal fold contact area and the nature of vocal fold vibration.

3.2.1 Electrolaryngography

Vocal fold vibration can be monitored non-invasively using a device known as an electrolaryngograph (Fourcin & Abberton, 1971; Fourcin, 1987; Abberton et. al., 1989). In laryngography, a small constant high frequency voltage is applied between a central conductor and an outer guard ring of one electrode. The other electrode acts as a receiver and picks up the high frequency current flow. The electrodes are strapped on either side of the larynx at the level of the vocal folds, and the current flowing between the electrodes is measured. In voiced sounds, for each vibratory cycle, as vocal fold contact area increases the current flow increases. This current modulation can be represented on an oscilloscope. A signal which increases in amplitude when the vocal folds are closing, and decreases in amplitude when the vocal folds are opening is called the Lx waveform and represents conductance. The USA version of the electrolaryngograph is the electroglottograph (E.G.G.). This measures impedance rather than conductance, and consequently the E.G.G. waveform is the inverse of the Lx signal. Vocal fold closure is represented by the steepest slope on both the Lx and E.G.G. signal. From this waveform the closed and open portions or phases of each vibratory cycle can be calculated. Figure 3.1 represents the output waveform of the laryngograph, the Lx signal as it varies with time (Abberton et al., 1989).

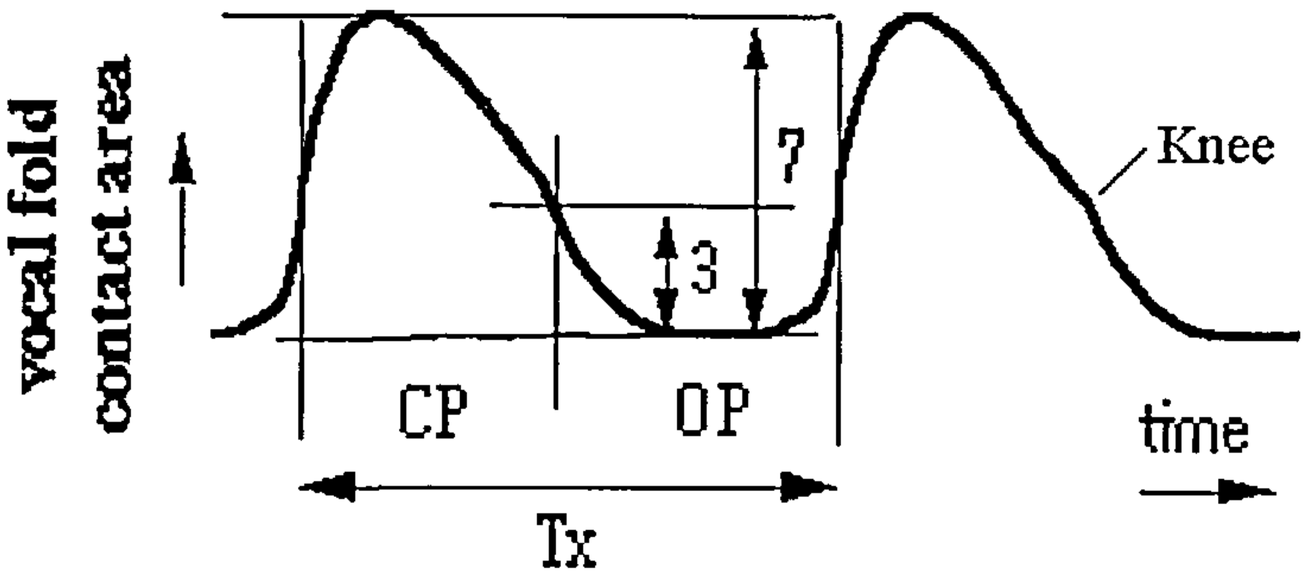


Figure 3.1. A typical Lx waveform with open and closed phases (after Evans & Howard, 1993).

There are four distinct portions of this waveform relating to the degree of conductance between the vocal folds, shown in figure 3.2. In normal modal phonation, the waveform can be divided into an open and a closed phase, corresponding to the degree of closure of the vocal folds. The closed phase is further divided into three parts: the closing phase, the peak closure, and the opening, or separation phase.

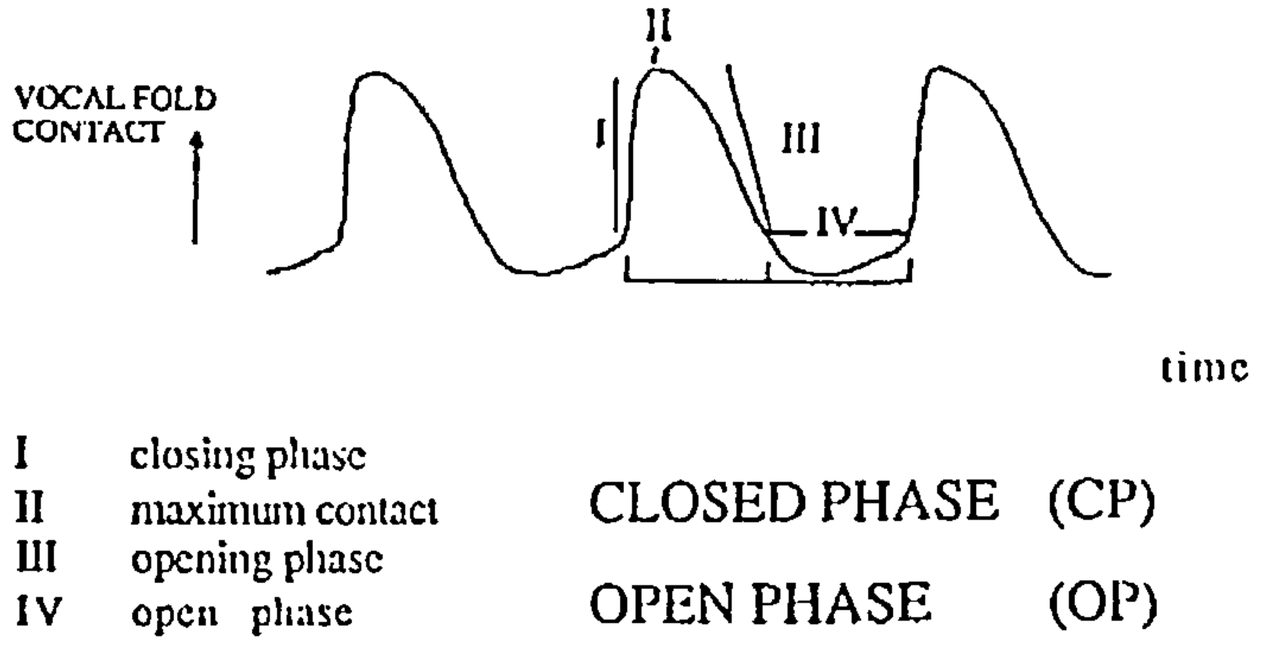


Figure 3.2. A few idealized cycles of a modal laryngograph output waveform (Lx) to illustrate the main phases in each cycle (from Abberton et al., 1989).

The closing phase is characterized by a steep rise in the amplitude of the Lx waveform. This corresponds to the snapping shut of the vocal folds from the bottom upwards. The period of maximum closure is represented by the peak in the differentiated Lx waveform. The opening phase corresponds to the slow peeling away of the vocal folds, resulting in a more gradual decrease in amplitude of the Lx waveform. It is a matter of definition determining the open and closed portions of the Lx waveform. The point at which the vocal folds separate is harder to determine than the point of closure, since the Lx waveform decreases in amplitude before any air can escape through the glottis (Breen, 1990). However, the point at which air flow starts, corresponding to the beginning of the open phase can be easily seen as a knee in the opening phase of the Lx waveform of many speakers (Breen, 1990). This knee can be seen in figure 3.1.

Successive points of closure of the laryngograph signal determine the duration of each cycle (Tx), from which fundamental frequency (F0) can be estimated. Other parameters which relate to the timing of pitch such as flutter, jitter, and diplophonia may also be calculated from the Lx signal. Laryngograph signals not only allow the estimation of pitch periods, but also the relative durations of the open and closed phases of each vibratory cycle which provide an indication of the nature of vocal fold vibration which are of great importance in characterizing vocal quality and voice pathology. One such method which gives an indication of voice quality is larynx closed quotient estimation.

Larynx closed quotient, or CQ, is defined as the percentage of each larynx cycle for which the vocal folds are in contact (Davies et al., 1986). It is calculated here as follows:

$$CQ = ((CP/Tx)*100)\%$$

CQ is often plotted against time (figure 3.3) or as a scattergram (figure 3.4).

.

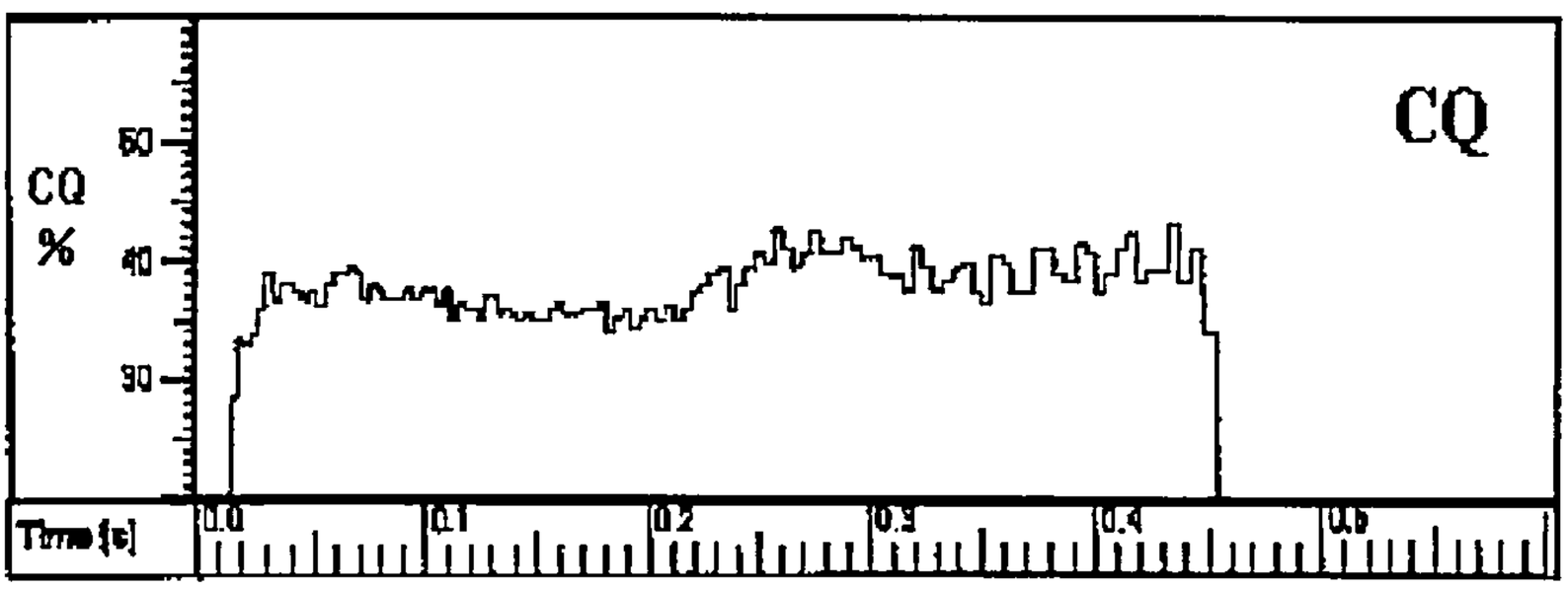


Figure 3.3. CQ plotted against time for a female speaking the word "bard".

A two channel electroglottograph (Glottal Enterprises) will also be used to measure larynx height. This device is described fully in chapter 5 since it is a relatively new invention and is not yet a standard speech technology item.

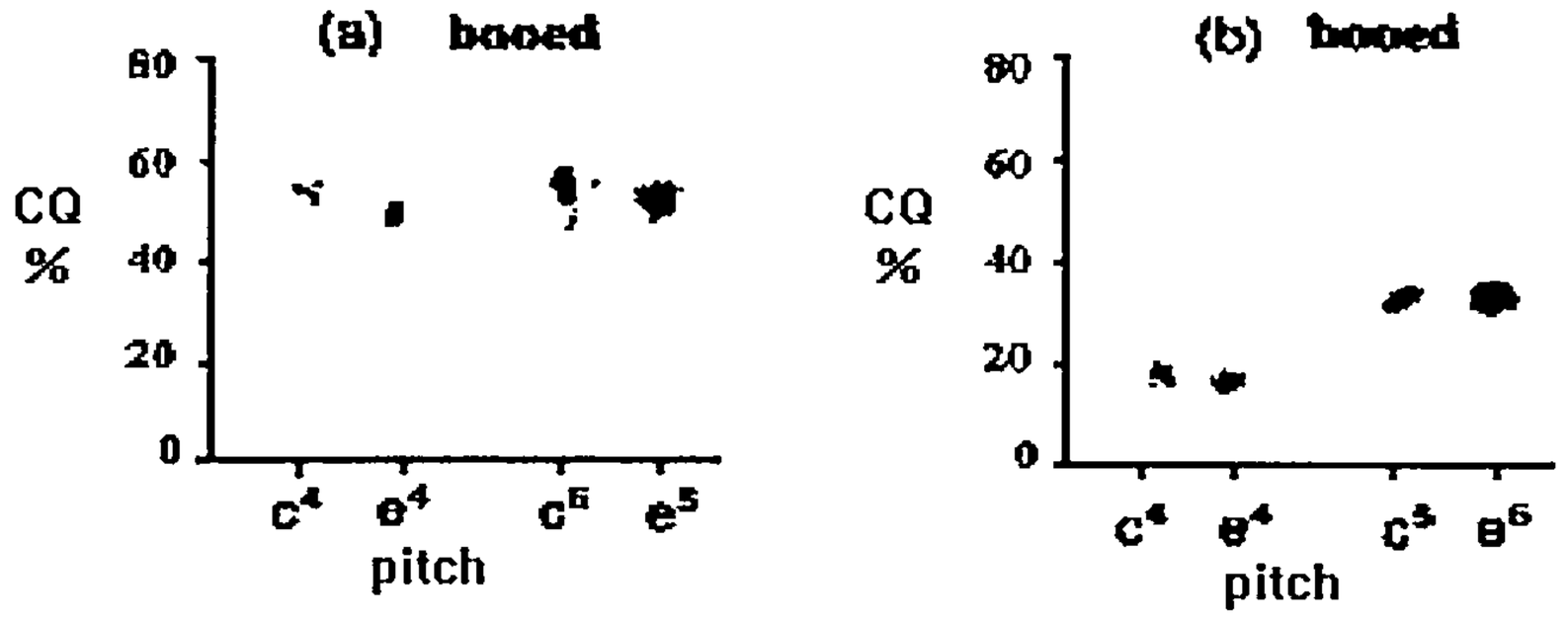


Figure 3.4. CQ (%) against pitch scattergrams for the vowel /u:/ (from the word "booed") sung in (a) belt quality and (b) opera quality by a soprano (after Evans & Howard, 1993).

3.3 Acoustic Signal Parameter Analysis

The analysis of the acoustic output for the purposes of this work divides into two areas; the overall energy of the acoustic signal, and the spectral content of the acoustic signal as it dynamically varies (spectrography) or as an average (average spectrum).

3.3.1 Sound Pressure Level Recording

A sound pressure level meter records sound level using an inbuilt microphone which adjusts to a selection of international and national standards frequency response weightings. The C-weighting is used here as it provides a flat response across all the audible frequency range and can be used to measure the acoustic output of an instrument such as the voice. The meter is calibrated in decibels (Rossing, 1990).

3.3.2 Fast Fourier Transfer Analysis

Fast Fourier Transform (FFT) is based upon converting continuous (analog) signals into discontinuous (digital) signals in the A/D converter. The resulting digital signals are stored then processed by the analyzer before being output to the screen. In the AND AD-3523 Sound Analyzer used in this research, the input signal is optimised by an amplifier then fed into an anti-aliasing filter which eliminates frequencies above the specified frequency range before being passed into an A/D converter. This is shown in figure 3.5 (from the AND AD-3523 manual).

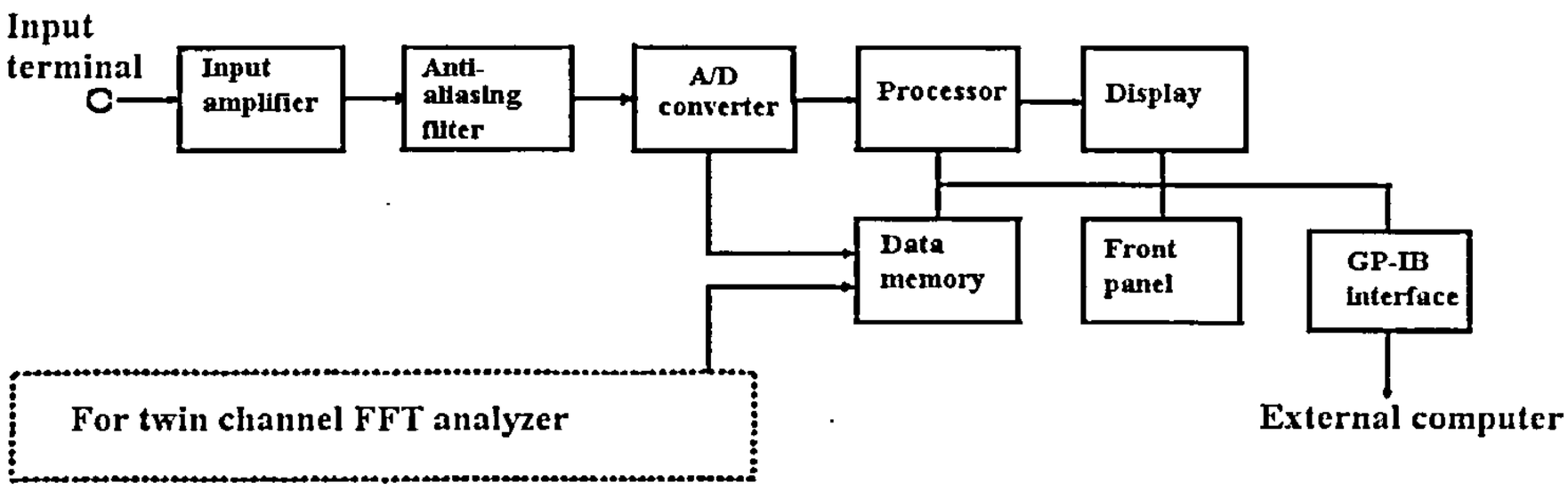


Figure 3.5. Typical Configuration of the FFT Analyzer (from AND AD-3523 Sound Analyzer manual).

FFT Analysis performs high-speed spectral analysis on an input signal by observing the signal for a discrete period of time, known as a frame. The duration of this frame is the frame time. This frame is split into a fixed number of points with associated signal values. These values are called samples,

and are necessarily an approximation of the input signal. The fixed interval between each sample is known as the sampling cycle. Its reciprocal is called the sampling frequency. This process is called Discrete Fourier Transformation (DFT). A knowledge of sampling theory and the above processes can reduce the approximation error when analyzing signals. DFT analysis provides the basis for both traditional spectrographs and spectra. Average spectral analysis has been chosen over spectrographic analysis since this work concentrates on the steady-state of a vowel only, so the average spectrum is adequate. However, spectrography will be reviewed.

3.3.3 Spectrography

A spectrograph is a device which produces a dynamic graphical representation (called a spectrogram) of the relative amplitudes of the spectral content of an acoustical signal with time. Different acoustic information is highlighted as dark horizontal bands, with frequency on the vertical axis, and time on the horizontal axis. The relative darkness of these bands corresponds to the intensity of the energy of that band within a specified bandwidth. In voice analysis, a narrow bandwidth gives individual partial information (seen as thin dark bands which are regularly spaced at intervals equivalent to the fundamental period), shown in figure 3.6, whilst a larger bandwidth will show formants of the signal, that is, high energy peaks (seen as broader dark bands), as seen in figure 3.7.

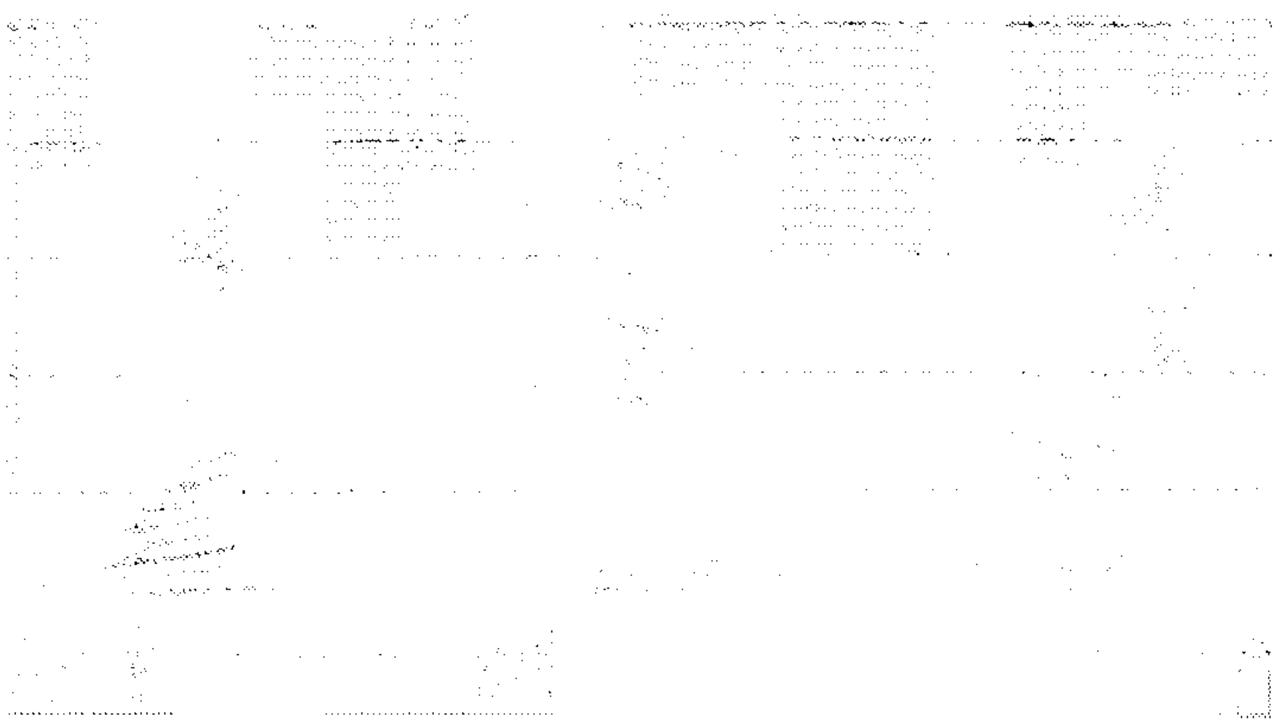


Figure 3.6. A narrow band speech spectrogram (from Curtis & Schultz, 1986).

By varying the analysis bandwidths, it is possible to get both individual partial information and formant information, though there will be some loss of clarity for both.

Spectrographs may be real-time, providing spectral information on the voice almost instantaneously, or non-real-time, working on a stored speech file. The spectral information from a specified duration of signal can sometimes also be averaged as a long-time average spectrum (LTAS) which can be useful in assessing the quality of a sustained sung tone. At any point or "frame" along

the duration of the time-varying spectrogram the darkness of marking shows the relative amplitudes of the acoustic components within a preset frequency range, normally well within the audible hearing range. This slice is called a spectrum.

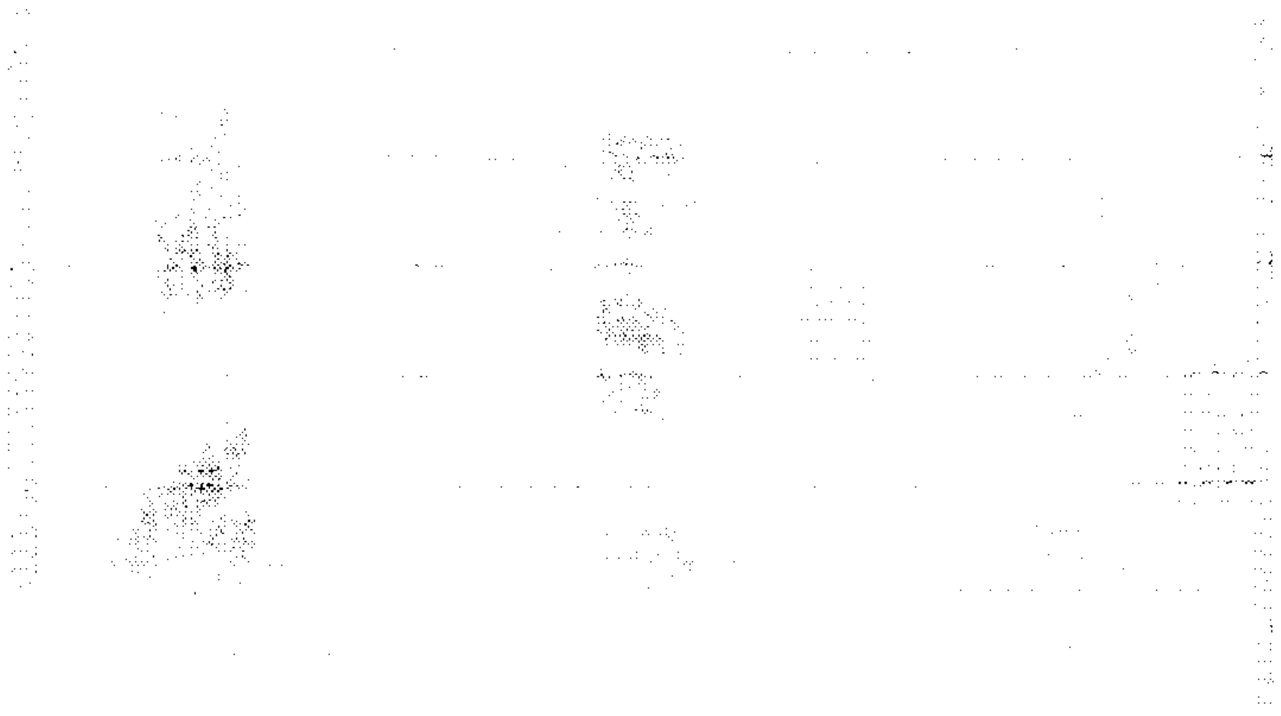


Figure 3.7. A wide band speech spectrogram (from Curtis & Schultz, 1986).

A spectrum may also represent the average spectral contents of a sound over a period of time. An example of this is given in figure 3.8 which shows the average spectrum for a sustained vowel /a:/ sung in opera quality at pitch E4 and at pitch E5 by a soprano. An AND AD-3523 sound analyzer was used for this purpose.

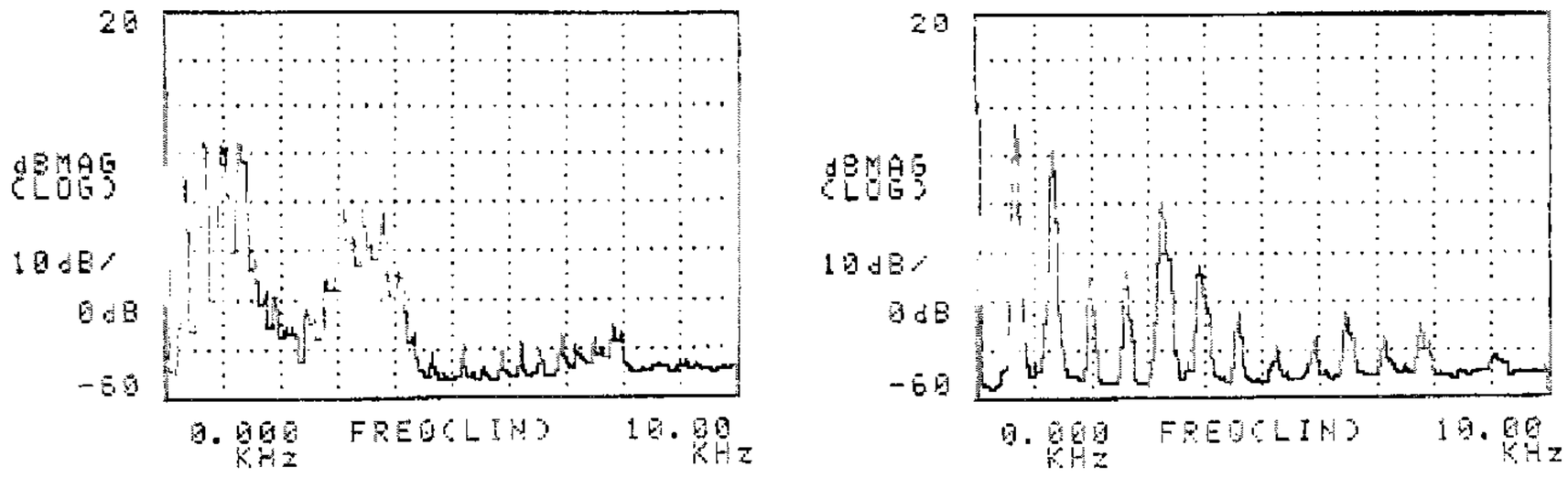


Figure 3.8. Average spectra for the vowel /a:/ sung at pitch E4 (on the left) and at pitch E5 (on the right) in opera quality by a soprano.

Speech signals decay at at least -6 dB per octave. This arises as a sum of combining the natural radiation characteristics of the glottis (-12 dB per oct) and the lips (+6 dB per oct). This means that the higher the frequency of a partial in the acoustic spectrum, the less energy it generally has associated with it in normal speech. Higher partials which are present in the voice signal may contain such little acoustic energy as to contribute very little to the acoustic signal. Articulatory settings and laryngeal settings can be drastically modified from the average speech settings to boost the high frequency energy. In order to see the total spectral content most spectrographs pre-emphasise the vocal signal multiplying the signal by a +6 dB per octave slope which balances out the

-6 dB per average slope on the speech. It is important, therefore, to know if the spectrogram shows a normal or a pre-emphasised signal. It is also important to use the correct algorithms and analysis window lengths depending on whether one wants to observe individual partial components or formant frequencies.

3.3.4 Inverse Filtering

Inverse filtering is a means of extracting the voice source waveform from either an acoustic signal from a microphone or from a pressure flow signal from a mask placed over the subject's nose and mouth (Rothenberg, 1973). In both cases, the voice source waveform is achieved by cancelling out the formants (resonances) of the vocal tract from the output signal thus in theory leaving the source signal. This presupposes that the formant frequencies and bandwidths can be found, and that there is no voice source-vocal tract interaction during phonation, that is, the voice source and the vocal tract are decoupled. This is true for the period of the vibrational cycle when the vocal folds are closed, but is not the case when the vocal folds open. This has to be taken into account in the measurements since not only is there the possibility of coupling between vocal tract and vocal folds if the open phases is sufficiently large, but also several phonatory types involve vocal fold settings whereby the folds themselves do not come into full contact with each other during any portion of the vibratory cycle.

The two methods complement each other when used together, as both have disadvantages which can be overcome by the other method. The Rothenberg mask is a device which records the oral air flow by measuring the pressure difference across a fine gauze over the mouth and nose. Inverse filtering of this signal provides low frequency information on the glottal airflow and can indicate how much leakage is present in the vocal folds. The inverse filtered signal contains spectral information on the voice source up to 1200 Hz due to microphone characteristics (Karlsson, 1986) and is a very useful tool for evaluating clinical voice function. Inverse filtering of the acoustic signal provides the necessary high frequency information (up to about 4000 Hz) and sufficient low frequency information for an acoustic analysis of the voice source.

3.3.5 Linear Predictive Coding

Linear Predictive Coding (LPC) is a particular inverse filtering method for representing speech waveforms as time-varying parameters related to the characteristics of the excitation and the transfer function of the vocal tract. The speech signal can be considered as the output of a linear system which consists of an input (modelled as a periodic pulse for voiced sounds, and noise for unvoiced sounds) which has been passed through a series of filters; the vocal tract, the glottis, and the lips.

The combined effects of the glottis and lip filters gives the source-radiation characteristic. In other words, the given signal is considered as the output of the dynamic system to which the input is unknown. A diagram representing the system is shown in figure 3.9.

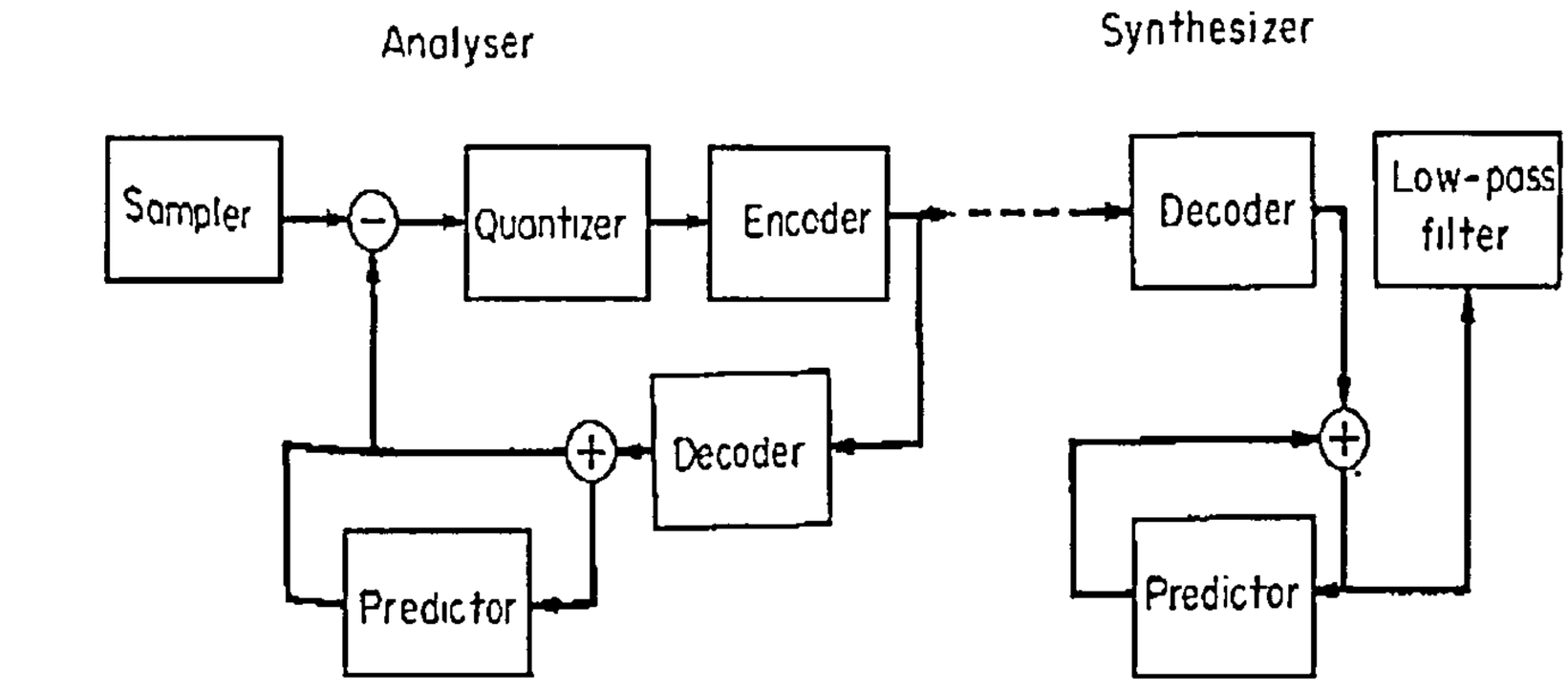


Figure 3.9. Block diagram of a predictive coding system (from Ainsworth, 1976).

The LPC models the speech signal as a linear combination of previous values. Model coefficients are generated by minimising the prediction error between real and predicted samples and recalculating these every 5 to 20 ms. Up to about 16 coefficients is sufficient to appropriately model speech. So an LPC speech sample can be defined as the predicted sample based on past values plus a prediction error. A fuller description of the mathematics behind LPC can be found in Makhoul (1975), Markel & Gray (1976), and Rabiner & Schafer (1978). It is an important tool in speech analysis and has many applications such as synthesis, formant frequency and formant bandwidth estimation, and spectral envelope determination. Figure 3.10 represents an lpc formant estimation for a female speaking the word "bard". The analysis bandwidth is set at 400 Hz.

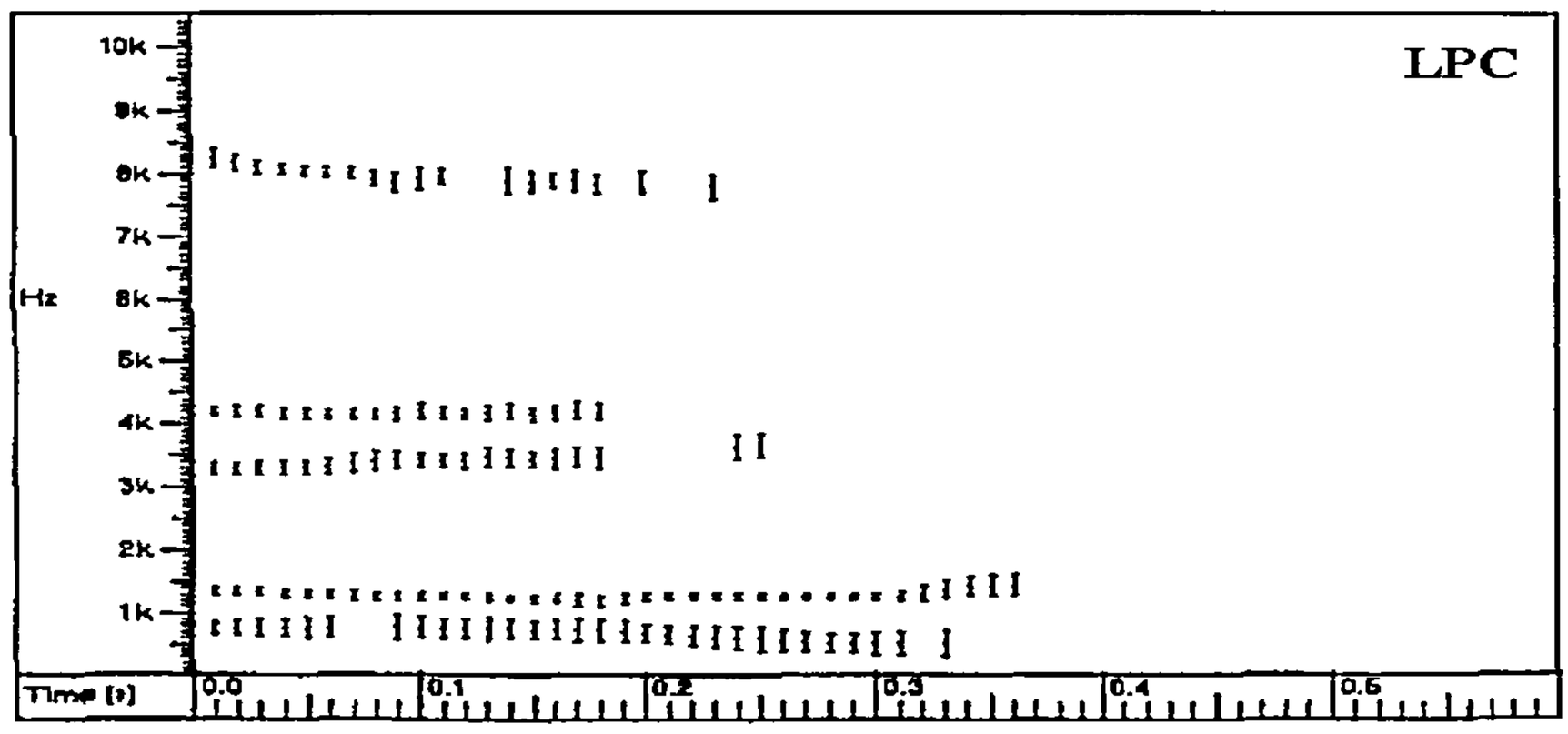


Figure 3.10. An LPC estimation of formant frequency and bandwidth of the word "bard" spoken by a female.

In the frequency domain, formant frequencies can be predicted by modelling the signal spectrum as a pole-zero spectrum. For modelling vowels, an all-pole model of the transfer function is required since the poles determine the resonances of the model tract.

Its advantages include simple algorithmic design, faithful spectral representation of the original signal which is derived directly from the analysis method, and automatic analysis (Rabiner & Schafer, 1978).

LPC analysis is prone to errors in estimation. In automatic formant detection methods such as LPC analysis it is crucial to model the overall spectral envelope of the original signal. If the analysis window is too short, only the pitch period will be detected. If the analysis window is too large, however, along with detecting the formant frequencies, it may also wrongly associate a partial as a formant and give spurious formant peaks which do not relate to original formant positions. Thus errors in formant estimation will occur if these spurious peaks are not eliminated. If the analysis window is reduced, it may eliminate the individual partial information but this could smear the formants bandwidths so making centre frequency location more difficult. LPC also has errors with the overall spectral gain and in formant amplitude estimation (Hughes, 1990).

Chapter 4

Vocal Qualities

4.1 Introduction

Voice quality is a perceptual attribute derived from a particular mode of production of the vocal system. This chapter gives an overview of the current issues involved in classification of vocal qualities in speech and singing. It is generally agreed that speech and singing registers/vocal qualities are separate though they may share common characteristics.

4.1.1 Voice Registers

A study of vocal quality in singing is not possible without first considering the controversial issue of voice registers. A consummate definition of voice registers, their terminology, number and mode of production has not been achieved, though there is now some agreement between voice specialists.

A useful definition of a register is that of "a phonation frequency range in which all tones are perceived as being produced in a similar way and which possess a similar voice timbre" (Sundberg, 1987). There appears to be two sources for vocal registers; the larynx and the supraglottal vocal tract (Hollien, 1983; Hollien and Schoenard, 1983). The phenomenon is common to both the speaking voice and the singing voice (Hollien, 1983).

Terminology is rather ambiguous, with some authors agreeing on two main registers for the male singing voice; modal register and falsetto (Sundberg, 1987; Wendler & Seidner, 1982) and others agreeing on three main registers (in order of increasing pitch); chest, middle, and head (Appelman, 1967); chest, falsetto and head (Garcia, 1840); and chest, head and falsetto (Vennard, 1967).

Catellengo et al. (1983) believes that the female singing voice also has two main registers; chest and head, though some believe that the female singing voice has three registers; chest, head, and an extra middle register with a register break at about C4 to E4 between chest and middle voice, and about an octave higher between middle and head register (Sundberg, 1987; Appelman, 1967). The middle register has not been proved or disproved scientifically but it is subjectively present (Hollien and Schoenhard, 1983).

Large compared the spectrums of tones sung in chest register and middle register belonging to female singing students singing at pitch E4. He observed that chest register contains more energy in the higher partials than middle register which is characterized by a stronger fundamental frequency (Large, 1968; Large, 1973).

It has been suggested that modal and chest registers have more energy in the higher partials than falsetto and head registers due to the vocal fold vibrations for modal and chest registers having a longer closed phase and steeper closing slopes, producing richer partials in the acoustic spectrum than for head and falsetto registers which tend to have fundamental frequency and lower partial dominance. For example, this has been observed in tenors (Hirano et al., 1989; Sunaga, 1971) and female singers (Large, 1968).

A myo-elastic theory for explaining register breaks is described in Titze (1989). It suggests that the muscles in the larynx behave like gears. An increase in frequency increases the tension on the vocalis muscle, and therefore the stress on the muscle fibres necessarily increases to the point of maximum physiological stress for the muscle at a certain fundamental frequency. The muscle has to "disengage itself" and release tension since it cannot sustain the stress required to produce higher fundamental frequencies.

Fundamental frequency ranges do not necessarily determine what register is being used. It is more a matter of mode of vibration of the vocal folds and the various muscle tensions associated with it which determines the type of glottal waveform being produced. Individual variations are great in terms of overlap of registers.

At the register transition, or "break", quality and pitch are detrimentally affected for one or more tones. The pitch positions of the register transitions depend largely on the singer's voice category. The lower the centre pitch of a person's vocal range, the lower the register transitions and vocal category to which he or she belongs (Sundberg, 1987).

Classical singing training enables singers to conceal register transitions where quality is affected, allowing them to "smooth" over the breaks and extend the range of their registers. This is carried out by adjustments in vocal tract shape, at the larynx, and subglottally.

4.2 Vocal Fold Vibration and Acoustic Quality

Differences in vocal quality arise from supralaryngeal settings, subglottal pressure and also different modes of vocal fold vibration. Vocal qualities in speech pathology are generally compared to ordinary healthy vocal quality which is called the modal voice (Hollien, 1974), described below.

4.2.1 Modal Voice

Hollien (1974) describes the modal voice as the phonation produced when the vocal tract is "in a neutral setting" and "no specific feature is explicitly changed or added". He states that "the modal register is so named because it includes the range of fundamental frequencies that are normally used

in speaking and singing (i.e., the mode)". It is generally agreed that the modal register corresponds to the "chest voice" register in singing (Appelman, 1973; Vennard, 1967), though some would argue that it also includes the "head voice" register (including Hollien). The physiological characteristics of modal voice for normal conversation (i.e., relaxed vocal tract producing low pitches) are as follows:

1. short and thick vocal folds
2. fully vibrating vocal folds (large amplitudes)
3. moderate muscle tension
4. no audible friction arising from incomplete glottal closure (Van den Berg, 1968).

Laver (1980) states that both the ligamental and the cartilaginous sections of the glottis function as a single unit, so the full glottis is involved, and vibration is efficient and periodic. He suggests that differences in quality arise from variations in laryngeal settings. His theory for vocal quality classification is described below. He suggests that five phonatory settings for speaking can be compared with modal voice. These are:

1. falsetto
2. whisper
3. creak
4. harshness
5. breathiness.

Laver (1980) divides the six phonatory settings into three categories which may or may not combine into compound phonatory types depending on their category.

In all, 20 laryngeal settings can be achieved. Modal voice and falsetto belong to the primary category. They can exist independently as simple phonation types or may combine with any other type in the second and third categories, but they are mutually exclusive.

The phonatory types belonging to the second category are whisper and creak. These may exist independently or combine with each other and with other types in the other categories.

The third category comprises of harshness and breathiness. These can only exist as compound phonation types. However, breathiness and falsetto are incompatible and therefore cannot combine. Other incompatibilities may also arise. Incompatibility arises from either redundant or conflicting acoustic requirements, or conflicting physiological requirements, which may arise from a "mutually exclusive specification of the phonatory settings involved in one or more of three muscular parameters concerned in adjustment of the vocal folds" (Laver, 1980). These are :

1. longitudinal tension - which is achieved by the vocalis and/or CT muscles
2. adductive tension - which is achieved by the IA and LCA which brings the arytenoid cartilages together closing the cartilaginous glottis and hence the ligamental glottis
3. medial compression - where tension of the LCA closes the ligamental glottis (Laver, 1980).

These tensions are represented in figure 4.1.

The five main phonatory types in speech which can be compared with modal voice are described below.

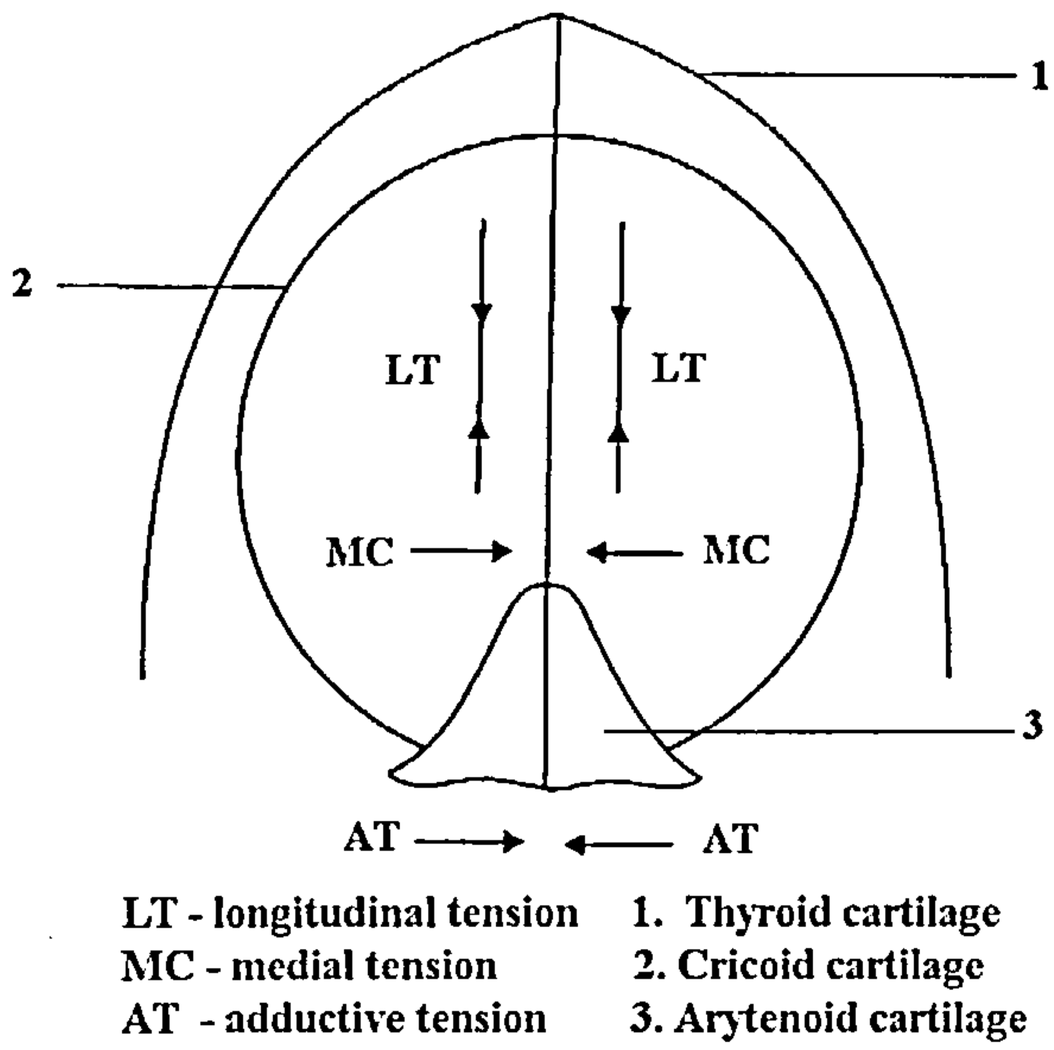


Figure 4.1. Geometric relationship between three laryngeal parameters (from Laver, 1980).

4.2.2 Falsetto

Modal voice and falsetto arise from completely different laryngeal settings (Laver, 1980). Falsetto has high tensions whereas modal voice only operates at moderate tensions (Hollien, 1971; Van den Berg, 1968; Laver, 1980). Falsetto is characterized by:

1. high adductive tension and medial compression. This arises from contraction of the IA and LCA muscles
2. high longitudinal passive tension of the vocal ligaments. The vocal ligaments are put under strong tension by the contraction of the CT muscle. This results in the vocal folds being maximally lengthened
3. only slight active longitudinal tension in the vocalis muscles, and they can generally be thought of as being relaxed along the glottal edge of the vocal folds. However, apart from the vibrating glottal edges, vocal fold mass is stiff and rather immobile. This is due to contraction of the outer vocal fold muscles, the lateral thyroarytenoid muscles. The result is a thin vertical edge to the vocal fold and a cross-section of the vocal folds shows them to be thin and triangular. The glottis is often slightly open, and the subglottal air pressure is often lower than in modal voice (Kunze, 1964).

Another important aspect of falsetto is that the average pitch range for male falsetto is higher than in modal voice, although there is some overlap. Hollien and Michel (1968) found it to be from 275 Hz to 634 Hz for male falsetto, whilst it was 94 Hz to 287 Hz for modal voice in males. The characteristically thin quality arises from this combination between high fundamental frequency and mode of vibration of the vocal folds. Sundberg (1987) puts the range of overlap between male modal and falsetto registers as in the region of 200 Hz to 350 Hz, which correspond to approximately pitches G3 to F4.

Several different types of falsetto can exist, depending on differing fundamental frequency control, glottal closure, and airflow rates. The untrained falsetto register exhibits a stable relationship between fundamental frequency and airflow. Fundamental frequency is determined by airflow, and an increase in airflow results in an increase in phonation frequency (Isshiki, 1964). This is not apparent in the falsetto register of trained singers.

One spectral feature distinguishing falsetto from modal is the amplitude of the fundamental relative to the higher partials. As seen in figure 4.1 the amplitude is 5 dB stronger in the falsetto register than in the modal register, and the difference between the fundamental and the second partial is also much smaller.

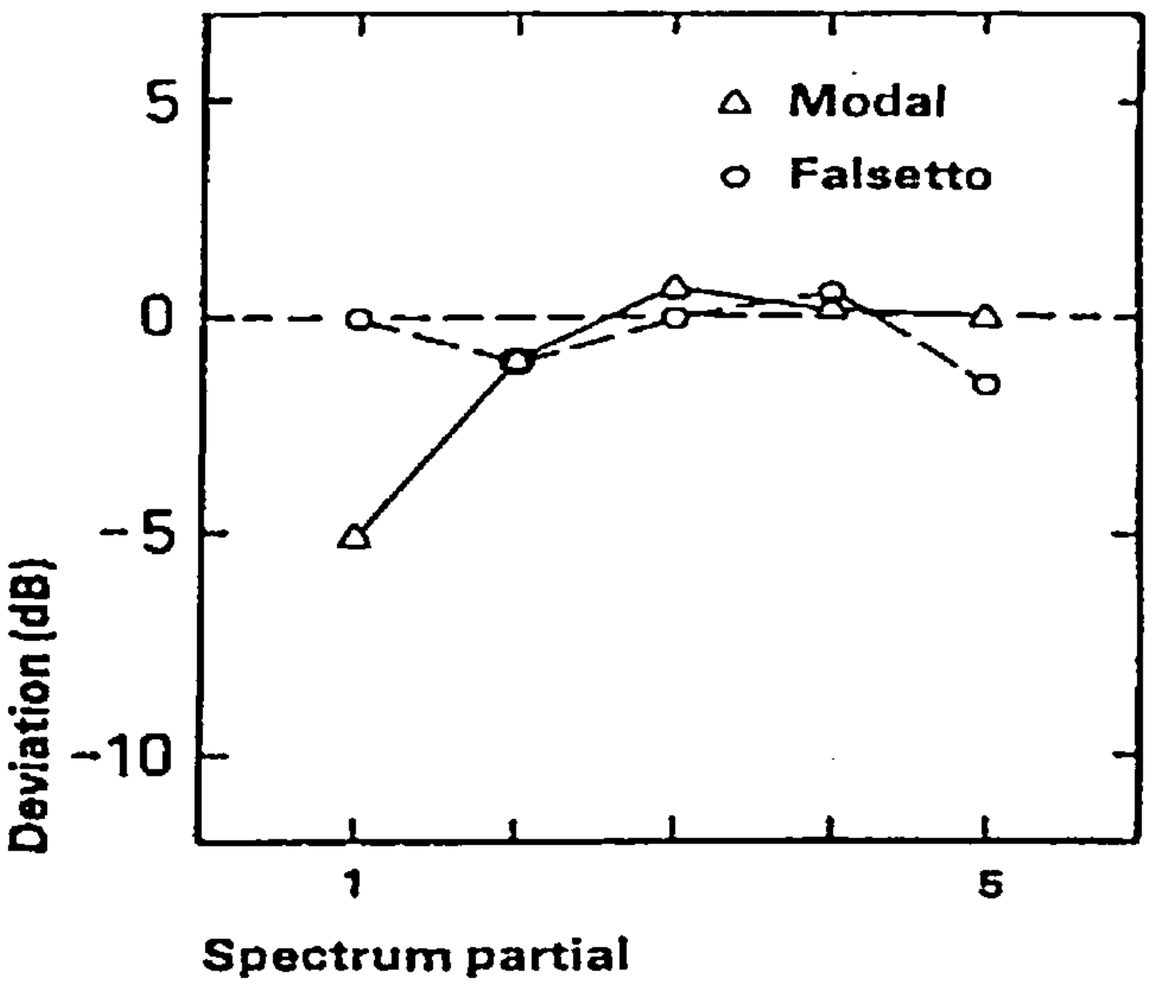


Figure 4.2. Average source spectra for three trained male singers pronouncing the vowel /α:/ on the same fundamental frequency in modal register (triangles) and falsetto register (circles). The spectra are represented by a curve showing how the spectrum contour deviates from the standard slope of -12 dB/octave (from Sundberg, 1987).

Also, in falsetto the spectral slope falls off more steeply (about -20 dB per octave) across the range than in modal voice, which falls off at about -12 dB, and gets steeper with increasing fundamental frequency (Monsen and Engebretson, 1977). The laryngeal waveform of the falsetto register has a steeper opening portion as opposed to that of the modal register which has a steeper closing portion.

4.2.3 Creak

Creak is characterised by very low fundamental frequency (in males, it ranges from about 30 Hz to 90 Hz) with an auditory effect of "a rapid series of taps, like a stick being run across a railing" (Catford, 1964). Wendahl, Moore and Hollien (1963) attribute this to a high dampening of the vocal tract between "glottal excitations". It is believed that this mechanical vocal tract damping arises from the action of the ventricular folds lightly coming into contact with the surface of the true folds. This explains the observation made by Moore (1971) that the closed phase of each cycle is lengthened.

The laryngeal setting of creak is not fully understood, but it is known that the vocal folds are adducted, thick and compressed; the ventricular folds are adducted; and

> "the inferior surfaces of the false folds actually come into contact with the superior surfaces of true vocal folds. Thus, an unusually thick, compact (but not necessarily tense) structure is created prior to the initiation of phonation" (Hollien et al., 1966).

Catford (1964) suggests that only a tiny portion of the anterior ligamental glottis is used. Control of fundamental frequency is different to modal voice, and the sub-glottal air pressure is lower than for modal voice. Cyclical vibration of the vocal fold is irregular, or aperiodic. Also, the spectrum falls off less steeply than in other types of phonation. Other synonyms for creak found in the glottal literature include vocal fry, glottal fry and laryngealization.

Lx waveforms of Laver's (1980) simple phonation types are shown in figure 4.3.

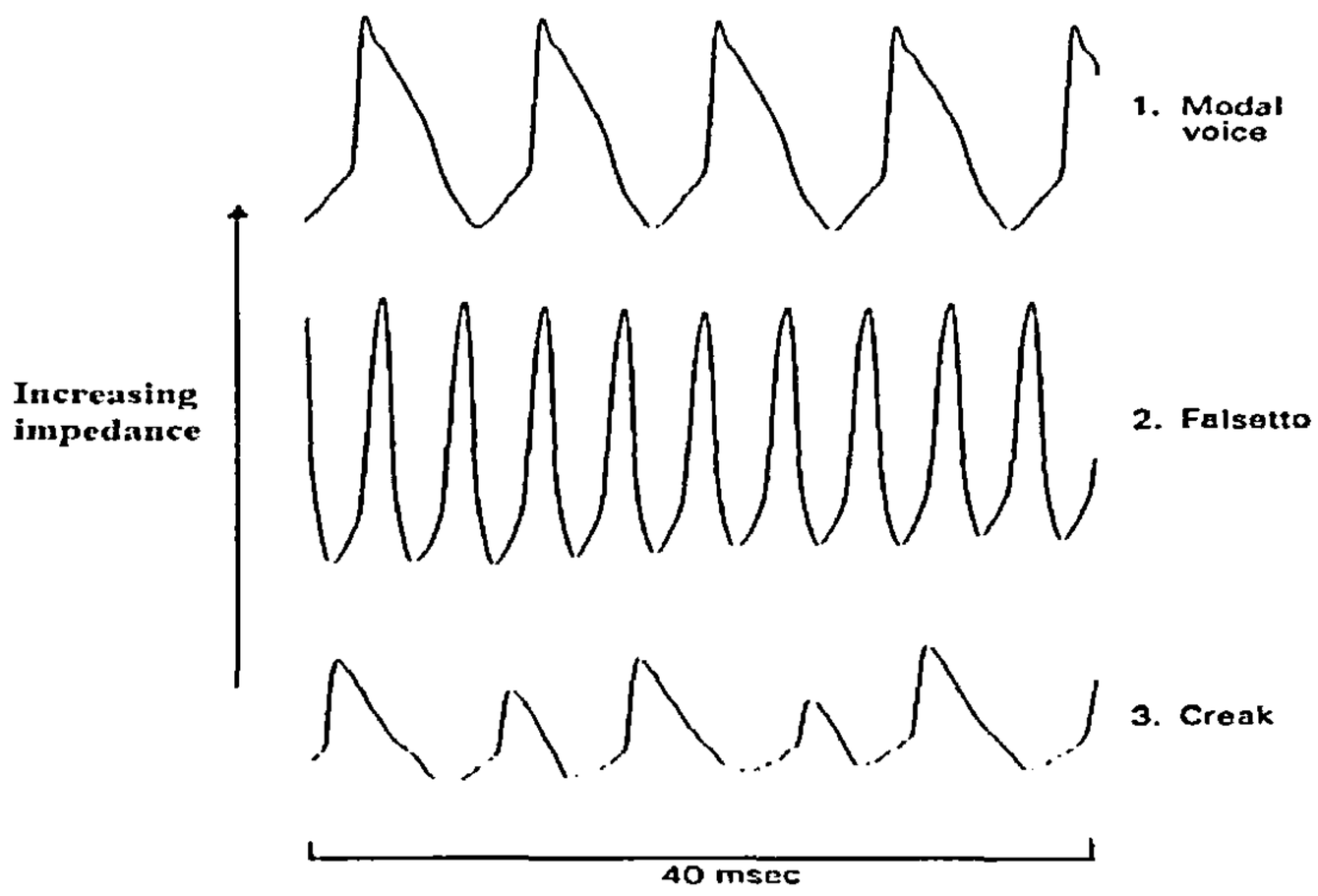


Figure 4.3 Lx waveforms of the simple phonation types (from Laver, 1980).

4.2.4 Breathiness

Breathiness is a quality combining modal phonation with inefficient glottal closure leading to a high amount of wasted breath. Catford (1977) describes breathiness as "the sound of voice mixed in with breath". The vocal folds narrow but do not close at any time. Breathiness is not limited to using a high rate of air flow. It is possible to produce a breathy quality with a low rate of air flow, as is the case in conversational speech at low frequencies. However, breathiness arises from an inefficient mode of vocal fold vibration, and as such, this small element of audible friction makes breathiness auditorily similar to whispery voice. However, whispery voice has a high amount of audible glottal friction. Physiologically, the two modes are incompatible since breathiness is produced with a low degree of laryngeal effort and whispery voice with its more constricted glottis, requires a greater amount of laryngeal effort (Laver, 1980).

The modal voice element in breathiness is always dominant. Breathiness is only compatible with modal voice since all other phonatory settings require too much muscular tension (Laver, 1980). Breathiness arises out of minimal adductive tension and minimal medial compression to enable the largish airflow to just cause the vocal folds to vibrate. Longitudinal tension is generally low, though it can be increased for the purposes of increasing fundamental frequency (Laver, 1980).

In terms of acoustic output, the general relaxation of muscles in breathy phonation contributes to the damping effect on the sound, leading to general energy loss and a broadening of the bandwidth of the first formant (Laver, 1980).

4.2.5 Whisper

Whisper arises from a partly constricted glottis. There appears to be a triangular opening of the glottis. For weak whisper part of the ligamental glottis may remain open. Increasing intensity results in an increase in glottal constriction until only the cartilaginous portion is open. This is the result of low adductive tension, with moderate to high medial compression. It uses a great deal of airflow, so is a very inefficient phonatory setting (Laver, 1980). The whisper characteristic is produced by "eddies generated by friction of the air in and above the larynx" (van den Berg, 1968).

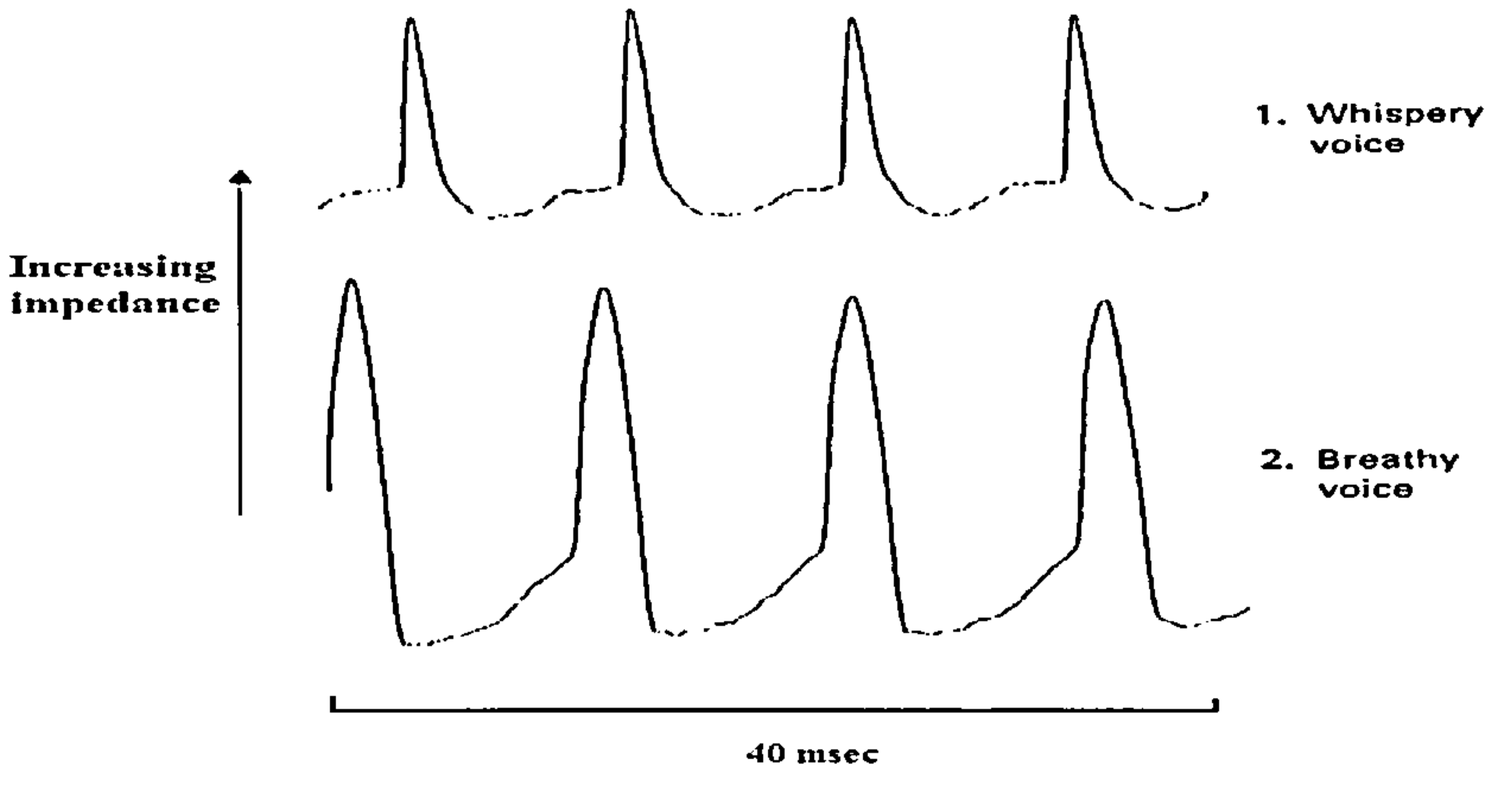


Figure 4.4. Lx waveforms of compound phonation types (from Laver, 1980).

4.2.6 Harshness

Harsh voice occurs in compound with other phonatory types and is characterised as including irregular and aperiodic spectral noise. As with creak, there are irregular cyclical vibrations in the action of the vocal folds. This is termed "jitter" (Cooper et al., 1957), and lends an auditory roughness, or rasping sound to the acoustic sound. Harshness sounds less severe in females than in males due to the higher fundamental frequency of females. This percentage deviation in fundamental

frequency due to jitter is therefore lower in females that in males (Hess, 1959), possibly explaining why harsh voice is perceived less commonly in women than in men (Laver, 1980).

Vocalisation is generally agreed to be initiated on a glottal attack and at lowish frequencies. Tensions from the extrinsic and intrinsic laryngeal muscles, that is both in the larynx and the pharynx, is high. Excessive tension in the vocal folds draws them too tightly together, and may account for the jitter and noise element in the acoustic output (Zemlin, 1964). This results in a louder intensity than in modal voice. Van Riper and Irwin (1958) suggest that "some of the apparent loudness may come from resonation effects due to the tenseness of the oral and pharyngeal cavities". Russell (1936) has observed that "as the voice begins to become strident and blatant, one sees the red-surfaced muscles which lie above the vocal cords begin to form a tense channel".

Laver (1980) suggests that the extreme tension in harsh voice is due to over-contraction of the muscles responsible for modal voice resulting in both excessive adductive tension and medial compression.

Figure 4.5 presents spectrograms of the six major phonatory categories defined by Laver.

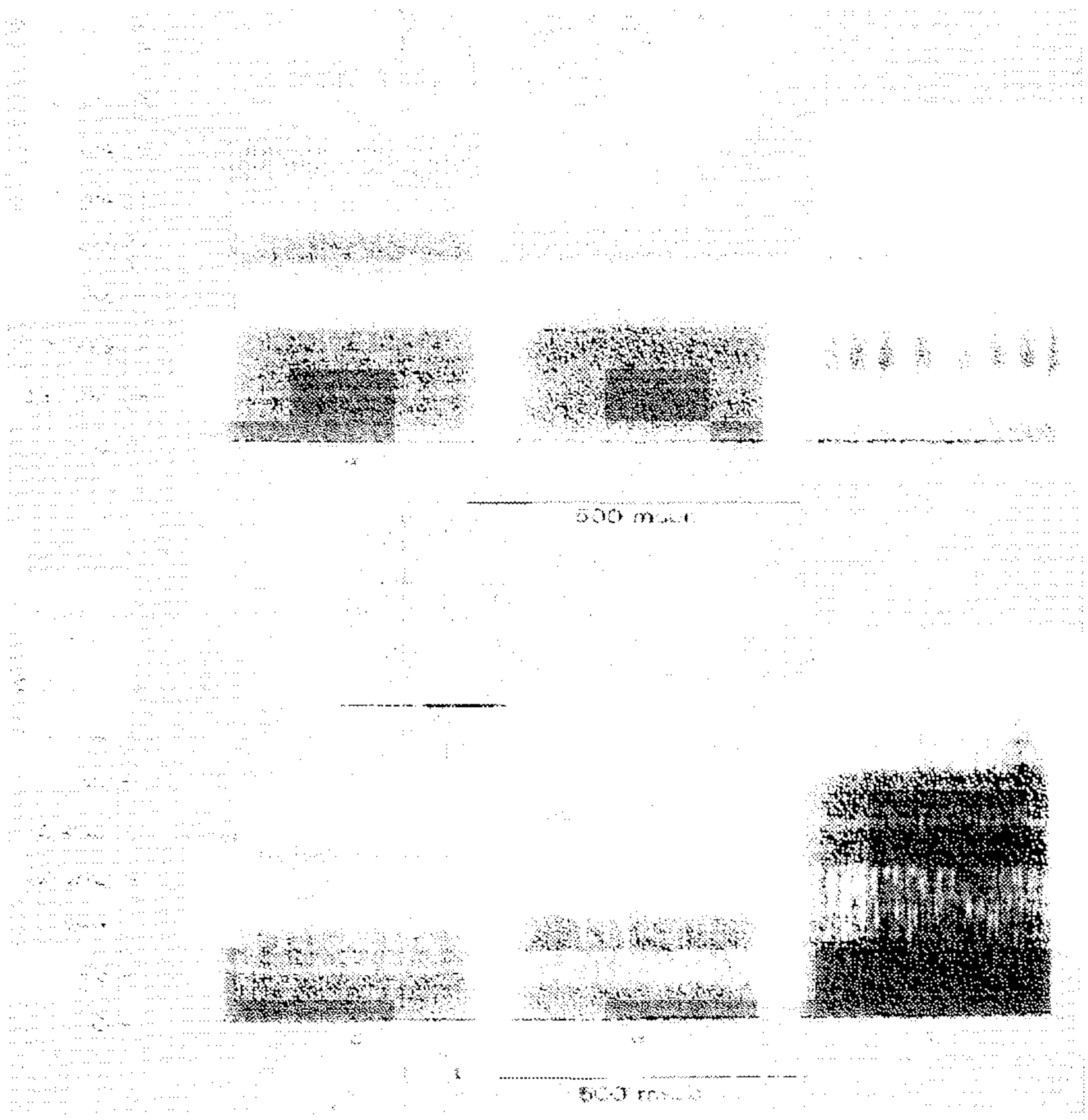


Figure 4.5. Spectrograms of steady-state vowels with six phonatory settings

a. modal voice
b. falsetto
c. creak
d. breathy voice
e. whispery voice
f. harsh voice

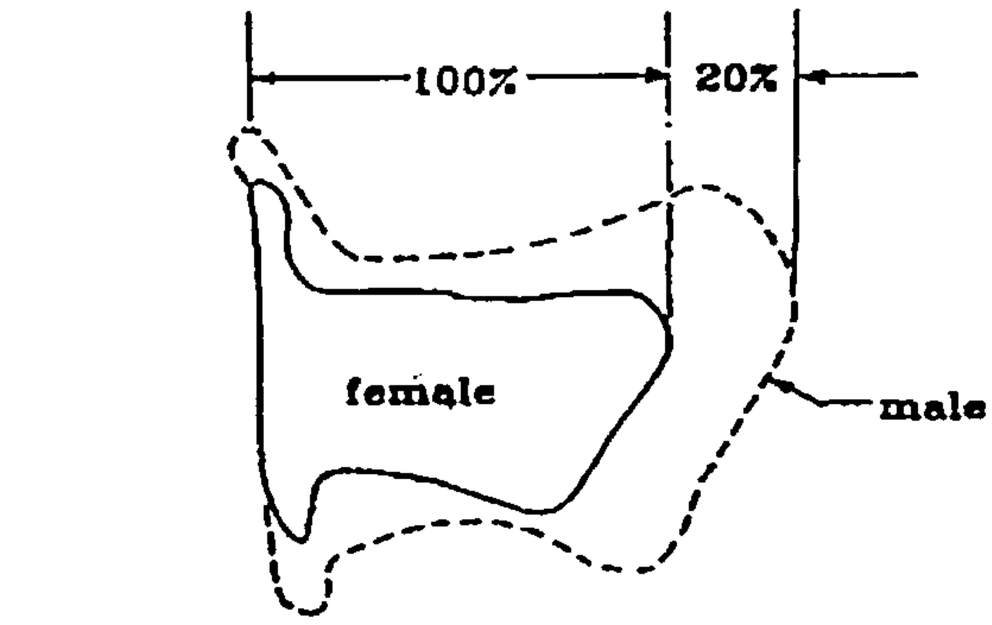
(from Laver, 1980).

4.3 Male and Female Voice Differences

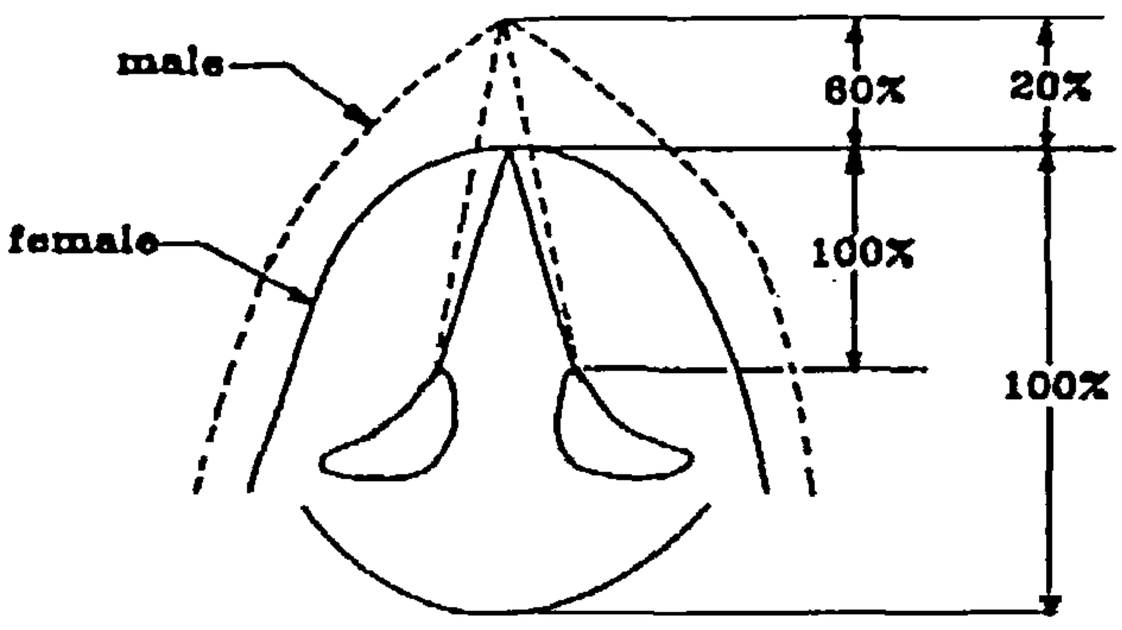
Adult male and adult female vocal tract dimensions differ in a number of ways. These are divided below into voice source (laryngeal) differences and supralaryngeal vocal tract differences. Differences between male and female vocal tract and laryngeal physiologies not only account for some of the male-female differences in voice quality, but also underlie some of the articulatory strategies used by professional singers.

4.3.1 Voice Source Differences

Voice-source differences between adult males and adult females can be attributed to the relative size of the larynx. The male larynx is 20% larger than the female larynx in all three planes: horizontal, vertical, and lateral (Kahane, 1978). However, in the anterior two-thirds of the larynx, the membranous vocal fold length is 60% larger in males than in females, shown in figure 4.6.



(a) Sagittal View



(b) Horizontal Section

Figure 4.6. Male-female comparisons of dimensions of the larynx. (a) Sagittal view of thyroid cartilage, (b) horizontal section showing difference in membranous length (from Titze, 1989).

Below is a list of main vocal fold differences which exist between adult males and adult females:

1. Vocal Fold Length and Fundamental Frequency

Vocal fold length changes as a function of fundamental frequency with the adducted vocal fold length in phonation always being shorter than the abducted length. There is a systematic lengthening of the vocal folds with increase in pitch (Hollien, 1960; Hollien & Moore, 1960).

2. Vocal Fold Thickness

Males have 20 % to 30 % thicker vocal folds than females (Hollien, 1960). Vocal fold thickness does not account for male-female differences in F0. Thickness and length have been observed to be inversely proportional (Hollien, 1960), proabably related to conserving tissue volumes (Titze, 1989).

3. Vocal Fold Tissue and Stress-Strain Curves

Female vocal fold tissue is slightly stiffer than male tissue (Titze, 1989) producing linear stress-strain curves, whereas the larger quantity of collagenous fibres present in the male vocal folds may account for the nonlinearity of the male stress-strain curve (Hirano, 1983; Fung, 1981). However, the difference between the two types is minimal.

4. Glottal Waveforms

The male glottal waveform is more asymmetrical than the female glottal waveform. This asymmetry is due to a slightly out of phase movement of the upper and lower parts of each vocal fold due to the increased male vocal fold size. Female vocal folds come into contact more as a single mass since they are relatively shorter and smaller (Titze, 1989; Monsen & Engebretson, 1977).

Male glottal waves approach the shape of female waveforms at higher fundamental frequencies (Monsen and Engebretson, 1977). That is, the glottal waveform differences between male and female speakers is the same as that between phonations at low and high fundamental frequencies in males.

The average female speaking voice is about -4 dB to -6 dB less intense than the average male speaking voice. However, due to a higher fundamental frequency of the female voice (about an octave higher), it has been predicted that the female voice should be 25% more efficient than the male voice (Schutte, 1980; Holmberg et al., 1988). The energy distribution in the female and male spectrum turns out to be equal due to the steeper tilt of the octave harmonics in the spectrum of the female glottal wave. This can be interpreted physiologically as the female having a much greater separation of the vocal processes than a male, that is having a posterior glottal "chink", or opening, in phonation. This too, is apparent in the glottal wave change which occurs when male speakers producing a rising glide of an octave (Titze, 1989). Monsen and Engebretson (1977) found that spectral envelopes fall off rather irregularly, but on average, the male voice spectrum has an initial tilt of -12 dB per oct, increasing to -15 dB per oct at higher frequencies. Females have a steeper slope. The only differences observed between the male and female glottal waveforms is a higher fundamental frequency and a larger open quotient in the female.

5. Breathiness from Incomplete Glottal Closure

One of the main acoustically perceived correlates of incomplete glottal closure is breathiness. It appears that the degree of perceived breathiness and incomplete glottal closure in female speaking voices is culturally determined, with English speaking females exhibiting higher breathiness with larger open quotient values than speakers of other languages (Klatt & Klatt, 1990; Sodersten & Lindestad, 1990; Karlsson, 1986) and conforming to subjective stereotypes (Henton & Bladon, 1985).

Both males and females can exhibit a visible posterior glottal aperture during the closed portion of a vocal period (Bless et al. 1986). In this study, this was observed in 20% of the males as opposed to 80% of the females. Holmberg et al. (1988) confirm this in a study of normal voices. They concluded that females have a more breathy voice than males, though both sexes have some breathiness when phonating vowels surrounded by voiceless consonants. They also found in the study that in normal voices, softer vocal effort is more breathy, with increase in open quotient and slow closure, and louder vocal effort is more laryngealized (pressed), with reduced open quotient. (Holmberg et al., 1988)

6. Glottal Models

Titze (1989) has drawn up both a static and a dynamic model of the glottis. The static model represents the prephonatory glottis, and the dynamic model represents the "time varying" glottis. This model contains two modes, a "horizontal string mode" and a "vertical ribbon mode". A phase delay quotient is built in because of the out of phase ribbon-like movement exhibited by the vocal folds. Figure 4.7 represents Titze's model showing this difference in settings between male and female vocal folds (Titze, 1989).

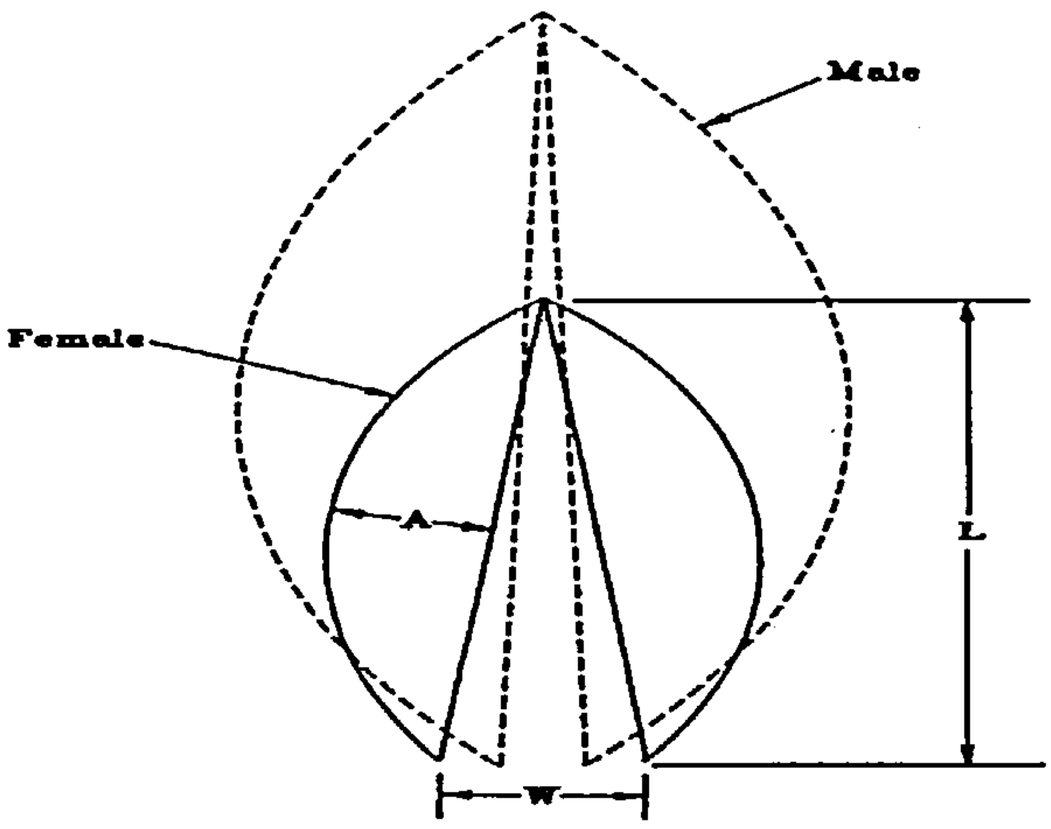


Figure 4.7. Scaling of the glottis in terms of length L, amplitude of vibration A, and separation of the vocal processes W (from Titze, 1989).

The main male-female difference in this model is the presence of a bulging factor in the medial surface of the male glottis, as opposed to the female glottis which does not bulge. A medial glottal

chink is present in the female model, allowing for continuous airflow. However, the presence of the medial bulge in the male glottis prevents a glottal chink from occurring. This bulging factor also accounts for the hump or "knee" in the opening phase of the simulated vocal fold contact area waveform (A_C) of the male model, represented in figure 4.8.

Titze (1989) believes that this bulging factor may be due to the greater contraction of the vocalis muscle in male speech. The female glottal flow (U_g) and female glottal area (A_g) waveforms also show a relatively longer open phase than the males. The open portion of the simulated female contact area waveform is also flatter, as opposed to the more rounded open portion of the male contact area waveform.

This is similar to the two mass model by Ishizaka and Flanagan (1972). In this model, a hump is generated in the opening phase of the glottal waveform when there is a low coupling between the upper and lower masses of the vocal fold. That is, an asymmetry is built into the model, as is the case in the male glottal waveform.

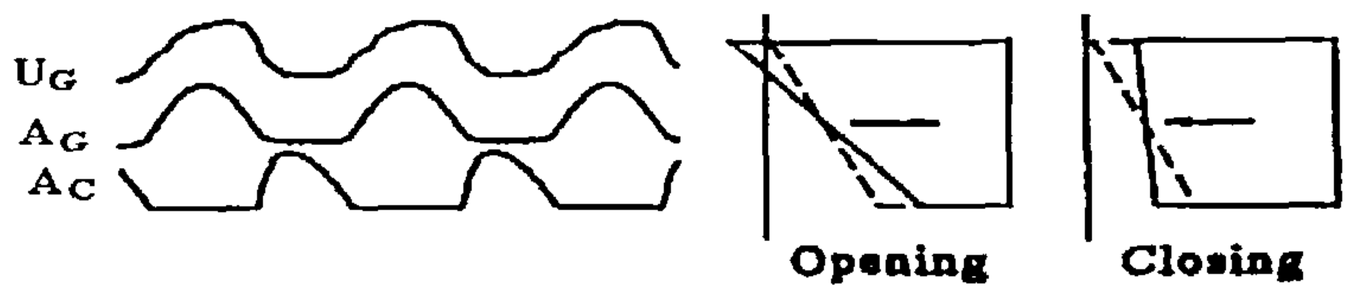


Figure 4.8. Differences in medial surface contour and corresponding glottographic waveforms. (a) Female-like with linear convergence and (b) male-like with medial surface bulging. A_C = vocal contact area, U_G = glottal flow, A_G = glottal area (from Titze, 1989).

An increase in the coupling between the masses results in the model becoming more symmetrical, similar to female glottal waves. Hirano (1975) has shown that for males the vocalis muscle is contracted far more than in females. He suggests that males tend to speak using a chest register whilst the female voice is more "falsettolike" in quality.

4.3.2 Supralaryngeal Differences

The relative size of the supralaryngeal vocal tract to the vocal folds can be related to some of the differences in vocal quality between males and females.

The dimensions of the average female vocal tract are not proportional to the corresponding dimensions of the male vocal tract. The average female adult has a mouth length which is 85% that of a male, but the average female pharynx is only 77% the size of the average male pharynx (Nordström, 1977). However, these values do not explain the actual formant frequency differences between males and females (Nordström, 1977). On average, women have higher formant frequencies than men. This is because of the smaller vocal tract lengths of women as comparedwith men (Sundberg, 1987). The higher the centre pitch of the vocal range the higher the formant frequency averages (Cleveland, 1977). Averaged across all vowels as a percentage, women's first three formant frequencies are higher than in men by 12%, 17%, and 18%, respectively (Fant, 1975).

Phonation frequency is perceptually much more important in identifying speaker sex than the three lowest formant frequencies (Coleman, 1976; Agren & Sundberg, 1978). The average female voice is about an octave higher than the average male voice. When a countertenor sings an alto part, which is associated with the modal voice range of a female, he sounds more like a female than a male, even though he has a male set of formant frequencies. The similarity to a female voice timbre may be explained by the voice source differences which are found between modal and falsetto registers in male singing (Coleman, 1976).

In the next section, the main physiological differences between male and female adult voices, discussed above, will be shown to be accountable for the various strategies used in the classical male and female opera voice, and differences between female opera and belting (Estill, 1992).

4.4 Opera Quality

The main feature which is common to both male and female opera singing is "formant tuning". Formant tuning allows the singer to project his or her voice. At any pitch, the relationship of the formant positions to the spectrum of the voice source accounts for many of the differences in the formant tuning techniques adopted by male and female operatic singers. The "singer's formant" is a crucial factor in the projection of the operatic voice. It is a parameter of the operatic sound mainly associated with male opera singers, though to a lesser extent can be observed in some female opera voices. However, the soprano adopts a different approach to vocal projection than the male. Rather than using a singer's formant, the soprano employs "pitch-dependent tuning" (Sundberg, 1987). This strategy arises as a direct consequence of trying to benefit from the harmonic effects of the first

formant which would often lie below the singing pitch, and therefore not contribute significantly to the spectrum of the sung tone. Different types of vibrato exist (Sundberg, 1987).

Vibrato is a phenomenon which is highly associated with the western operatic singing tradition. It seems to have become an integral esthetic parameter of the professional operatic voice, and develops as operatic singing training progresses even though it is not consciously learned (Bjørklund, 1961). Different types of vibrato exist (Sundberg, 1987), though the main attributes of opera vibrato are that it consists of a very low frequency quasi-sinusoidal modulation of the sung pitch at a rate of between 5 Hz and 6 Hz (Sundberg, 1987).

4.4.1 Male Opera Quality

The singing voice of Western male opera and concert singers exhibit a spectrum amplitude peak at around 3 kHz. This peak is known as the "singer's formant" (Sundberg, 1974). It arises from a migration/clustering (Sundberg, 1987) of the 3rd, 4th and 5th formants (Sundberg, 1974). By reducing the distance between these formant frequencies, the ability of the vocal tract to transfer sound is greatly increased in the region of these formant frequencies, shown below in figure 4.9.

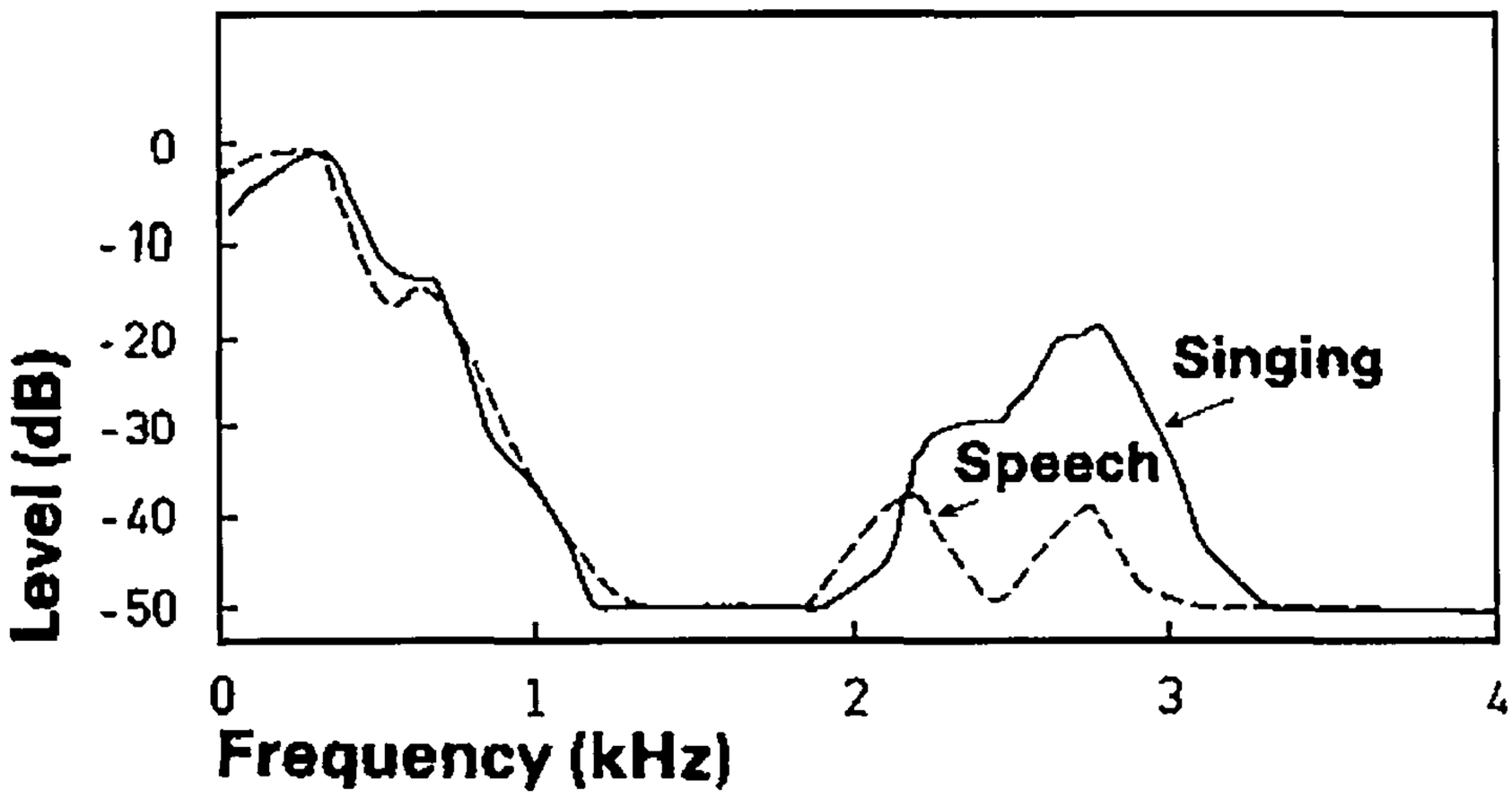


Figure 4.9. Spectrum envelopes for the vowel /u:/ as sung and spoken by a male professional opera singer. The peak in the spectrum envelope near 3 kHz is typical of all voiced sounds in singers except sopranos; it is called the singer's formant (from Sundberg, 1987).

A spectrum amplitude peak is generated, the amplitude of which depends on how apart the 3rd, 4th, and 5th formant frequencies are (Sundberg, 1987) and also on the amplitudes of those partials in the voice source. The amplitudes of the voice source partials indicate the rate at which the glottis closes. A low singer's formant amplitude is indicative of a low glottis closing rate.

The amplitude of the singer's formant is also dependent on other factors. One factor influencing the level of the singer's formant is loudness of phonation. A louder tone has stronger partials than a softer tone (Sundberg, 1987) resulting in a less steep spectral roll-off. It is expected, therefore, that, as figure 4.10 shows, the amplitude of the singer's formant in louder tones is greater than in soft tones.

This also applies to pitch. High notes are generally sung louder than low notes, and therefore, their voice source spectra has higher partials. Consequently, the singer's formant belonging to the spectra of high pitches is greater in amplitude (Hollien, 1983).

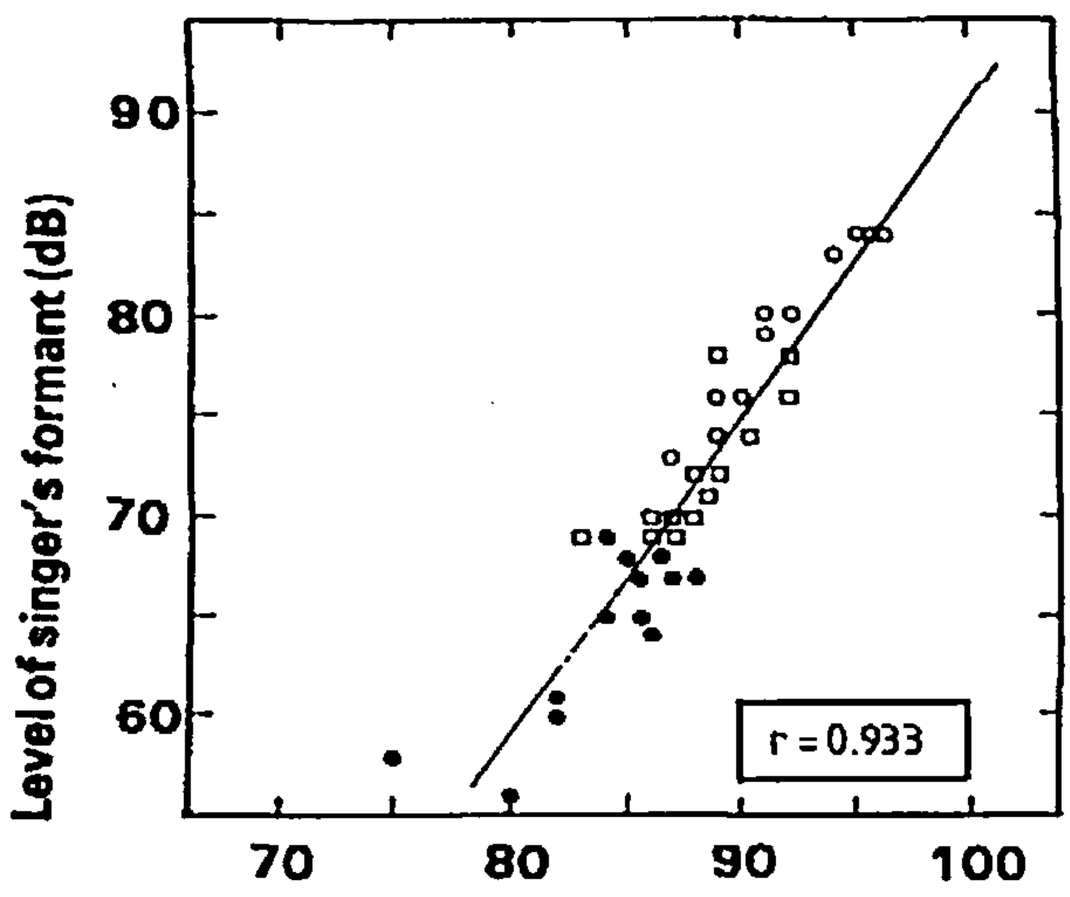


Figure 4.10. Sound level of the singer's formant in a baritone singing a chromatic scale on the vowel /ae:/ in soft (filled circles), middle (squares), and loud phonation (open circles). The gain in the level of the singer's formant is greater than the rise in the overall sound level (from Cleveland and Sundberg, 1983).

Formant frequencies depend on articulation. This implies that type of vowel will also determine the level of the singer's formant. Bloothooft (1985) found that in professional male singers, the singer's formant of vowels such as /i:/ and /e:/ which have high second formant frequencies is about 12 dB weaker than the SPL of the tone, but in vowels with a low second formant, e.g., /u:/ or /o:/ the singer's formant is about 20 dB weaker.

However, the singer's formant is independent of vowel articulation since it is always present in the professional singer's voice regardless of vowel (Sundberg, 1987). Rather, one way of achieving it is by lowering the larynx. Lowering the larynx lengthens the pharynx and the bottom part surrounding the larynx tube, the sinus piriformis, and widens both the sinus piriformis and the laryngeal ventricle, shown in figure 4.11.

Lengthening the pharynx increases the 2nd formant frequency of front vowels. Lengthening and widening the bottom of the pharynx and the laryngeal ventricle lowers the 4th formant frequency only when "the cross-sectional area in the pharynx at the level of the larynx tube opening is more than six times the mean of that opening" (Sundberg, 1987). This satisfies the conditions required for achieving a singer's formant, resulting in a grouping of 2nd, 3rd, and 4th formant frequencies. The 4th formant frequency can be lowered from about 3.5 kHz to about 2.8 kHz in adult males.

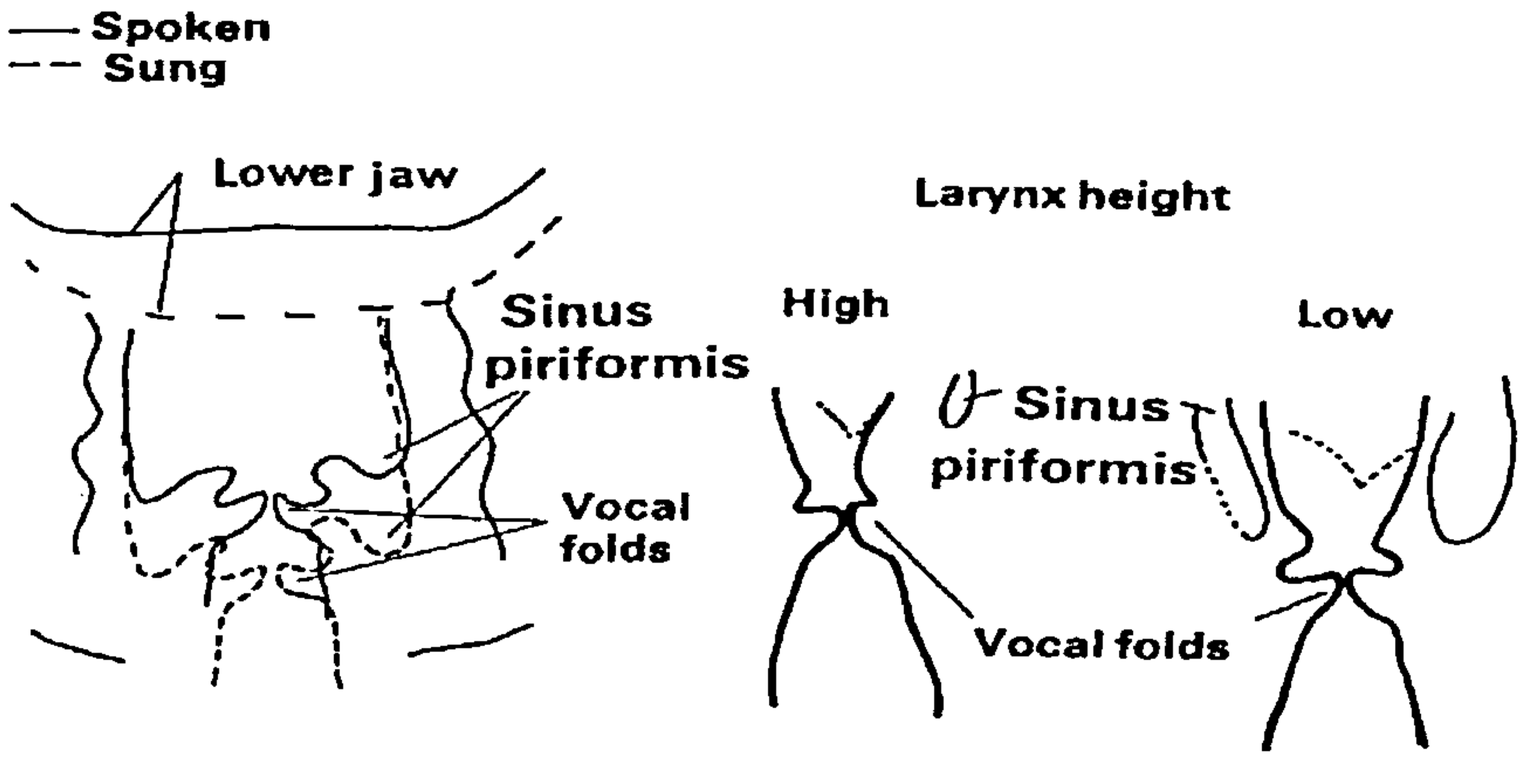


Figure 4.11. Left: Tracings of frontal X-ray pictures of a male singer singing and speaking the same vowel (solid and dashed contours). In singing, the larynx was lower and the piriform sinuses were wider. (from Sundberg, 1970, cited in Sundberg, 1987). Right: Contours shown in frontal X-ray pictures of the deep pharynx, when a subject deliberately raised and lowered his larynx. The laryngeal ventricle and the sinus piriformis expanded considerably when the larynx was lowered (from Sundberg, 1987).

There are several acoustical advantages which can be attributed to the presence of a singer's formant. One is that a voice can be heard over an orchestra since the loudest partials in the singing voice are found in the 3 kHz region as opposed to about 450 Hz in the orchestra or normal speaking voice. Also, the higher partials will be radiated directly towards the audience rather than be diffused upwards and to the sides of the singer.

Tuning of the two lowest formants to match partials of the voice source is also an established technique in classical singing. Miller & Schutte (1990) studied formant tuning in a professional baritone. They found that the baritone tuned his first two formants by modifying vowel articulation in order to amplify his voice. Accurate formant tuning of the first and second formants to voice source partials in the sung tone typically results in those partials having the highest sound pressure level. The first formant, followed by the second formant largely determine the overall SPL of a tone. As fundamental frequency increases, accurate formant tuning necessarily becomes more important as the distance between partials increases. This is relevant to female opera singing, described below.

4.4.2 Female Opera Quality

A greater difference in sound level exists between female singers and non-singers than between male singers and non-singers. Female singers have a greater maximum sound level than female non-singers. However, the singer's formant has a lower amplitude in female voices than in male voices (Seidner et al., 1983). This is especially so in sopranos (Bloothooft, 1985). The first formant frequencies for vowels having a narrow jaw opening are comparatively low.

The first formant frequency in the vowel /u:/ in a female speaker is about 350 Hz. However, female singers have to sing in excess of 700 Hz. At that pitch there are no partials to excite the first natural resonance of the vocal tract. In order to use to advantage the first formant frequency which would otherwise not be used, female singers, especially sopranos widen their jaw opening at higher pitches. This raises the first formant frequency up to, or close to that of the pitch. This tuning of the first formant to match the frequency of the pitch where the 1st formant would normally drop below the phonation frequency. This is shown in figure 4.12.

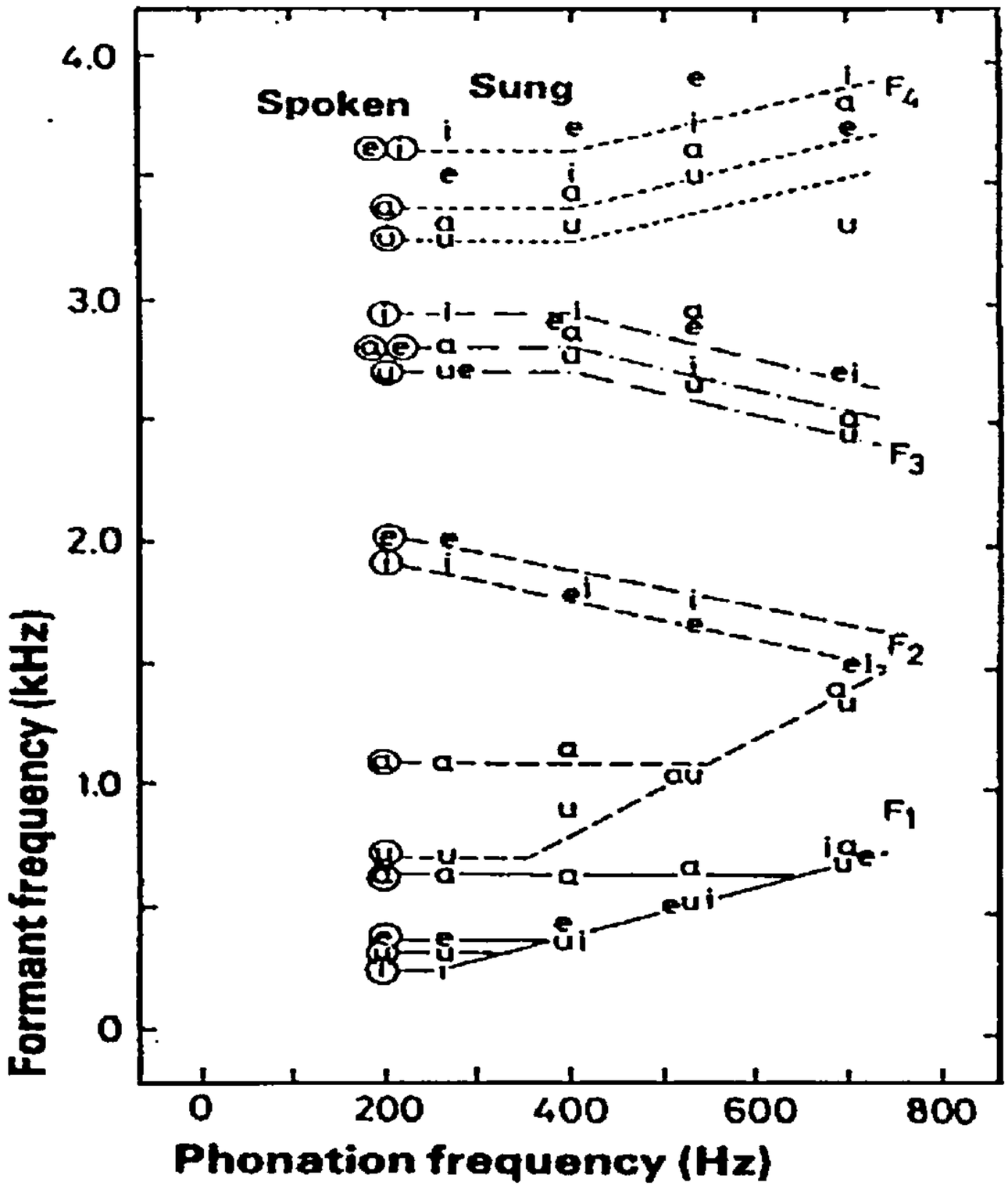


Figure 4.12. The four lowest formant frequencies (F1, F2, F3, and F4) in the vowels indicated used by a professional soprano singing at various pitches. The circled values pertain to the subject's speech. The lines illustrate the trends; the first formant is not allowed to be lower than the phonation frequency; the second formant of back vowels /u:/ and /α:/ rises and that of front vowels /e:/ and /I:/ drops with rising phonation frequency (from Sundberg, 1987).

The gain in amplitude can be as much as 30 dB. The skill involved in tuning the first formant frequency accounts for the difference in maximum SL reached between female singer and non-singers. This is a very economical way of singing. Sundberg (1987) states that "the pitch-dependant tuning of formant frequencies gives the singers' vowels a high loudness at a low price in terms of muscular energy".

Sopranos do not have to work as hard as altos to be heard over an orchestra, since they frequently sing above 450 Hz, the loudest partials in the orchestral sound, whereas altos have to compete in this region. According to Seidner et al. (1983), and Bloothooft (1985), altos use the help of a singer's formant, whereas it is much smaller in sopranos. Further investigation is needed in order to observe the articulatory strategies used by altos for producing their singer's formants. Figure 4.12 also shows the relative positions of the first four formant frequencies with rising phonation frequency for a professional soprano. With increasing phonation frequency:

1. in front vowels such as /i:/ and /e:/ the 2nd formant frequency drops;

2. in back vowels /u:/, /o:/, and /α:/, the 2nd formant frequency is tuned close to the 2nd partial;

3. the 3rd formant frequency drops after the phonation frequency of 440 Hz;

4. the 4th formant frequency rises after the phonation frequency of 440 Hz;

It also shows that the 1st and 2nd formant frequencies of the vowels studies are all very similar at the highest pitch. Johansson et al. (1983) found that at 960 Hz the tongue shapes for vowels /u:/ /α:/, and / i:/ are practically the same, as shown in figure 4.13.

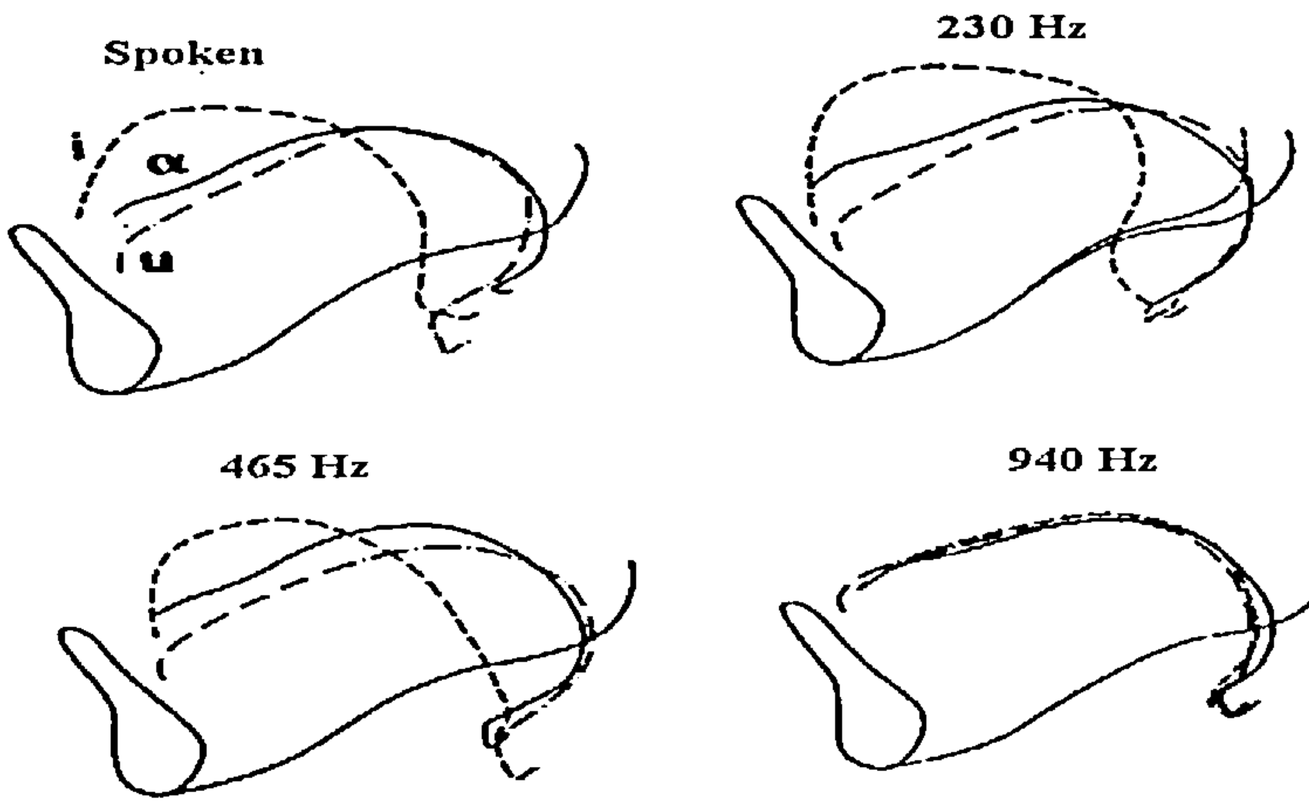


Figure 4.13. Mid-sagittal contours of the tongue body: dashed, solid, and chain-dashed curves pertain to the vowels /i:/, /α:/, and /u:/. The upper left family of contours are from spoken vowels; the others were sung at the phonation frequencies indicated..
The same tongue shape was used for all vowels at the top pitch (after Sundberg, 1987).

Other articulatory devices apart from jaw opening used by professional sopranos also include pitch dependent retraction of the corners of the mouth and also pitch dependent vertical positioning of the larynx. Raising the larynx increases the 1st formant frequency.

The effect of formant shifting is most apparent in high pitched singing where often projection of the sung tone is often achieved at the expense of vowel intelligibility (Sundberg, 1987).

Rothenberg (1985) has modelled a situation where efficiency of voice production for a soprano can be further enhanced by combining vocal tract tuning with sharp vocal fold adductive control.

Sopranos sing very loud high pitches with high subglottal pressures. In order to reduce the risk of vocal abuse, the vocal tract and larynx interact in order to maintain low average air flows whilst avoiding excessive vocal fold adduction, which will cause strain and fatigue (Rothenberg, 1985).

A reduction in airflow combined with the production of a harmonically rich tone can be achieved through the interactive effects of vocal tract tuning with almost complete vocal fold closure for a significant part of the glottal cycle (Sundberg, 1975). This excludes nazalization which appears to increase air flow (excessive air flows can lead to the risk of drying out the mucosal linings) possibly due to F1 damping which reduces the interactive effect (Rothenberg, 1985). Other mechanisms could be used for nazalized vowels or the efficient production of tones at lower pitches. A reduction in

average air flow is achieved through first formant tuning. This increases the amount of energy in the fundamental frequency therefore increasing the amount of interaction with F1. A harmonically richer tone is produced by keeping open quotient values for the glottal cycle above 50%. With OQ values greater than 50% the subglottal pressures become negative just as the vocal folds begin to open and close. This increases the glottal air flow at the beginning and end of the glottal pulse producing an air flow pulse with sharper onsets and offsets. The resultant spectrum has strengthened higher partials.

4.5 Belting

"Belting" has been described as "yelling set to music" (Yanagisawa et al., 1989), and can be heard in musicals, rock and gospel singing, and in much ethnic world music. Famous exponents of this technique, known as "belters", include Ethel Merman, Judy Garland, and Liza Minelli. The term "belting" is used synonymously with the term "belt" in the literature (see Schutte & Miller, 1993).

Several speech pathologists, doctors, and scientists have claimed that this quality is dangerous to vocal health (Lawrence, 1979; Osborne, 1979). It is a sad truth that with the demand for louder and louder singing, many rock singers and music theatre singers end up with vocal traumas such as throat strain, hoarseness, or even vocal nodes and ruptured vocal folds by either singing very loudly in the wrong register (such as falsetto in men), or by singing too loudly too low, in women (Howell, 1978). The term "belting" has had an ambiguous and confused history. One definition has described it as chest voice range extended upwards and over the break (Ruhl, 1986). Howell (private correspondence) uses the term "belting" to describe "the forced vocal muscles, pushed-air oversinging of anyone in any material, whether it's opera or rock".

The misunderstandings surrounding the quality are finally being dispelled through the scientific work of Jo Estill and her associates. It is now slowly becoming regarded by western voice scientists and voice pedagogues as a legitimate form of singing which can be undertaken in a healthy manner.

The following passages aim to describe the physiological and acoustic characteristics of belting thoough experimental evidence. Most of the papers compare the functional role of one or a group of related physiological components in a number of voice qualities, belting being just one of several qualities under scrutiny. Her work also challenges some of the strongly held beliefs about opera singing, and shows that there are physiological settings which are common to the production of both opera and belting qualities (Estill, 1992). Estill has worked her experimental results into a single theory for voice training, called "Voice Craft" which is rooted firmly in the understanding of the physiology of vocal production. It relies on the learning of a set of exercises called the "compulsory figures for voice". When mastered, these exercises allow singers and speakers control over individual parts of their vocal apparatus, overcoming the natural functions of the vocal tract, whilst enabling them to create any desired vocal quality effectively and efficiently (Kmucha et al., 1990).

4.6 Belting and Opera Quality Comparisons

This section relates belting to opera in the female singing voice by describing the differences between the qualities and then the similarities.

4.6.1 Differences Between Belting and Opera

Experimental evidence for the physiological differences existing between belting and opera are summarised below:

1. the activity pattern for extrinsic and intrinsic muscles differs between the two qualities (Estill et al., 1983; Estill, 1988).

2. the levator palatini, the muscle associated with the raising of the palate, is more active in belting than opera (Estill et al., 1983), but the genioglossus posterior, the tongue muscle, works harder in opera than in belting, possibly to create the "roundness" associated with the opera tone, since tongue compression darkens the sound (Estill, 1988).

3. in opera with "squillo" (the "ringing component"), the pharynx is wide laterally, and wide front to back because the tongue is compressed, whilst in belt, the pharynx is constricted (Estill, 1988;Sundberg, Gramming, & Lovetri, 1993).

4. in belting the larynx is higher than in opera, resulting in higher first formants (Schutte & Miller, 1993). In opera quality with "squillo", the larynx is pulled in two directions - up for the twang quality in order to constrict the AES which provides the "ringing" quality, and down, to provide the width required for the depth in the quality of opera. It is at a relatively neutral to mid-high setting (Estill,1988). For other schools of opera singing in the middle range, the larynx is stabilised in a comfortably low position, resulting in low first formants (Schutte & Miller, 1993). Lowering the larynx, as in opera, creates the perception of a fuller and darker sound by creating a frequency spectrum with odd partials at higher amplitudes than the even partials and by lowering the 2nd formant frequency (Yanagisawa et al., 1990).

5. the greater effort required to belt reduces the vibrato (Schutte & Miller, 1993).

6. belting exhibits far more glottal adduction across the range than any other quality, (Estill 1988; Sundberg, Gramming, & Lovetri 1993, Schutte & Miller 1993). Glottal adduction can be linked to closed quotient measures, described in chapter 3 (Abberton et al., 1989). Estill (1988) suggests that different vocal qualities have associated with them characteristic patterns of larynx closed quotient across the vocal range. Belting tends to have a consistently high CQ value, whilst for female opera quality in the middle register, the value is low for the lowest pitches and increases as pitch rises. This is shown in figure 4.14.

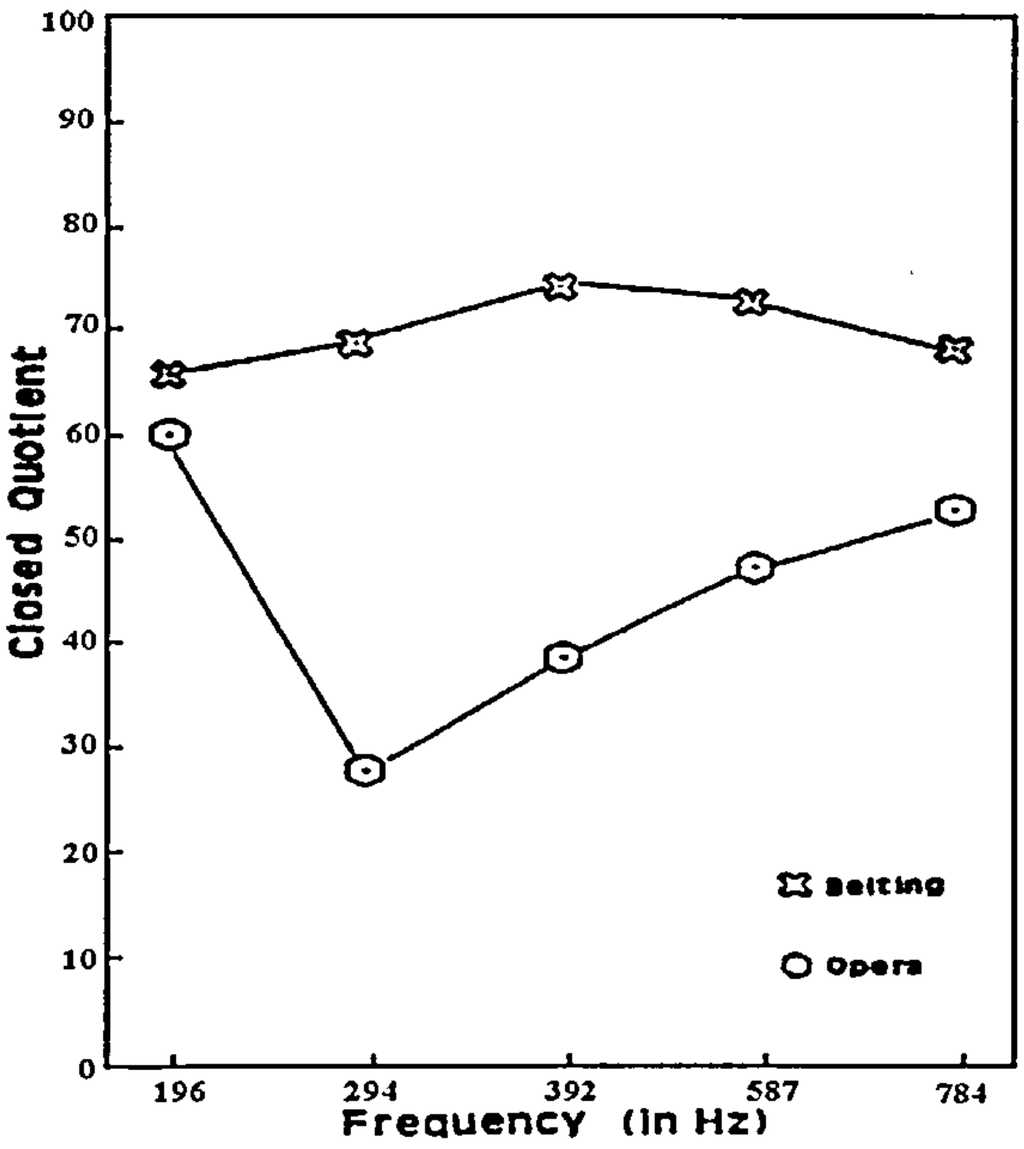


Figure 4.14. A comparison of the closed quotients of opera and belting in a female singer. Data points represent the average of all tokens recorded for each condition (after Estill, 1988).

7. except for the tongue muscle, muscle activity is higher in belting than for opera in the head, neck and torso, and increases with rising pitch, (confirming the observation that belting is a far more strenuous quality than opera quality to achieve (Estill, 1988)). One explanation for the higher larynx position, the greater adductive effort and the greater subglottal pressure required to belting is that it is in order to raise the frequency of the first formant up to the frequency of the second harmonic .This results in "a loud sound with a bright, somewhat harsh quality that conveys the excitement of high tension" (Schutte & Miller, 1993).

8. belting has been shown to contain high energy in the upper partials above about partial 8; much higher than in opera quality, as seen in figure 4.15 (Yanagisawa et al., 1990; Estill, 1988).

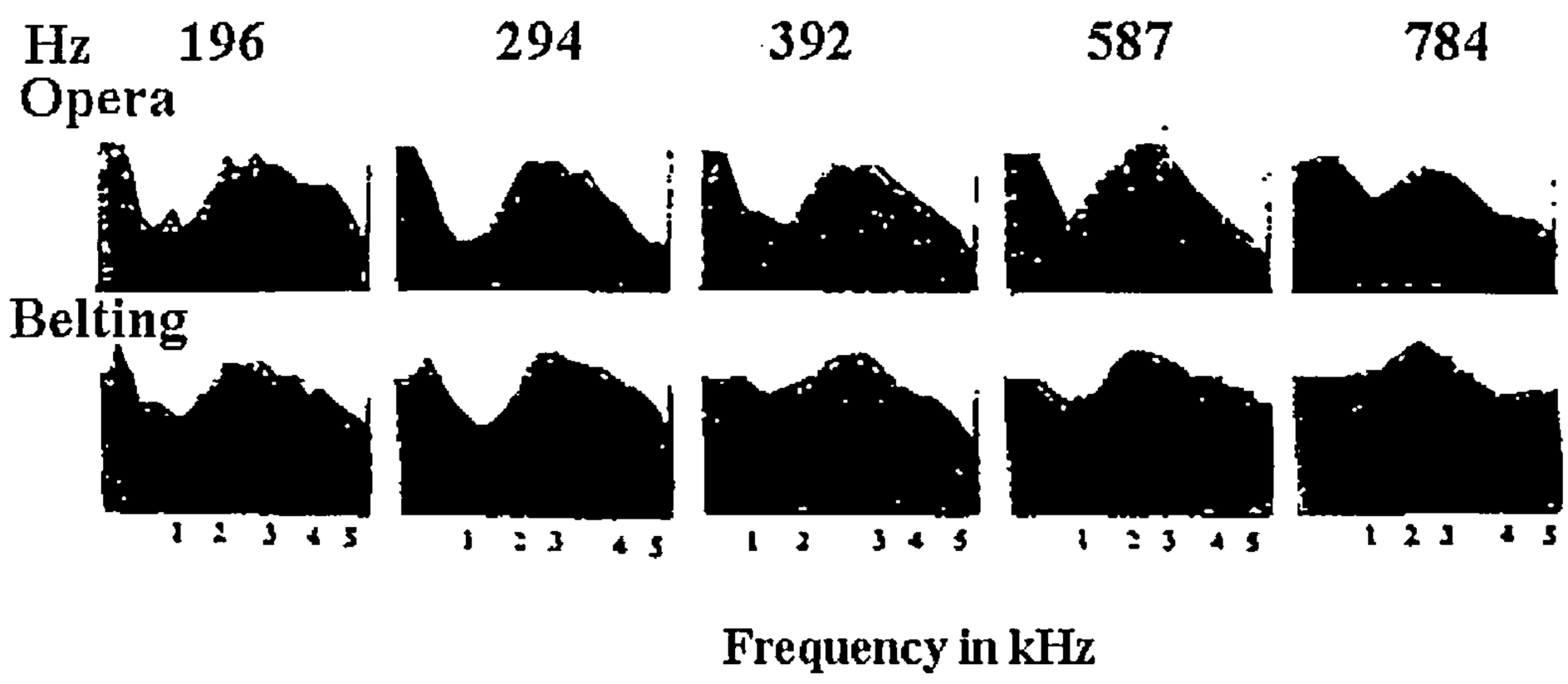


Figure 4.15. Comparison of average spectra for opera and belting at five frequencies (top line) (after Estill, 1988).

4.6.2 Similarities Between Belting and Opera

Belting has been shown to share several physiological characteristics with opera with "squillo", or "ring" (as heard in the Italian tradition). High intensity vocal production in both belting and opera can be achieved with:

1. the tight constriction of the aryepiglottic sphincter (AES). This results in "twang", which is typically heard in the country and western voice (Yanagisawa et al., 1990) and contributes to the ringing quality known as the "singer's formant" (Sundberg, 1974) by increasing the amplitude of the spectral partials in the region of 3 kHz (Yanagisawa et al., 1989; Yanagisawa et al., 1990). Constriction of the AES creates an extra resonator from the rim of the aryepiglottis to the vocal folds. The small size of this resonator accounts for the amplitude increase of the partials in the 3 kHz region (Yanagisawa et al., 1989). The twang sound is most perceptable in an oral twang /i/ where the amplitude of adjacent partials from 2-3 kHz is nearly equal. Oral twang is a main component of oral belting which exhibits an amplitude plateau of spectral partials from 2 kHz upwards to 4 kHz (Yanagisawa et al., 1990);
2. the avoidance of harmful endolaryngeal constriction, i.e., the constriction of the ventricular folds and vocal folds (Kmucha et al., 1990);
3. the high pressed tense tongue position (which has been shown to not interfere with the vocalis muscles, as thought of previously (Estill, 1983);
4. a raised larynx (for opera with "squillo");
5. an increase in supralaryngeal muscle activity with increasing fundamental frequency. Geniohyoid muscle activity (the geniohyoid muscle is found under the jaw and inserts into the hyoid) is highest at all frequencies for both qualities (Estill, 1983);
6. good posture (Estill et al., 1983).

Schutte & Miller (1993) believe that there are performance related reasons why nonclassical singing (of which "belt" is a part) differs from classical singing:

1. the texts of the songs have a more important role in nonclassical music than in classical music. Since texts should be understood, vowel modification is minimal;
2. vocal individuality and naturalness are valued more than in classical singing;
3. songs maybe adapted to the "strengths and weaknesses of the individual voice and temperament" (Schutte & Miller, 1993).

Figure 4.16 compares the spectra of "classical" and belting on the same note. Their definition of belting is quoted below:

> "Belting is a manner of loud singing that is characterized by consistent use of "chest" register (>50% closed phase of glottis) in a range in which larynx elevation is necessary to match the first formant with the second harmonic on open (high F1) vowels, that is ~G4-D5 in female voices" (Schutte & Miller, 1993).

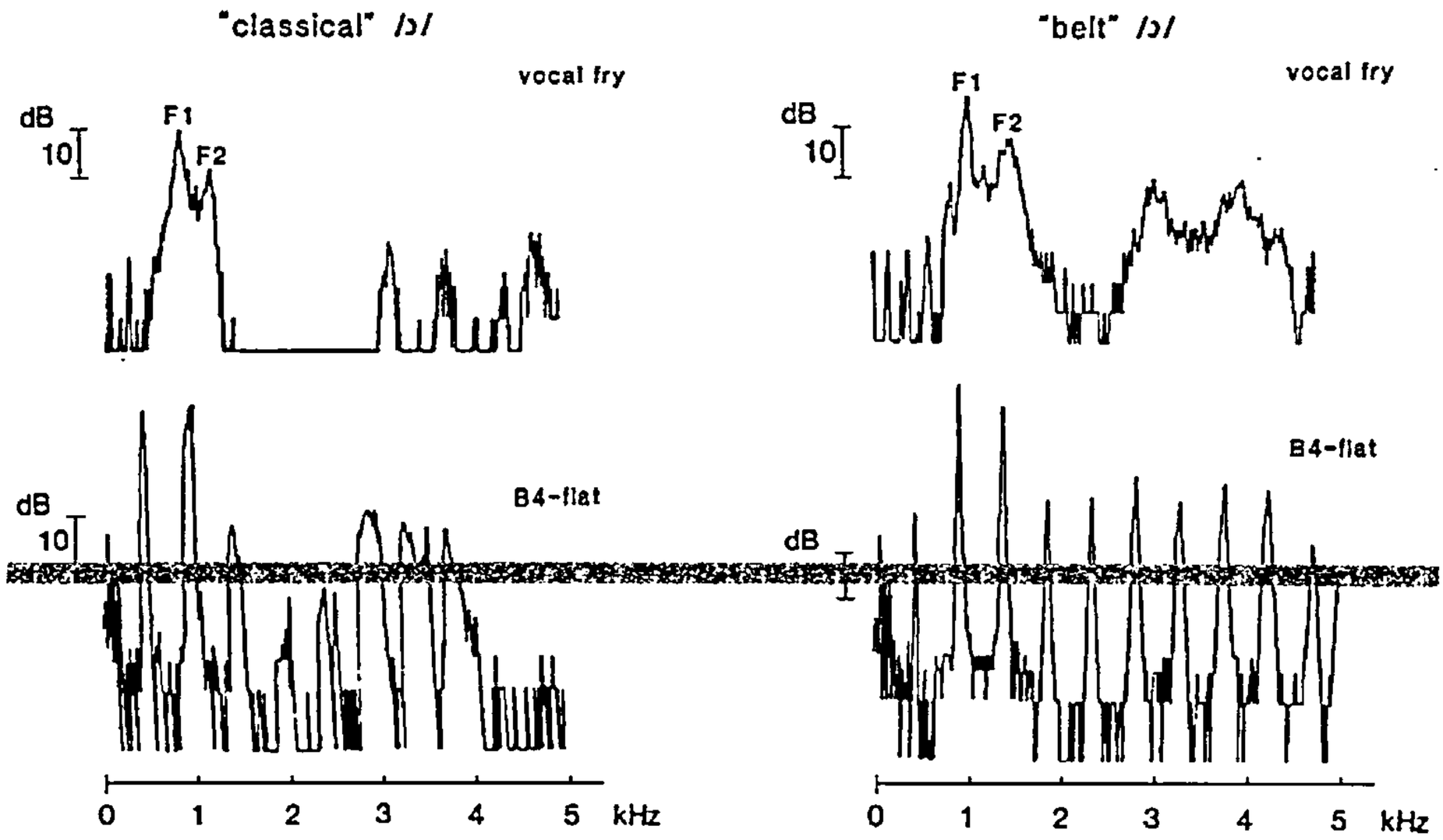


Figure 4.16. Spectrograms sung in "classical" mode (left) and "belt" mode (right) on the same pitch. For the classsical tones, "both F1 and F2 are low and the first harmonic is also prominent". The perceptual result is a "round" and "dark" tone. For the belt tones, which Scutte & Miller equate with being sung "in "chest" register (> 50% closed phase), F1 and F2 are higher, with F1 following the second harmonic. Vibrato is diminsihed and higher-frequency component is increased. Perceptually this is a loud, bright, "edgy" sound" (after Schutte & Miller, 1993).

4.6.3 Larynx Height Differences

Vertical larynx position, called larynx height naturally varies in speech. For example, larynx height rises considerably for an exclamation. Varying the larynx height changes the shape and length of the vocal tract leading to timbral differences and vowel differences. Therefore varying the larynx height would be expected to contribute to the production of different singing qualities.

Raising the larynx not only shortens the pharynx, but also constricts the lower part of it, by causing the tissues of the side and back walls to compact together. This occurs by constricting the lower and middle pharyngeal muscles which run from the cricoid and thyroid cartilages and the hyoid bone around and upwards to the back wall. Lowering the larynx stretches the pharynx walls which widens the pharynx (Sundberg, 1987).

Larynx height influences the voice source. A raised larynx position has been associated with an increase in adduction of the vocal folds perceived as pressed phonation, and a lowered larynx height has been perceived as flow phonation which is generally associated with good opera quality technique (Sundberg & Askenfelt, 1983). Raising the larynx can also stiffen, stretch and thin the folds (Shipp, 1977). All these will affect the voice source (Titze, 1988; Gauffin & Sunberg, 1989).

Differences in larynx height position will also change vocal tract length, which determines formant frequencies (Sundberg & Nordström, 1983). The combination of changes in formant

frequencies and voice source will change the voice and vowel timbre. If larynx height is raised both vowel timbre and voice timbre brightens and if it is lowered timbre darkens (Sundberg & Askenfelt, 1983). In speech and singing, perceptually the vowel [u:] produced with a low larynx, is perceptually a lot less bright than [i:] (produced with a high larynx) and [a:] (Wang, 1985). Larynx height appears to be vowel dependent regardless of singing style (Wang, 1983; Pabst & Sundberg, 1992; Seidner et al., 1983).

Generally, the smaller the vocal tract, the higher the voice tessitura. Tenors have smaller vocal tracts than basses and hence have higher formant frequencies. The percentage difference in formant frequencies between male and female voices is comparable to those found between tenors and basses (Cleveland, 1977), and those between raised and lowered larynx height (Sundberg & Nordström, 1983). Lowering the larynx slightly reduces the distance between the 3rd and 4th formant frequencies across vowels. In the study by Sundberg & Nordström (1983) the 4th formant frequency averaged across vowels reduced by 17 % as opposed to 11 % for the 3rd formant frequency when subjects were asked to phonate from a high to low larynx position.

Different schools of singing advocate different vertical larynx positions for correct qualities. There are two strands of thought concerning larynx height and correct singing. The traditionalists believe that for good operatic singing the larynx should be anchored at or below the larynx rest position throughout the singing range since this contributes to the production of the singer's formant and minimises possible variations in vocal quality across pitch due to changing formant frequencies (Shipp, 1974; Sundberg, 1974). It also reduces excessive vocal fold adduction. A low larynx height changes the formant frequencies of the vowels (Sundberg, 1987). For example, a vowel pronounced with a low larynx position becomes similar to the vowel /oe:/ (Sundberg and Nordström, 1983).

On the other hand, several recent studies have shown that professional opera singers do not necessarily maintain a low larynx height throughout their pitch range, but allow it to rise with pitch without causing any harm to the voice and with little disruption to vocal timbre or the intensity of the singer's formant (Johansson et al., 1983; Pabst & Sundberg, 1992). This is shown in figure 4.17.

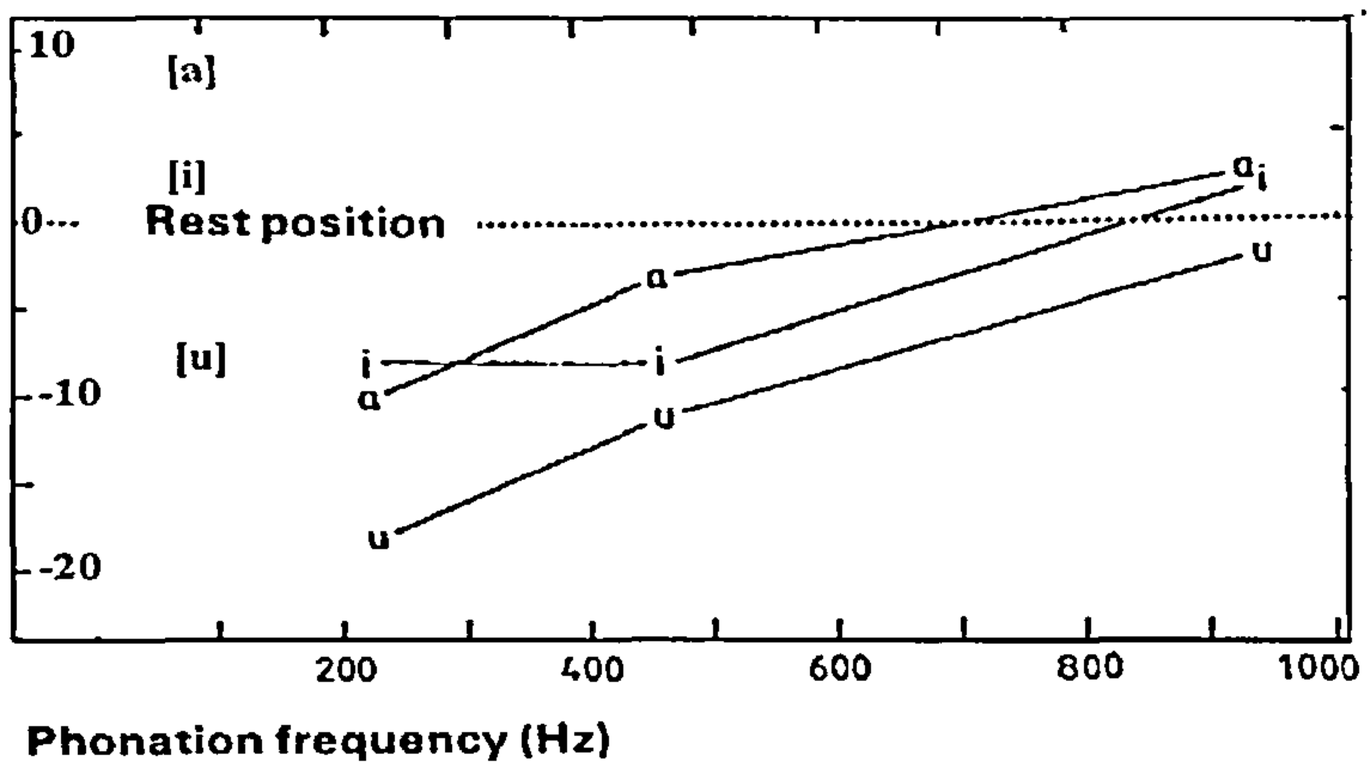


Figure 4.17. Vertical larynx position observed in a professional soprano. The bracketed symbols refer to speech; the unbracketed to singing (from Sundberg, 1987).

Yanagisawa et al. (1989) argue that the ring produced by a lowered larynx, as is the convention in operatic singing is harder to achieve since lowering the larynx "masks the tone", making it darker and softer.

Studies on singing styles other than opera quality have also shown that they are also produced with a raised larynx. These include belting, discussed above (Estill, 1988; Yanagisawa et al., 1989), and a Scandinavian female herding singing style called Kölning which shares the same pitch range as sopranos (Johnson et. al., 1983). Strategies are used to minimise the risk of vocal abuse. For example, first formant tuning by pitch dependent jaw widening has been observed in Kölning, a method shared by operatic sopranos to increase the SPL with the minimum of effort (Sundberg, 1982). As with operatic singing, Kölning is reported to be efficient for its function, vocal economy being the key. Kölning is usually sung in shortish bursts, therefore the very high subglottal pressures and elevated larynx position are not a possible vocal abuse problem (Johnson et. al., 1983).

A study of tenors by Wang (1983) has also shown that Chinese and Western early music is sung with an elevated larynx which rises with increasing pitch. This was as part of a study to show that bright timbre can be produced in different singing styles with similar spectral features at differing larynx heights.

Chinese and early music styles were compared to Western operatic style. Larynx height remained below the larynx rest position and decreased with increasing pitch for the operatic style only. Perceptually all styles were considered bright, though the Chinese style was considered to be the brightest. For all three singing styles spectral peaks were found in the region between 1.8 kHz and 3.8 kHz, termed the "Bright Timbre frequency Range", or BFR, (seen in figure 4.18, comparing Chinese voice and Western Operatic voice spectra) and the relative amplitudes of these peaks increased with increasing fundamental frequency. However, the formant positions for the Western operatic voices were lower than the other two styles.

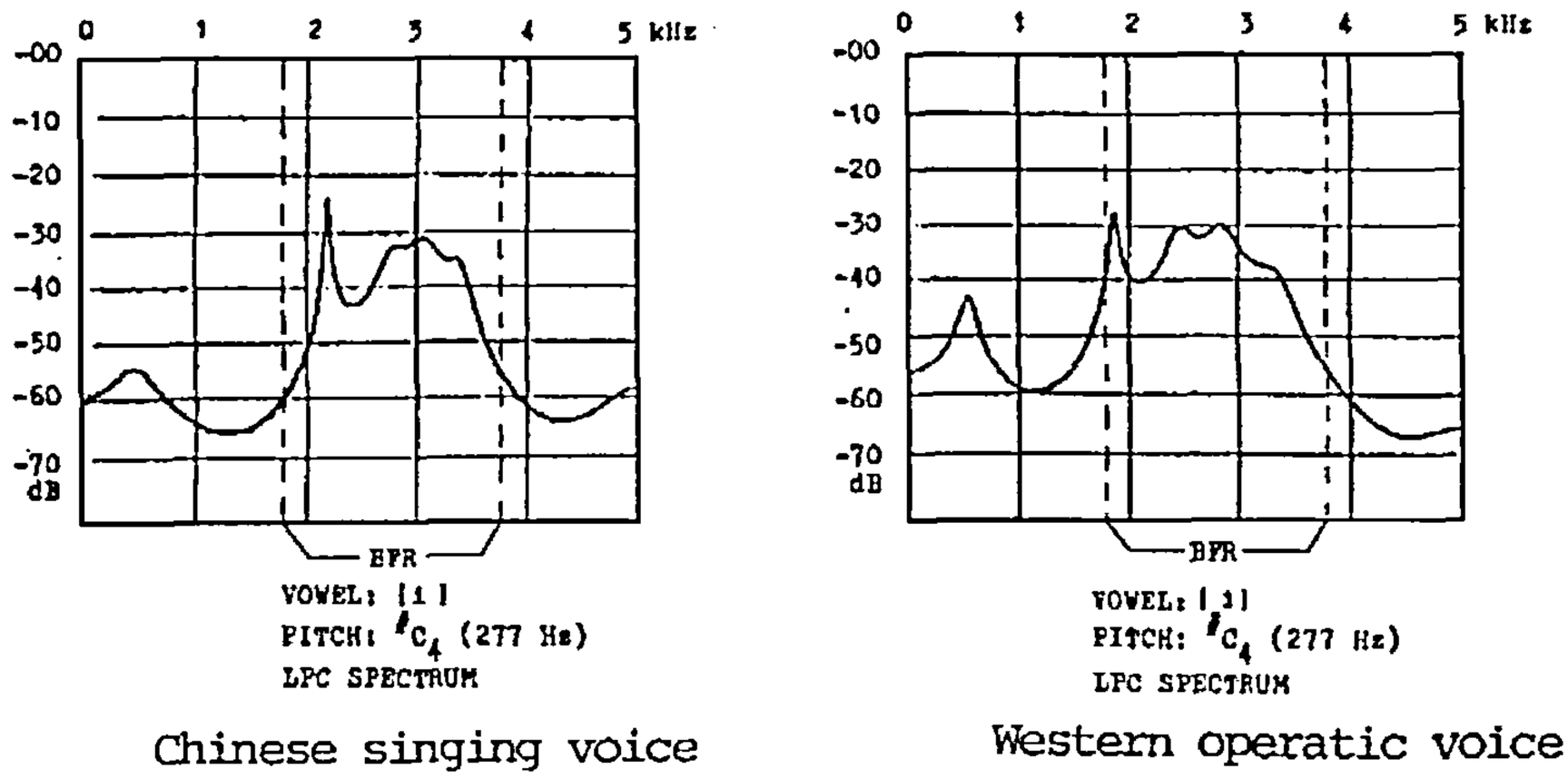


Figure 4.18. A spectral comparison of the Chinese singing voice and the Western operatic voice (from Wang, 1983).

Larynx height tends to rise with increasing phonation frequency in untrained singers. The studies above have shown that larynx height measurements are not necessarily a good way of distinguishing between trained and untrained voices since trained singers can sing efficiently with an elevated larynx, and brightness can be achieved with elevated as well as lowered larynx height (Wang, 1983). Since many singers have been shown to sing with an elevated larynx, Wang (1983) concluded that singing with an elevated larynx does not necessarily lead to poor vocal health.

It appears, then, that larynx height does not need to remain independent of pitch in order to maintain an efficient and good singing technique.

4.7 Conclusions

In summary, singing qualities can incorporate various attributes of spoken qualities; modal (chest), breathiness, harshness, and falsetto, though harshness and breathiness are seen to be inefficient and potentially dangerous especially when singing. Classical female opera quality arises from blending the two (or three) registers within the natural female voice (chest, (middle) and head) as one sings up the scale usually in order to derive a homogenous single quality throughout the range, though this is not a rule; female belting perceptually incorporates elements of chest quality, though it appears that some aspects of mode of production are different.

The conclusions to the literature review reveal several prominent areas which can be considered for investigation using available speech science resources:

From the acoustic output (microphone output):

Voice quality - frequency spectrum characteristics and dynamic perturbations

From the voice source (laryngograph signal):

Closed quotient;

Dynamic perturbations of the voice source;

Larynx height - though standard speech analysis techniques do not tend to use larynx-height measurements, it appears to be worthy of study in singing quality discrimination, as shown above.

Chapter 5

Experimental Procedure

5.1 Introduction

The applicability of standard speech technology to the study of singing science is the basis of this thesis. The experimental techniques described below have been chosen to reflect this. A number of the technological tools described previously are incorporated in a computer software package, thus reducing the amount of expensive bulky hardware equipment that would otherwise be needed. This outcome of these experiments will show how useful these techniques are in determining different vocal qualities in female singing.

5.2 Recording Location

The recordings were made in a sound proof booth in the Sound Acoustics Laboratory within City University's Department of Clinical Communication Studies. The sound booth has "dead acoustics" (National Sound Archives, personal communication with Allen Hirson). The dimensions of the booth measure 4m * 3m * 3m. The walls comprise of 8 layers of fibre glass sound insulation with an innersurface consisting of perforated customised acoustic tiles. The floor is carpeted. Minimum sound transmission occurs around the door and through the power points where there are no acoustic tiles. The sound booth is raised off the cement floor to minimise vibration transmission.

5.3 Subjects

West End musical singers and sopranos were recorded. From this, four sopranos and five West End musical singers were chosen. Three of the West End musical singers had been trained in both belting and opera qualities.

5.4 Equipment and Experimental Method

A multi-channel recording setup including a cardioid microphone, a Glottal Enterprises MC2-1 two-channel electroglottograph (described below) and an eight-track Alesis ADAT recorder to capture the acoustic output, the averaged Lx data and larynx height data from the voice source, was used. The equipment used in the recording procedure is listed below:

Sennheiser cardioid microphone MKH 40P 48U3

Beyerdynamic microphone stand

Microphone Phantom Power

Shure Prologue 200M microphone mixer to attenuate Lx signal

Glottal Enterprises MC2-1 tow channel electroglottograph

Alesis 8-track professional digital audio recorder (ADAT)

Two B+K Precision 3020 sweep/function generators

Crrus CRL 252 sound pressure level meter

Cirrus sound level meter calibrator

Chromatic pitch pipe

Glottal Enterprises conductive gel

Shure SM58 microphone

ITT instruments OX 7520 metrix oscilloscope

5.4.1 The Two-Channel Electroglottograph

This device has several advantages over the standard single-channel laryngograph; for example, it has ouptut channels which provide information relating to larynx height, average electroglottograph signal (EGG), fundamental frequency, extended low frequency limit signal, differentiated EGG, as well as the normal EGG outputs.

The two-channel electroglottograph can either be operated as a single-channel unit, with a single-channel cable, or, as in this experiment, as a two-channel unit using 34mm diameter two-channel electrodes. The device has a number of controls.

The normal outputs: These outputs are called the electroglottograph (EGG) signal, or electrolaryngograph (LX) signal. With the electrodes positioned so that the cable is at the bottom, the upper pair of electrodes provide outputs from channel A (CH A), whilst the lower pair provide outputs from channel B (CH B).

The average output: This is the average of the normal outputs, calculated by adding the waveforms together and then dividing by 2, i.e., (CH A + CH B) / 2.

The laryngeal tracking output: This output is a dc voltage signal taken from the laryngeal tracking meter on the front of the device. This output records vertical larynx movements. To

calibrate correctly vertical larynx movement, it is advised that the electrodes should be placed on the larynx such that the meter reads zero when the subject is vocalising continuously after taking a deep breath. The location of the vocal folds corresponding to zero on the tracking meter is called the "larynx resting position". At this point, the vocal folds are midway between the upper and lower pairs of electrodes. With the electrodes in this position, a vertical movement of the larynx will result in a positive voltage change, and a lowering of the larynx will result in a negative voltage change.

The extended LF limit outputs: These outputs are the EGG signals with an increased low frequency response. Abductory movements are represented by a positive increase in voltage.

The DEGG outputs: These signals represent the differentiated EGG signal and track glottal opening and closing points. They record the rate of change of the signal up to 3 kHz.

The F0 trigger output: This signal represents vocal fundamental frequency. A positive spike at each glottal cycle is recorded, the time between spikes is the period (Tx). This can be converted into a fundamental frequency contour (F0=1/ Tx).

The electroglottograph has no output gain controls and the DAT recorder has no input gain controls. The averaged EGG signal was passed through a microphone mixer before being recorded in order to optimise the EGG signal amplitude.

Both the laryngeal tracking signal and the extended LF limit signal are dc voltages. These signals were passed into the function generators which converted the dc signals into ac sine wave signals of an appropriate frequency such that they could be optimally modulated within the frequency response range of the DAT recorder (between 20 Hz and 20 kHz). The carrier frequencies were set to 6 kHz for the laryngeal tracking signal and 10 kHz for the extended LF limit signal. For every 1 V change the function generators frequency modulated the carrier signals by 2 kHz. The recording connections are shown in figure 5.1.

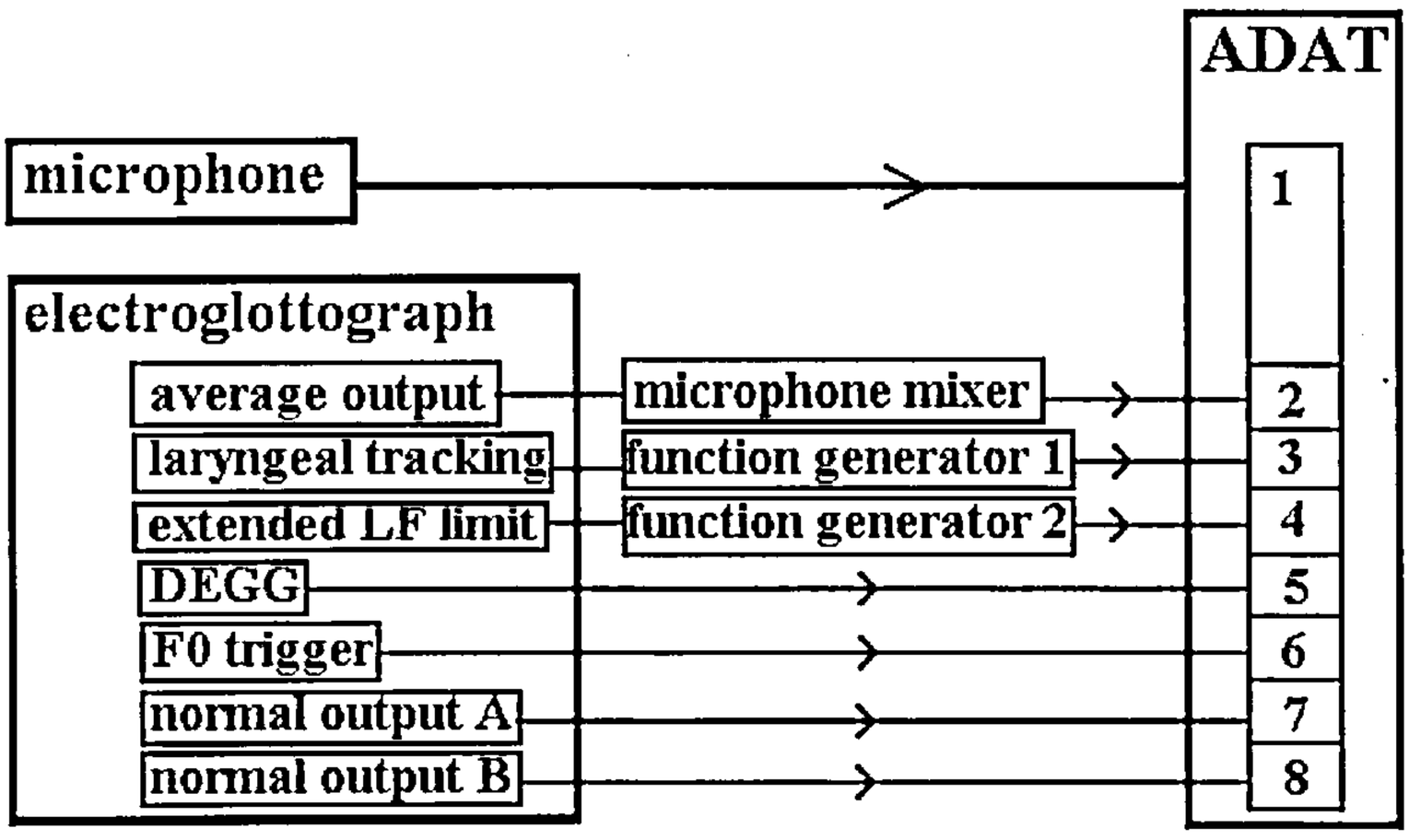


Figure 5.1. A diagram of the connections between the microphone and electroglottograph output channels and the input channels of the 8-track ADAT recorder.

5.5 Recording Procedure

At the start of each recording several procedures were carried out.

1. the sound level pressure (SPL) meter was set to mid-range, C-weighting and was calibrated with a 94 dB 1kHz test tone.

2. it was then set to max, with a 80-140dB range.

3. the electrolaryngograph electrodes were cleaned with water and a thin film of conductive gel was applied to the electrodes.

The subject was then asked to carry out some tasks to set up the recording levels, in the order outlined below. These were:

4. to strap on the electrodes and become comfortable with wearing the electrodes.

5. to find her larynx resting position by monitoring the laryngeal tracking meter whilst also producing a maximum amplitude Lx signal, monitored on the oscilloscope.

6. to stand 10 cm infront of the microphone and sing the loudest pitch she could. All signals were then adjusted so that they would not overload the ADAT recorder. The singer was requested to keep movement to a minimum and try and keep the same distance from the microphone during recording.

7. the SPL meter was positioned next to the microphone at arms length. The singer was asked to sing a sustained high pitch vowel whilst being recorded by the SPL meter. The maximum amplitude reading on the SPL meter was noted.

The singer was then asked to read a spoken passage then sing through a series of vowel exercises at different pitches in opera or belt qualities.

5.6 Digital Recording

The recordings were made on Ampex 489 super VHS tape with an Alesis ADAT professional eight-track digital audio tape recorder. The data was transferred into a Viglen 486 PC and stored in files using The EdDitor and Soundblaster software.

5.7 The Speech Filing System (SFS)

The computer software used for manipulating the recorded audio data is the Speech Filing System (SFS) (Edgington et. al., 1992). This tool allows convenient storage and handling of an original single speech data file which also contains the processing history of any subsequent manipulations by adding a header to the original file. A variety of utility programs are present in the SFS. The main programs provide spectrographic analysis, formant frequency estimations using LPC

analysis, larynx closed quotient analysis, and fundamental frequency estimation. This project does not use the spectrographic analysis program and LPC formant frequency estimations provided by SFS as these tools are not suitable for singing analysis at present; a stand-alone sound analyzer, the AND AD-3523 described in chapter 3 was used to create average spectra of the sung tones. The general purpose graphics program allows for SFS data display of files no longer than 2 seconds duration. The data files captured from ADAT into the computer are converted into SFS format which adds a header and arranges the stereo patterning.

5.8 Method for Extracting Larynx Height Data

The Larynx height data (referred to as Lx-height) in the form of a varying voltage was converted by a frequency modulator into a frequency varying sine wave on tape. In order to derive a larynx rest position (referred to as LRP) the subjects were told to relax, breathe in and on an outbreath phonate a sustained schwa vowel with relaxed vocal tract. The electrodes were then positioned so that the LRP was lined up as zero on the electroglottograph monitor. The frequency modulator was set to 6 KHz. The subjects were required to periodically repeat the larynx rest phonation and to line up the electrodes so that the signal on the monitor was zeroed. It was assumed that there would be some movement of the electrodes during the course of the experiment, so for comparison, the frequency signal which corresponded to the LRP phonation just prior to the above phonations was used and averaged. Any deviation in frequency denoted a change in Lx-height from this position, assuming that the electrodes did not move. This was captured by SFS and presented as a single frequency sampled at 100Hz. It was found that a frame size of 10ms duration gave the smoothest results. The spoken and sung phonations were compared to the LRP average.

Chapter 6

Results

6.1 Introduction

This chapter reports the results of several related investigations. It is divided into a number of analysis sections each with results and discussion work, concluding with a section drawing the several strands of work together. The analysis results will be presented in this order:

1. an investigation of CQ patterns between opera and belting qualities;

2. an investigation of the relationships between CQ, F0, vibrato, and lx-height;

3. a spectral analysis comparing opera with belt;

4. an investigation of larynx height with respect to opera and belting qualities.

The data has been chosen from a group of four opera singers (referred to by their initials AG, SS, SW, and TT) , and five West End musical singers (MC, KK, CM, VP, and AW), of which three have also been operatically trained (MC, CM, AW).

Each singer was required to speak and then sing a passage of words in belting quality or opera quality at different pitches. The passage was "booed, bead, bad, bud, bed, bird, bard, board" repeated at the pitches C4, E4, G4, C5, E5, and G5 (comprising a C major arpeggio). The analysis data are all extracted from this passage.

6.2 CQ Differences Between Opera and Belting

The aim was to investigate whether there is any difference in CQ measures between female opera quality and belting quality for a database of singers. The recorded laryngograph data from 4 opera singers, 4 West End musical singers, and one West End musical singer who also sings operatically was used. The Lx data was extracted from the steady-state portions of three vowels /3:/, a:/, and /i:/ from the carrier words "bird", "bard" and "bead" at the pitches described above. The data was subjected to CQ analysis and statistical analysis using the Wilcoxon rank sum Test.

6.2.1 Results

Table 6.1 contains the average CQ values for individual singers and a total average across singers for each pitch-vowel token. The subject's initials are followed by either "(o)" denoting that the

subject was asked to sing operatically, or "(b)", denoting that the subject was asked to sing in belting quality. The sung pitch-vowel tokens are grouped by vowel, and are listed as a rising arpeggio.

opera quality

token	AG(o)	SS(o)	TT(o)	SW(o)	CM(o)	group average
C4/i:/	41.02	33.38	51.83	40.65	32.76	39.928
E4/i:/	24.71	32.69	44.67	40.34	36.83	35.848
G4/i:/	25.77	24.96	35.16	40.4	40.66	33.39
C5/i:/	21.99	29.89	34.88	40.73	44.57	34.412
E5/i:/	23.93	27.08	41.69	36.72	42.25	34.334
G5/i:/	20.47	26.44	46.36	37.47	45.94	35.336
C4/3:/	33.29	32.01	51.52	34.63	38.24	37.938
E4/3:/	23.02	29.09	47.79	33.95	29.3	32.63
G4/3:/	20.68	26.36	29.84	37.26	33.33	29.266
C5/3:/	18.75	31.31	29.77	36.23	42.69	31.75
E5/3:/	22.26	27.48	39.09	35.51	44.26	33.72
G5/3:/	24.75	30.84	45.21	39.79	44.76	37.07
C4/a:/	34.82	31.21	47.65	35.01	39.3	37.598
E4/a:/	23.67	29.49	47.3	35.51	31.68	33.53
G4/a:/	21.6	25.49	29.73	34.88	34.63	29.266
C5/a:/	18.31	30.28	30.69	35.17	42.72	31.434
E5/a:/	21.91	29.03	38.17	32.61	45.57	33.458
G5/a:/	25.05	30.89	46.79	37.22	44.92	36.974

belting quality

token	CM(b)	MC(b)	KK(b)	VP(b)	AW(b)	group average
C4/i:/	40.57	52.5	47.92	41.03	42.87	44.978
E4/i:/	41.28	55.16	56.7	47.94	36.9	47.596
G4/i:/	47.07	57.8	63.69	53.64	40.71	52.582
C5/i:/	40.18	59.55	44.92	50.44	46.11	48.24
E5/i:/	45.07	58.08	63.49	49.91	39.76	51.262
G5/i:/	45.42	46.07	48.02			46.503
C4/3:/	43.4	44.96	50.41	40.23	41.61	44.122
E4/3:/	46.54	51.4	50.72	50.02	37.53	47.242
G4/3:/	53.8	53.9	63.15	54.39	36.55	52.358
C5/3:/	41.87	56.73	37.85	50.06	44.33	46.168
E5/3:/	44.77	59.39	59.65	40.91	47.97	50.538
G5/3:/	40.64	46.2	50.81			45.883
C4/a:/	46.84	46.54	51.61	42.4	41.04	45.686
E4/a:/	47.42	54.33	45.27	47.04	34.69	45.75
G4/a:/	51.67	56.24	60.32	54.44	36.54	51.842
C5/a:/	44.81	56.62	37.86	55.47	43.72	47.696
E5/a:/	42.04	59.74	64.51	39.91	45.07	50.254
G5/a:/	42.66	44.48	48.3			45.146

Table 6.1. CQ values for sung vowels at different pitches.

6.2.1.1 Average CQ of Opera Set versus Belting Set

CQ data from the opera group and belting group were averaged for each pitch-vowel token (final column of table 6.1). This is presented in figure 6.1. The most striking feature of this graph is the difference in patterns between opera and belting. The opera CQ patterns are lower in value than for belting, with a prominent dip then rising upwards as pitch increases. The opera patterns display a pivotal dip at G4 and the belting patterns display a dip at C5.

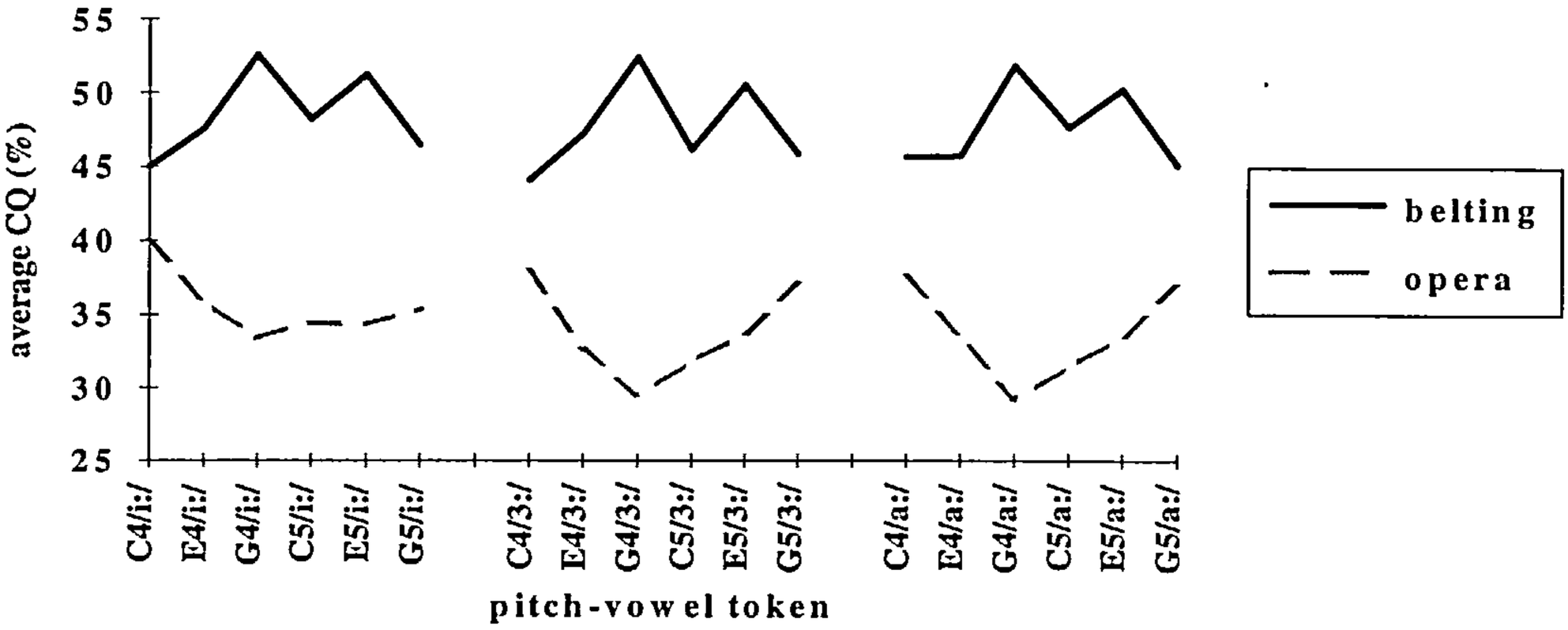


Figure 6.1. Average CQ patterns for the opera and belting sets (from final column of table 6.1).

6.2.1.2 Average CQ Statistics of Opera Set versus Belting Set

A Wilcoxon rank sum test statistical analysis of the results in figure 6.1 is shown in figure 6.2. Most of the results are at or above the 5% significant level, meaning that the vowels requested to be sung in belting have a significantly different value of average CQ than for those requested to be sung in opera, and the two sets can be described as mutually exclusive for most of the singing range. Vowels at pitch G5 were not considered so the sets could have an equal number of samples.

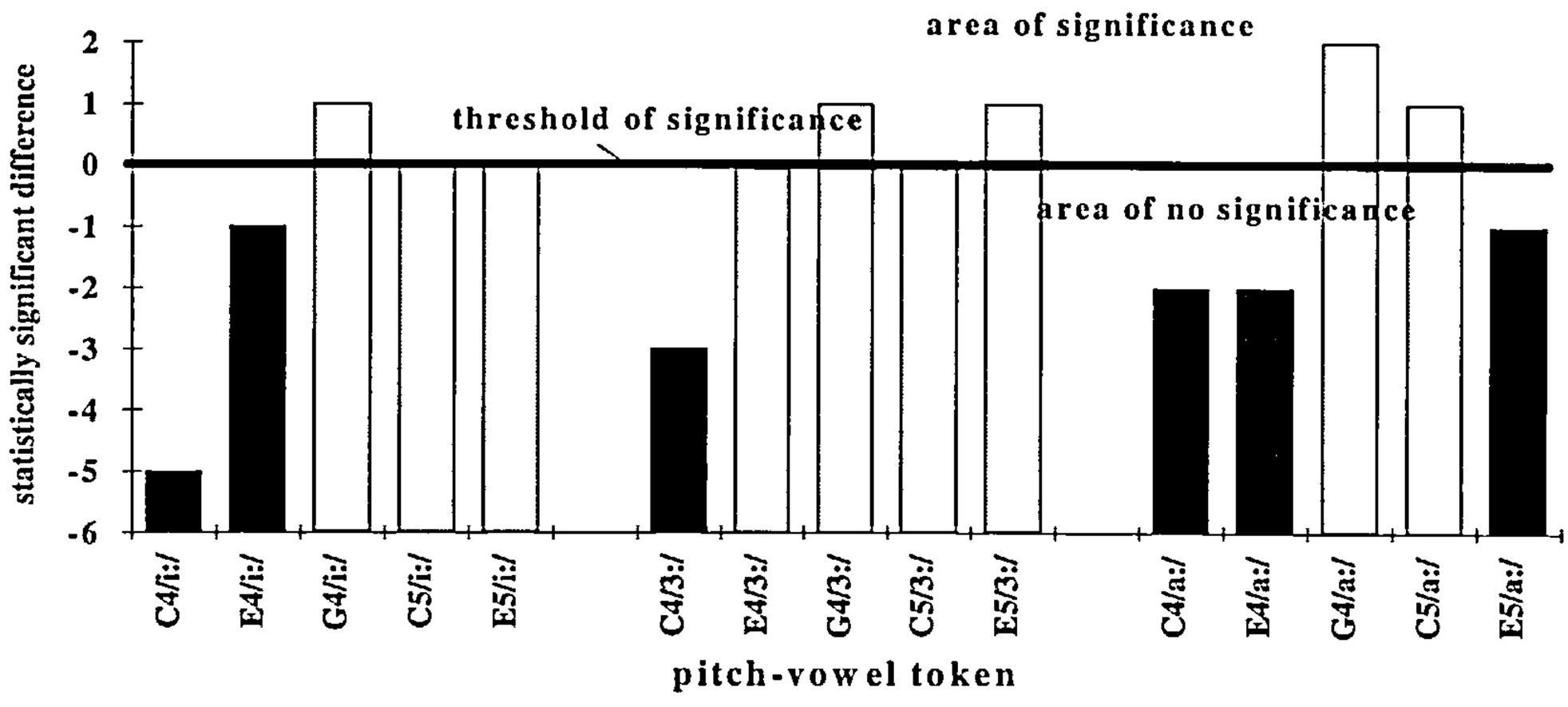


Figure 6.2. Statistically significant differences (at 5% level) in CQ between the opera and belting sets. The taller the column the more significant the difference is between the two sets.

The statistical results show that there is no significant difference for all pitch-vowel tokens at C4, and pitch-vowel tokens E4/i:/, E4/a:/, and E5/a:/. However, a number of the pitch-vowel tokens lie on the significant threshold. There is a low level of significance for pitch C4. The extremes of the range (C4/i:/, plus all vowels at G5 which are not represented in figure 6.2) show no significance.

6.2.1.3 Individual Average CQ Patterns of Opera Set

Figure 6.3 charts the CQ patterns for each singer, taken from the results in table 6.1. Differences exist between the singers, though each singer's pattern remains relatively consistent across vowels. There is a good deal of variablity in CQ patterns across the singers, which is not exhibited in the average trends shown in figure 6.1.

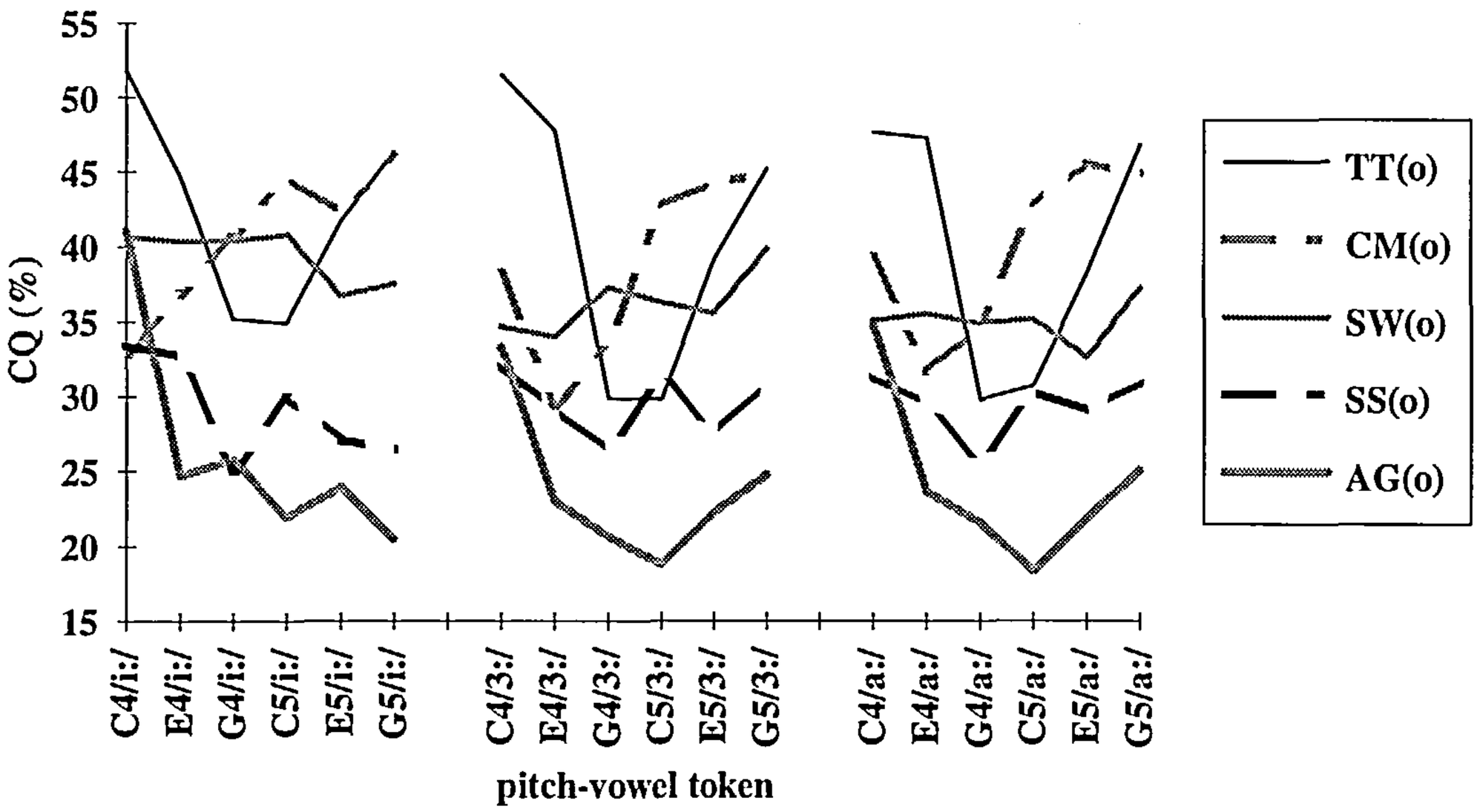


Figure 6.3. Individual CQ patterns for the opera set.

Figure 6.3 shows that there appears to be no consistency in CQ value, CQ pattern or CQ range across pitch for each singer, though there is reasonable consistency between vowels. Each singer's average CQ pattern can be characterized by its range, its degree in percent, and its pattern. These attributes appear to be unique to the individual, in other words, each singer has her own personal CQ pattern, CQ range and CQ position. These individual trends do, however, follow a general trend for opera quality, shown in figure 6.1. For the opera tokens, all vowels show a general increase in CQ after G4, up through the middle register. For the highest pitch G5, the more open vowels /3: / and /a: / have higher CQ values than the closed vowel /i: /.

6.2.1.4 Individual Average CQ Patterns of Belting Set

The CQ values cover a large range of nearly 30%, extending from 34% to 63%, the lowest values being much lower than expected. These results are also reflected if one eliminates the influence of the extreme values on the results, by taking the medians for each CQ pattern, as shown in figure 6.4. There is inter-subject consistency even though the range in values appears larger than one would expect for belting.

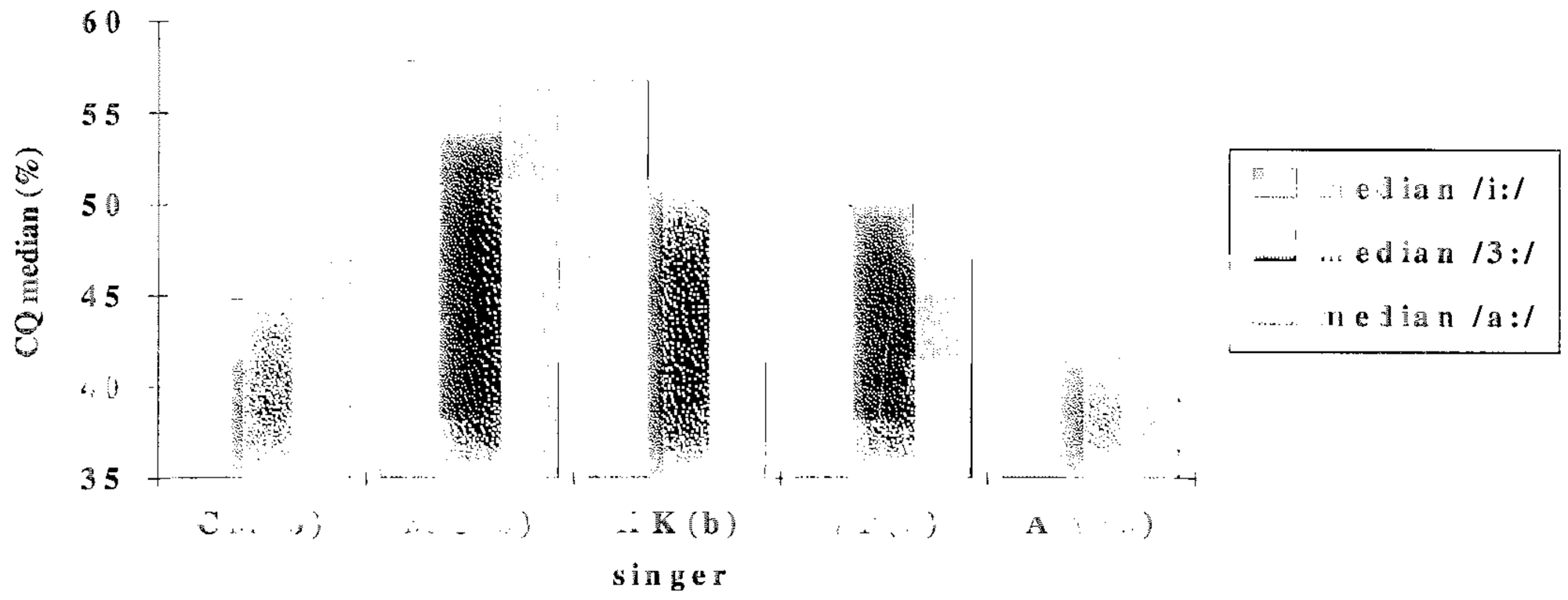


Figure 6.4. CQ medians for the belting sample.

The CQ patterns can be divided into those that display a CQ dip at C5, and those that do not. Figure 6.5 presents those singers whose CQ patterns display a CQ dip at C5. The patterns are highly irregular.

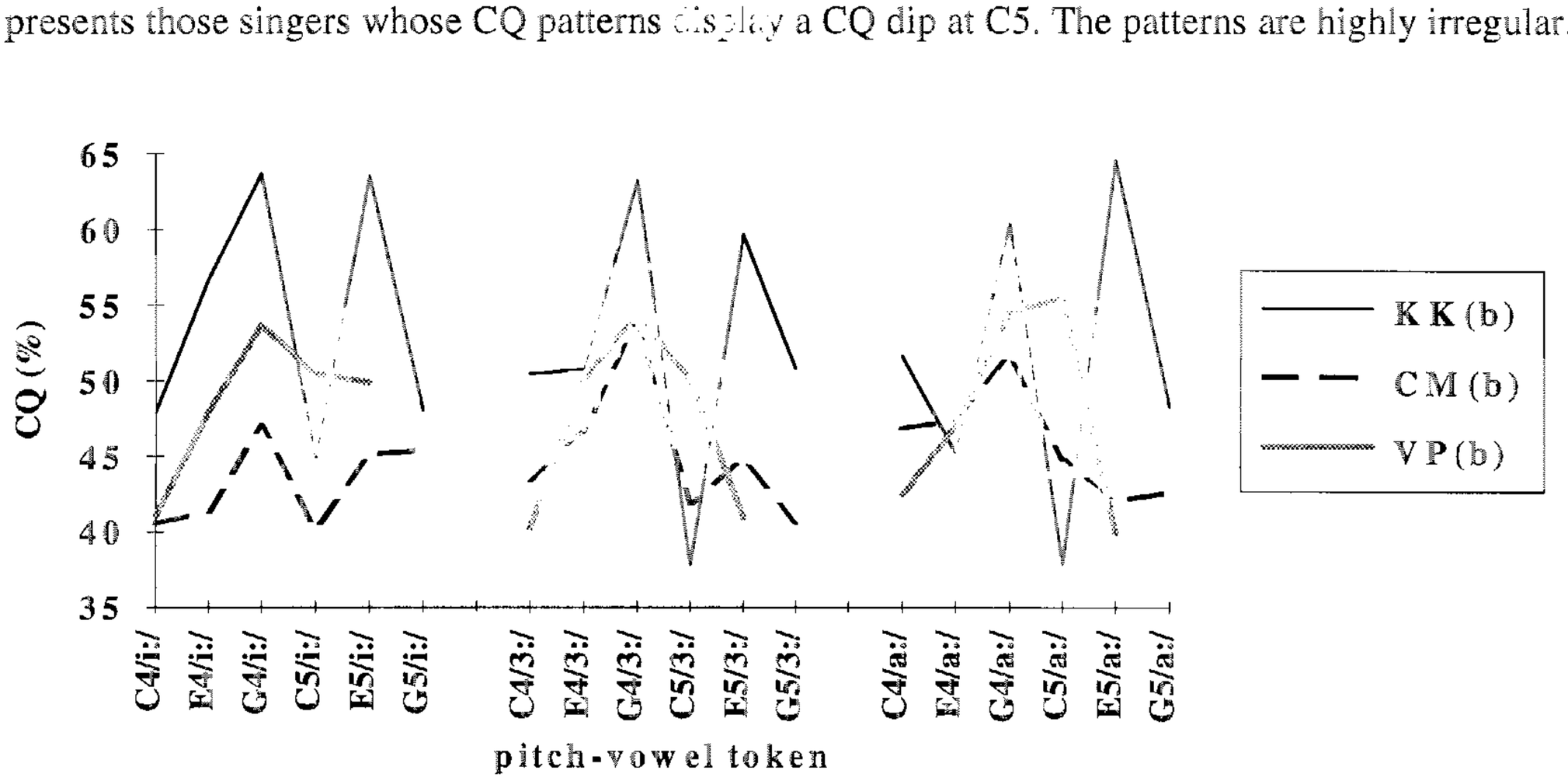


Figure 6.5. Belting CQ patterns which display a dip at C5.

Of the two singers that do not display a dip at C5 in their belting CQ patterns, shown in figure 6.6, singer AW's belting pattern more closely resembles the opera pattern displayed in figure 6.1. Of the patterns in MC's belting, C5 and E5 have the highest CQ values.

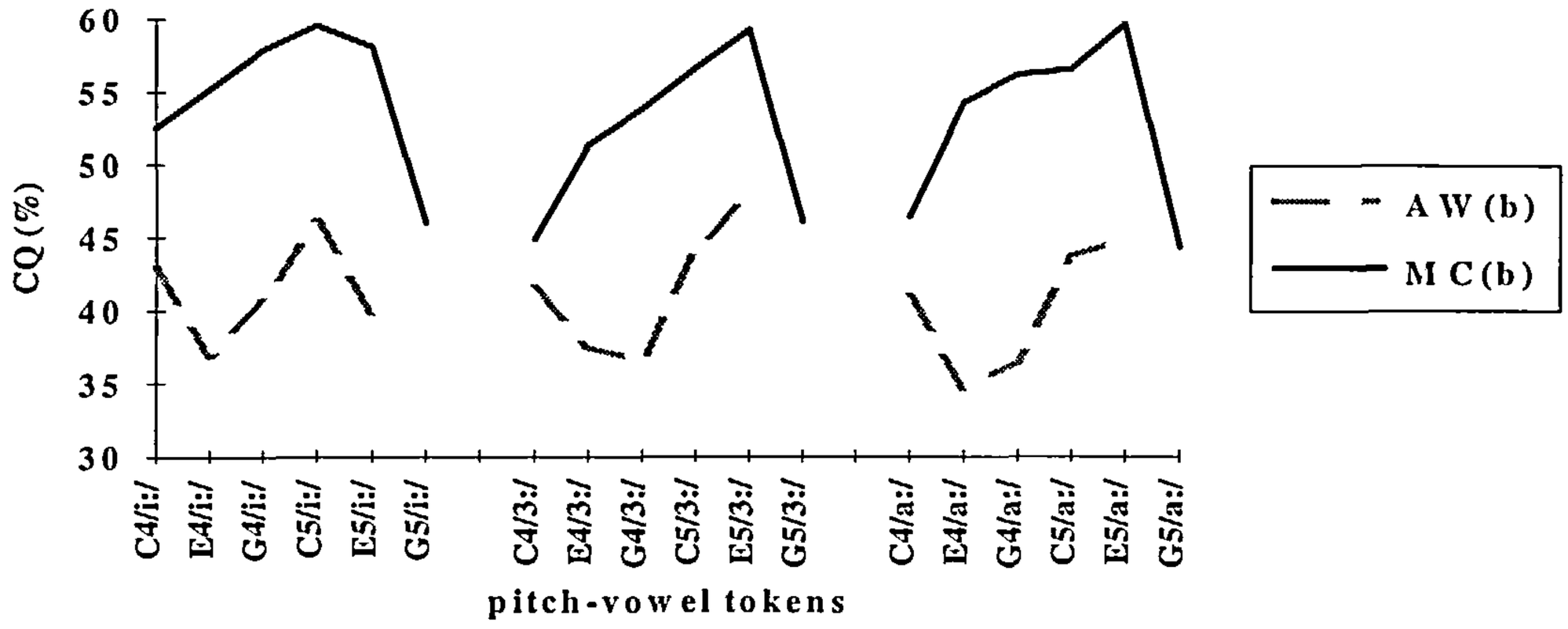


Figure 6.6. Belting CQ patterns which do not display a dip at C5.

6.2.1.5 Comparison of Opera and Belting CQ Patterns of one singer

Figure 6.7 below shows that the CQ differences in opera and belting for subject CM who is trained equally in opera and belting do show a difference in trends similar to the Estill (1988) study. However, CM's CQ values are lower in both patterns, the opera pattern has a smaller CQ range, and the belting pattern is markedly angular with a prominent peak at G4.

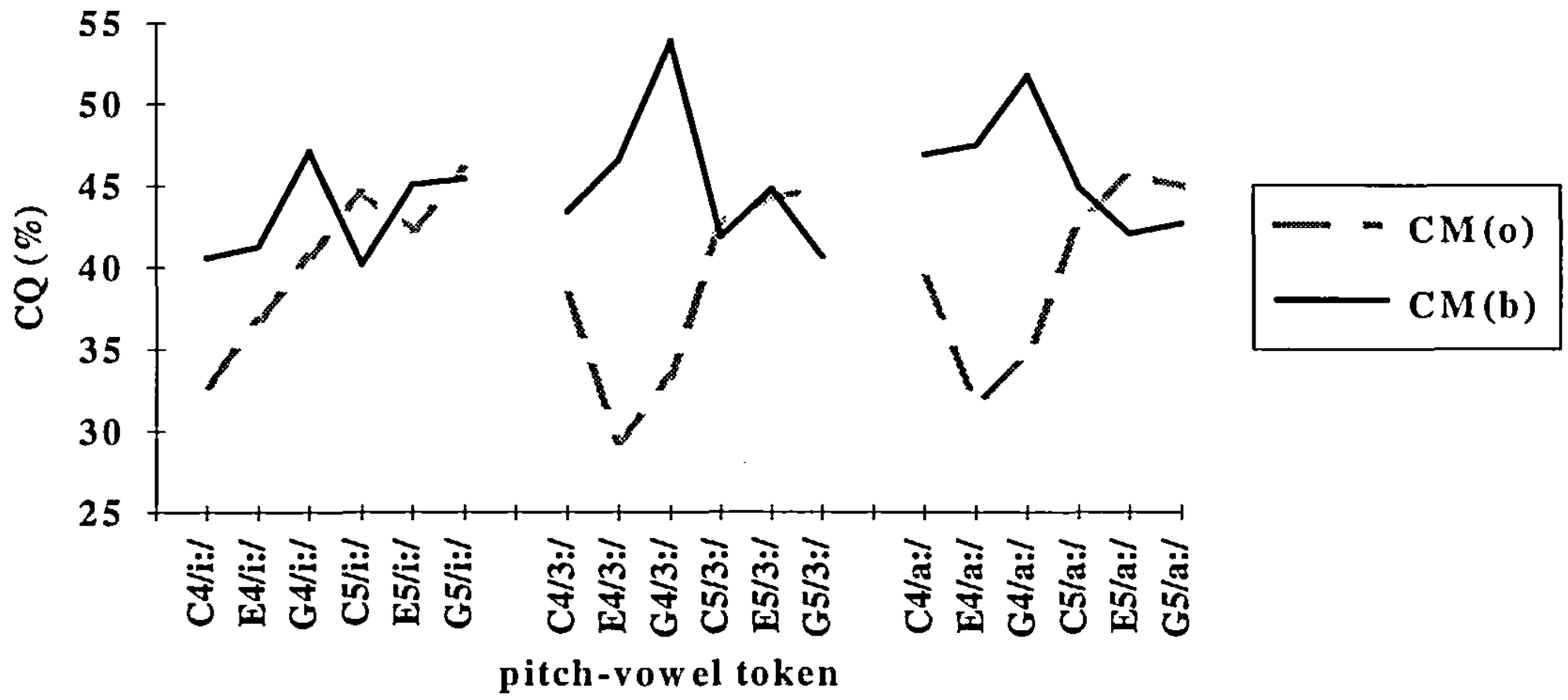


Figure 6.7. Comparison of opera and belting CQ patterns of singer CM.

6.2.2 Discussion

This section is divided into a seperate discussions of the opera and belting CQ results, followed by a general discussion of the two sets.

6.2.2.1 Summary of Results

The analysis so far has resulted in these observations on CQ:

a) on average, different patterns of CQ exist between opera and belting-

- belting has generally higher CQ values in the middle pitches (E4-C5) than opera;
- in opera: pivotal dip in CQ at G4; rise in CQ after G4;
- in belting: dip in CQ at C5 in some singers; highly erratic CQ pattern (unlike results from Estill's study (1988));

b) extremes of CQ show little difference (statistically no significant difference);

c) both opera and belting: not much influence from vowels on CQ (however, spectral analysis needs to be done - the singers may have modified the vowel quality).

d) each singer's CQ set can be characterized by its scalar position, range, and pattern.

6.2.2.2 Discussion of CQ Results

These results partly resemble the CQ patterns for opera and belting observed in the single-case study by Estill (1988), shown in figure 4.14. She suggests that the differences in trends between the two qualities arise from their different production mechanisms. She attributes the consistently higher belting values to there being no register differentiation in belting, whereas in opera quality, the low values for CQ at around D4 in her study represent the chest-middle register transition for that subject.

a) discussion of the opera CQ results

The general trend for the opera CQ results is a dip at G4 followed by a rise (see figure 6.1) which conforms to the observations found in the literature. Referring to figure 6.1, the pivotal dip at pitch G4 exhibited by the opera CQ patterns indicates that at around this pitch, there is a possible register transition from chest to middle register, chest register being characterized by higher CQ values than for middle register (Sundberg, 1987). This register break is higher here than in Estill (1988). The results reported by Howard (1995) supports this theory:

> "A number of [trained] female subjects exhibit turning points in their Qx plots (mean = 404 Hz, or approximately G#4). This could be indicative of a chest-middle register break point. The results were plotted as scattergrams (Qx) of CQ(%) against F0 (log Hz). All the Qx plots exhibited an overall pattern of variation with an essentially linear variation of CQ with log(F0) within two F0 sub ranges approximately either side of G4" (Howard, 1995).

For the opera group, the higher than average CQ values across the vowels at C4 suggest that, on average, the opera singers are mainly using their chest register. All of the opera singers are sopranos and it is not uncommon to have a chest-middle register break around E4. Extending the middle register down below the register break is difficult and results in an unclear tone with a low average CQ value. These singers may be predominantly using chest register because at these lowest pitches, the action of the set of muscles used in the chest register can be stronger than those used in the middle register. A higher CQ value (in the region 35-50%) is usually associated with a good frequency spread of harmonics which can be transmitted by the vocal tract, resulting in a strong sound. A low CQ (below about 25%) is

usually associated with slow vocal fold closure, suggesting that the glottal waveform spectrum could be lacking in components in the high frequency range, thus making it potentially more difficult for the vocal tract to transmit the higher frequency information required for voice projection at the lowest range.

Opera singers are trained to mix registral qualities on and around the register breaks in order to achieve a homogeneity of sound. It is likely that these singers are mixing a little middle register quality in with the chest sound on pitch C4.

For the opera tokens, all vowels show a general increase in CQ after G4, up through the middle register. For the highest pitch G5, usually sung in the "head" register, the more open vowels /3:/ and /a:/ have higher CQ values than the closed vowel /i:/. One may speculate that the mechanisms used to project the sound at high pitches in opera quality favour more open vowels and a rise in CQ. However, singers tend to sing louder at high pitches. This may be responsible for the rise in CQ across the opera singers' range. More investigation into the effects of loudness on CQ is required. The singers SS and AG have opera patterns which do not rise consistently with pitch (accounting partly for the drop in average CQ value for vowel /i:/ in figure 6.1). It should be noted that not all opera singers exhibit this CQ rise with pitch after G4. Singer AG has an average CQ rise at C5 combined with an unusually low CQ range. In the authors' opinion, this does not appear to detrimentally affect the quality of the operatic tone with rising pitch.

There remains a good deal of variability in the CQ patterns between subjects, but is relatively consistent within singers on different vowels (see figure 6.4). Figure 6.3 highlights the individuality in CQ characteristics for each singer. There is greater variability within CQ for opera quality than is suggested by the average CQ patterns across singers. Conclusions on CQ and vocal quality should be drawn not only from averaged data but also in conjunction with individual singers' CQ patterns.

b) discussion of the belting CQ results

A discussion of the results for this particular set of singers is difficult due to the fact that some of the singers may not be belting consistently across the pitch range. It is unclear from looking at these CQ results in isolation whether these tokens are representitive of belting quality.

The results discussed below depart from results found in the current literature on belting:

The dip in CQ at C5 in the belting patterns in figure 6.1 and figure 6.4 indicate that at this pitch there appears to be a break in quality for some of the singers. A closer look at table 6.1 will reveal that CQ values vary widely within the individual belting sets. These results depart from Estill (1988) and Evans & Howard (1993) which have shown that belting values are at a consistently high level across the singers' ranges and attribute this to there being a consistently high adductory force on the vocal folds. The singers in this study may be taking their chest quality up too high through the middle range (from E4 to C5) resulting in a register transition at this pitch. This suggests that the vocal production of these singers is not characteristic of belting production. These CQ values must be looked at in conjunction with the corresponding acoustic spectrums before any reasonable conclusion can be reached.

Singer AW's belting pattern (see figure 6.5) does not remotely resemble the model pattern for belting, as proposed by Estill (1988), though in the author's opinion, she perceptually sounds like

belting. It is possible that a spectral analysis will reveal that this singer is singing in a quality which is inbetween opera and belting called "mixed", which has been described in chapter 4. The CQ pattern for this singer is closer to that for opera quality, yet the percept is closer to belting. This is evidence that one must be aware of classifying vocal qualities based on CQ evidence alone.

The Estill subject can reach notes in belting quality (G5) in a consistent fashion (in terms of CQ values) which the singers in this study fail to reach. Even pitch E5 proves to be difficult. The results in table 6.1 suggest that the belting tessituras for most of the belters do not extend up to pitch G5. This is confirmed from listening to the samples. The singers appear to be using a vocal quality which has a lower CQ value than is normal for belting for the upper extremes of the range, possibly, again, "mixed" quality.

The CQ values for belting are lower than those found in the Estill study. The disparity in results between this study and the Estill study (average CQ is around 70%) could be the result of a different technique with harder glottal adduction in the case of the Estill subject. If one eliminates the influence of the extreme values on the results, by taking the medians for each set, as shown in figure 6.6, there is inter-subject consistency even though the range in values appears larger than one would expect for belting. This is an indication that along the CQ continuum for belting, different singers may occupy their own individual space within a relatively large range (perhaps the upper level being occupied by the Estill subject), in much the same way that the opera singers do. In other words, there may be some freedom in the degree of glottal adduction required to produce belting. The natural strengths of each singer's laryngeal and associated musculature could be the origin of this. Assuming the correct production, it is possible that there is an individual lower limit in glottal adduction below which the singer's vocal tract may not be able to compensate for the lack of higher frequency energy, and belting quality cannot be produced. Conversely, there may be an individual upper limit above which the voice is too strained. CQ values above 60 % with no stress on the vocal system may be easily reached in one singer (for example, the Estill (1988) subject) but not in another without harm to the voice. This could also apply to opera singing, though it has been shown that the stress on the vocal system is somewhat less in opera singing than in belting (Estill, 1988).

The single subject results (see figure 6.7) show a highly erratic belting pattern. At the time of recording, CM was singing in a music theatre production which required her to sing in opera quality. It is possible that with a singer who is equally trained in different qualities, the quality which is currently in use is the strongest quality; rather like a bilingual speaker. It is suggested that CM's belting quality is possibly weaker than her opera quality due to her being out of practice with belting.

Of the five belters, singer MC appears to have the most consistent CQ values across the pitch range.

c) general discussion

The non-significance in CQ difference between the opera and belting sets for pitch C4, E4/i:/ and E4/a:/ has been attributed to the use of chest register in the opera voices which has a higher CQ value than for middle register. This means that the CQ values for chest quality are very similar to belting quality. It has been stated in chapter 4 that belting makes use of elements of chest production. Sound

pressure level must be taken into account. It is possible that the belters are in fact chesting C4 and E4 but with greater loudness due to greater subglottal pressure in combination with a higher adductory force (hence a possible higher CQ value), since belting is supposed to be "loud". It is difficult to say whether someone is belting or chesting using CQ measurements alone. However, Estill's theory for belting versus chesting would be that if there is a break in the CQ pattern for belting, then the production is incorrect.

Chest versus belting production is an important issue which cannot be easily clarified from looking at CQ measurements alone. However, deviations in a CQ pattern across range are good indications that some other production mechanism may be in use. Within a single voice, it appears that an anomalous CQ value is a good indicator that a register transition may have occurred in the same voice. For example, see the opera patterns - there is definitely a difference between chest and "middle", although there is no apparent difference between middle and head registers in terms of CQ. It is a continuum, possibly reflecting an increase in loudness. CQ in itself does not indicate the amount of stress on the larynx, yet a variation of CQ from, say, a professional singer's normal pattern of behaviour for CQ could indicate an error in production which could then be further investigated. CQ as a parameter in visual feedback displays in singing training is attributed to David Howard and Paul Garner at the University of York where CQ feedback systems for vocal training is being developed.

The results here suggest that within sets, the CQ range is large. This may also be due to the belting set including a large portion of non-belted tokens, hence the need for spectral analysis and perceptual testing as well. The very definition of belting may be different for different people.

One explanation for the lack of real belting data (in terms of Estill's work) in these results and would account for the lower than expected CQ values, could be that the type of sound required by London musicals seems to have moved away from the traditionally American hard belt sound to a softer more naturalistic sound which has elements of the belting sound in it (the "mixed" quality). It is possible that the absence ot twang in (southern) British speakers makes belting harder than for Americans who have this natural element in their accents. This study has shown that the concept of belting for most of these singers is not as rigid as the author originally thought. The CQ dips in the belting patterns may be a result of poor singing technique. If the singers practised this pitch, then the quality may be more continuous through the range, and the CQ pattern could be more linear. However, these singers produce tones which sound like belting but which may not fit the definitions given in the literature on belting. If they can produce a tone which sounds like belting without damaging their vocal instruments, to what extent does the production matter?

CQ as a measure of length of vocal fold closure does not really give any indication as to the true source of the production since CQ is one measurement derived from another measurement, Lx. Vocal fold closure is determined by a combination of subglottal pressure and glottal adduction through tensing muscles in the larynx. The CQ measurement alone cannot provide information on the respective degrees of laryngeal tension and subglottal pressure in a vocal production. These are very important factors which must be known if vocal production mechanisms are to be understood and especially if CQ is to be used as a parameter in vocal training. It is suggested that Lx shape should be looked at closely with its corresponding CQ value, since laryngeal tension can be more easily assessed from the shape of the Lx

waveform. CQ does appear to suggest general differences in production between different qualities and has a role in indicating erroneous vocalisations within a pattern. The most consistent results across singers are the CQ patterns in the opera quality, though it has also been pointed out that a CQ pattern may be that of one quality yet the percept is of another (in the case of singer AW, the CQ pattern of opera, yet the perception of belting). It appears then, that in singing, the CQ measurement alone is not an absolute indicator of vocal quality, yet it has a useful role in indicating voice-source differences which may then be further looked at. It is suggested that CQ would be of far greater value when studied in conjunction with a number of other parameters such as subglottal pressure and the degree of tension in the larynx (for example, from the Lx waveform).

6.3 CQ, F0, Vibrato, and Larynx Height Relationships

Figure 6.8 shows the relationship between F0, CQ, and Lx-height for each singer singing "bard" on pitches E4 and E5.

6.3.1 Vibrato Differences Between Opera and Belting

As observed in previous studies (Schutte & Miller, 1993; Estill, 1988) the opera tones are sung with vibrato, whilst the belting tones generally have very little vibrato or are sung completely straight. The observations above are also apparent in this study. Vibrato rate in opera appears to be quite consistent between subjects and across the subjects' ranges, varying between just over 5 Hz to 7 Hz. For example, it is observed in figure 6.8, in the F0 analysis sections for each utterance, that vibrato rate is between 5 Hz to 5.5 Hz for TT(o)E4 (5 Hz) and TT(o)E5 (5.5 Hz); 5.5 Hz for CM(o)E4 and CM(o)E5, 6 Hz for AG(o)E4 and AG(o)E5, SW(o)E5, and SS(o)E4, and a little more for SS(o)E5 (7 Hz).

Sundberg (1987) states that most singers find it difficult to change their vibrato rate. It appears that for two of these opera singers TT and SS, vibrato rate increases slightly (0.5 Hz) with an octave increase in pitch. Whether this slight increase is a function of an increase in loudness remains to be investigated. It should be noted that these vibrato rates have only been calculated from the 1-2 second tokens shown in figure 6.3. It is also apparent that vibrato frequency modulation differs between each opera singer, with singer SS having very little, whilst singer TT has a pronounced modulation in both the E4 and E5 utterance. It has been suggested that vibrato may be a function of the voice system being under optimum stress, and also a function of the system monitoring pitch (Graham Welch, personal correspondence 1995).

In the belting tokens, there is little or no vibrato present. Each singer exhibits similar patterns of vibrato within their own ranges, though the amount of vibrato varies between singers. For example, singers KK and CM sing with almost no vibrato whilst singers MC and AW sing with clear modulating vibrato. The vibrato rate is the same (at 6 Hz) for MC who was asked to sing in both belt and opera, and

for AW it is 6 Hz for her attempt at belting and nearly 6 Hz for her opera attempt on E5 (her opera attempt on E4 appears to be a little wobbly), further adding to Sundberg's argument stating that when vibrato is present, it represents a set feature of a singer's vocal mechanism and cannot be changed easily.

The results here suggest that vibrato rate may be independent of vocal quality. It could also reflect what is considered appealing for the Western ear, which, as it happens, corresponds to the rate at which vibrato is produced when it develops as a natural consequence of vocal training in Western culture. Of the three singers who were asked to sing in both qualities, only singer CM distinguishes between the two qualities by singing her belting tones almost straight. A little vibrato is added at the termination of the word in belting for singers CM, KK, and VP suggesting that these singers can "switch on" their vibrato at will, probably for the sake of adding a little colour and variation to an otherwise flat sounding tone.

The results here suggest, then, that when asked, opera singers will sing a tone in opera quality which has a stable vibrato rate between about 5.5 Hz and 7 Hz, implying that vibrato is an important feature of opera quality, but singers who are asked to belt, produce tones which vary between containing no vibrato at all, to those which have vibrato at a rate similar to that sung in opera quality, suggesting that vibrato is not so important in defining belting, but it's inclusion is more probably left up to interpretation of the text and character by the singer.

6.3.2 Relationship Between CQ and Vibrato

For TT(o) both tones on E4 and E5 display CQ moving in synchrony with vibrato. The vibrato has a large modulation amplitude, as does the CQ line.

For SW(o) and AG(o) only the tone at pitch E5 shows this CQ-vibrato synchronicity. Both singers' E4 tone have less vibrato and very little CQ movement.

SS(o) has the opposite; CQ-vibrato synchronicity is more evident for the lower note E4, even though there is vibrato in the E5 tone.

It appears, then, that CQ movement can exist as a function of vibrato production, since it can vary in synchrony with vibrato, or it can be independently stabilised at a particular value (or, as in singer AW, interestingly in both opera and belting qualities at pitch E4, CQ moves out of synchronicity with the vibrato towards the end of the utterances, showing that it can do its own thing, too). It seems to depend on the individual singers. For those singers displaying CQ-vibrato synchronicity, a possible explanation is given below:

Vibrato involves modulation in frequency below and above a mean. The frequency deviation would not be great enough in order to appreciably alter the length of the vocal folds, so an explanation that the longitudinal stretching and relaxing of the vocal folds due to the vibrato frequency modulation may alter the adductory force (and hence CQ) is not appropriate here, since adduction involves medial compression; however, there is also an intensity amplitude modulation due to harmonics moving in and out of the formant peak that is in synchrony with the vibrato - see the Lx waveforms for the whole utterance. It is possible that the whole larynx mechanism tenses and relaxes with the production of vibrato. CQ-vibrato synchronicity is possibly derived from a cyclical mass tensing and relaxing of the

vocal mechanism (evident in SPL measurements of the acoustic output and at the larynx level), as opposed to some localised phenomenon, and it is possibly dependent on the way the singer produces her vibrato. It is doubtful whether CQ modulation of this order can be perceived since it is hard enough being able to correctly determine the CQ value for a well-produced tone. What it does show is that vibrato production can involve a great deal of the vocal mechanism, if not all of it.

6.3.3 Relationship Between Vibrato and Lx-Height

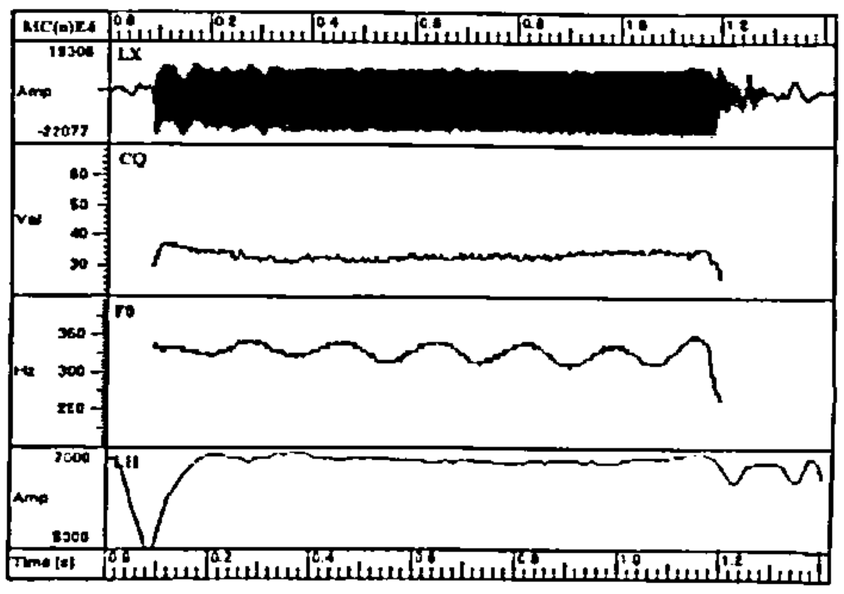
Again, as for the relationship between vibrato and CQ, there appears to be no correlation for half of the singers, and correlation for the other half. For most of the tones Lx-height is held constant throughout the steady-state portion of the tone. This suggests that Lx-height is not appreciably altered during the production of vibrato for these singers. However, there appears to be some movement as observed in four singers: CM, SW, AG, and MC at pitch E5, which is in synchrony with the vibrato.

6.3.3.1 Discussion

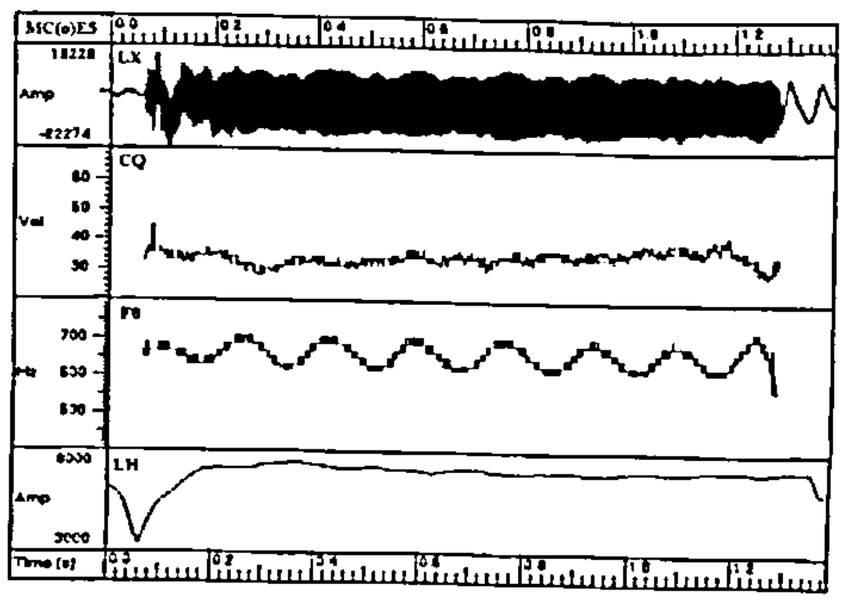
One suggestion for the synchronous movement of Lx-height and vibrato is that vibrato involves vertical variation of the larynx, and this can be related to the effects of subglottal pressure upon the larynx mass. As shown previously in figure 2.5, subglottal pressure varies in synchrony with fundamental frequency at a vibrato rate of 6 Hz (Sundberg, 1987). It seems, then, that not only is the vocal tract and larynx involved in the production of vibrato, but that the vibrato subglottal pressure variations suggest that the subglottal system, and possibly the diaphragm and/or stomach muscles are also utilised. This vibrato variation in subglottal pressure could cause the larynx to move up and down with the rise and fall of the pressure upon it, thus accounting for the synchronous movement of the larynx with vibrato.

These measurements are beyond the scope of this thesis, and further work combining subglottal pressure, Lx-height movement, vibrato and other features is needed. It is also not possible from these results alone to assess whether vibrato is produced intentionally or unintentionally, other than by asking the singer. It is also not possible to address what the exact causes of CQ, F0, and Lx-height movement really are, and the causal relationships between them (for example, which parameters are direct functions of others, which are determinants, and which are secondary observations).

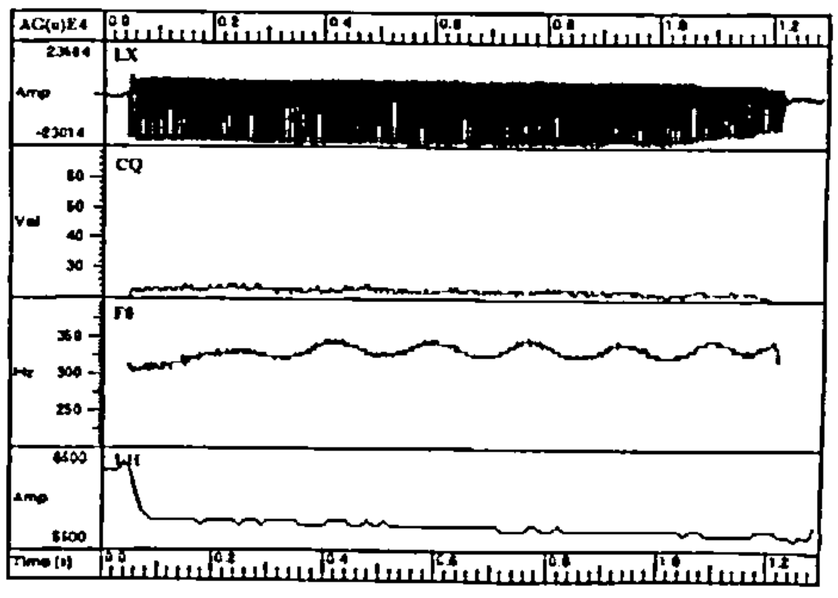
The observation of the larynx moving up and down in synchrony with vibrato implies that certain muscles belonging to the hyoid complex must be relaxing also. One can speculate that these will be the ones connecting the thyroid, such as the stylopharyngeus muscle, the palatopharyngeus muscle, and the sternothryoid (muscles 13, 15, and 17 respectively on figure 2.9) (Graham Welch, personal correspondence, 1995).



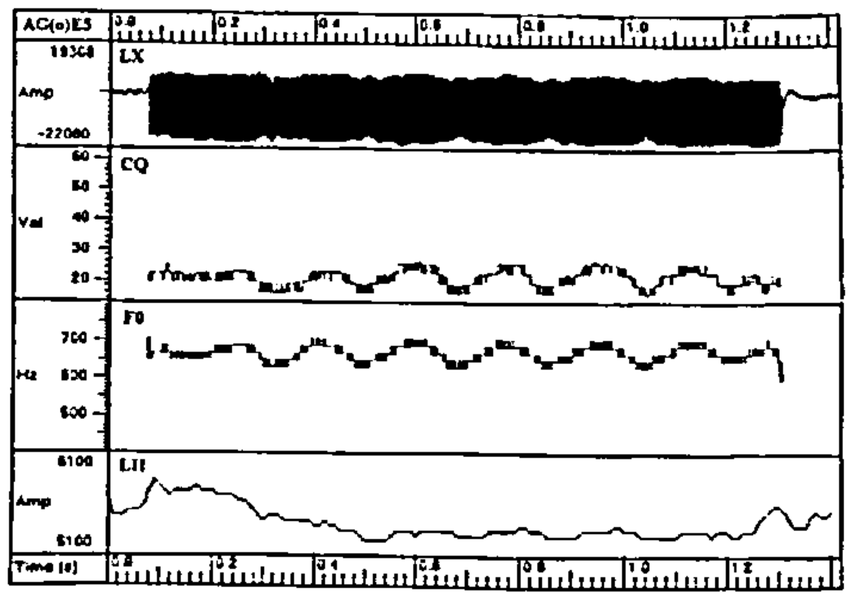
MC(o)E4



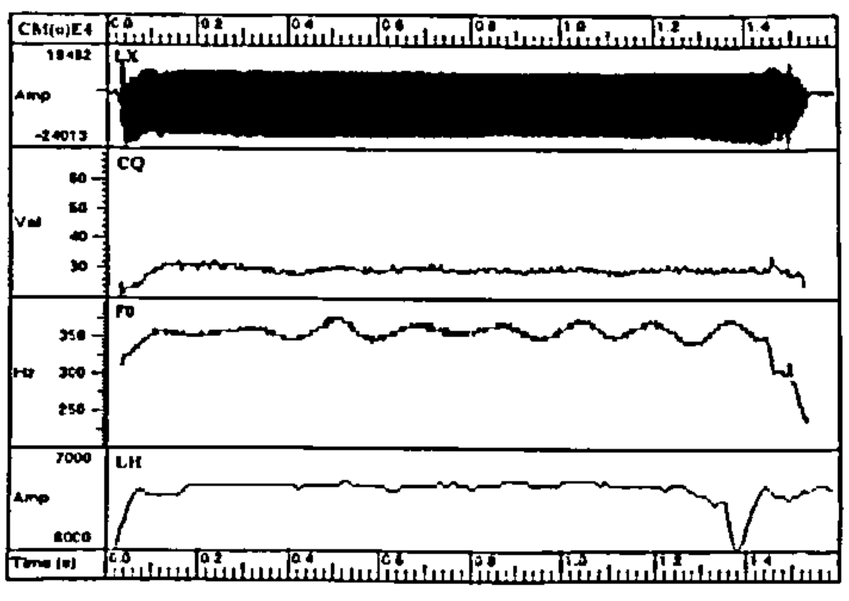
MC(o)E5



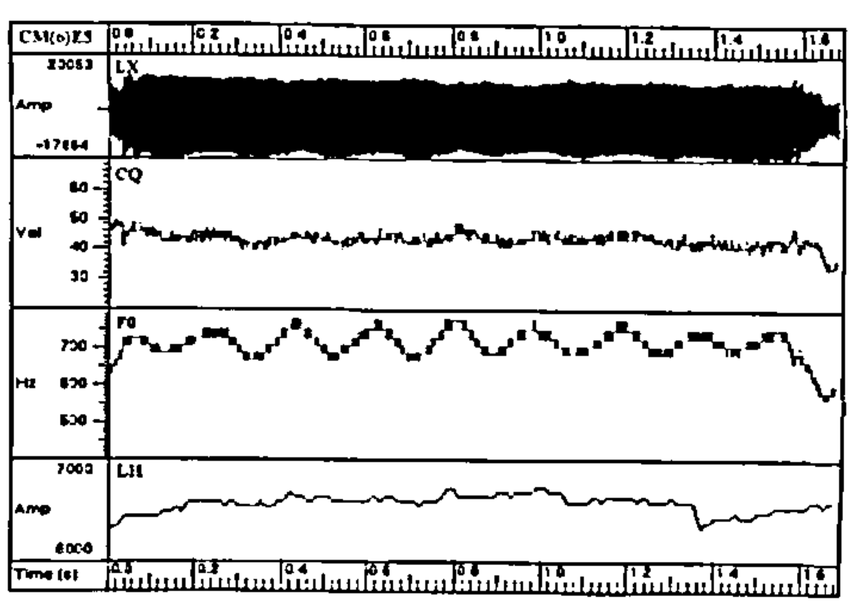
AG(o)E4



AG(o)E5



CM(o)E4

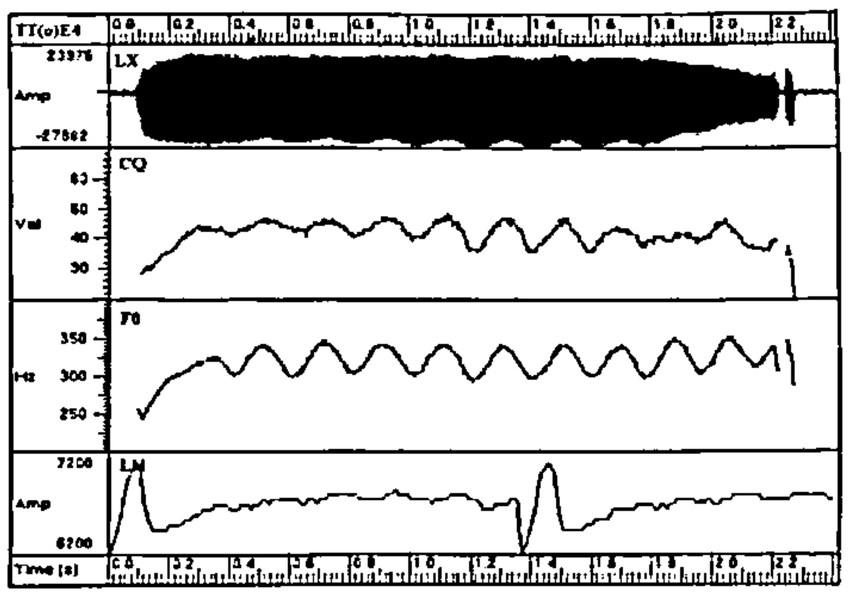


CM(o)E5

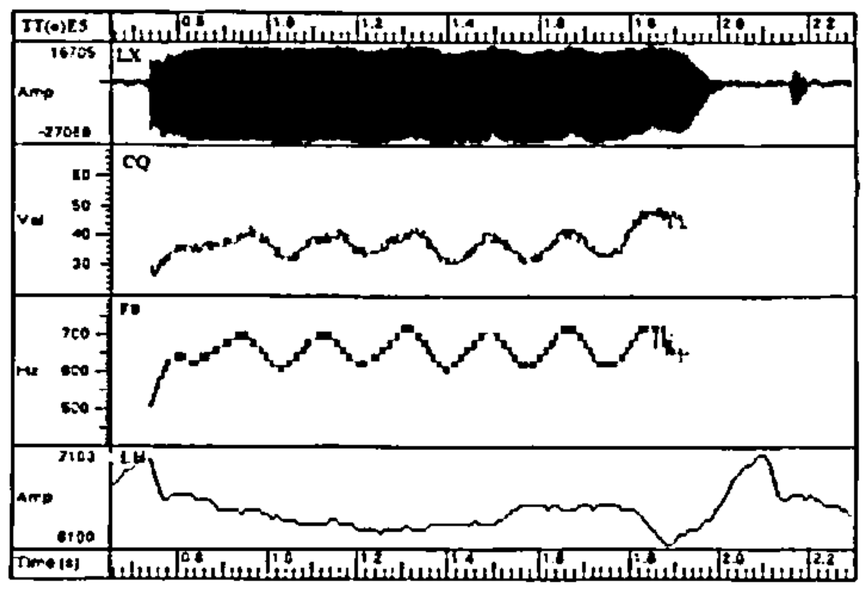
SS(o)E4

SS(o)E5

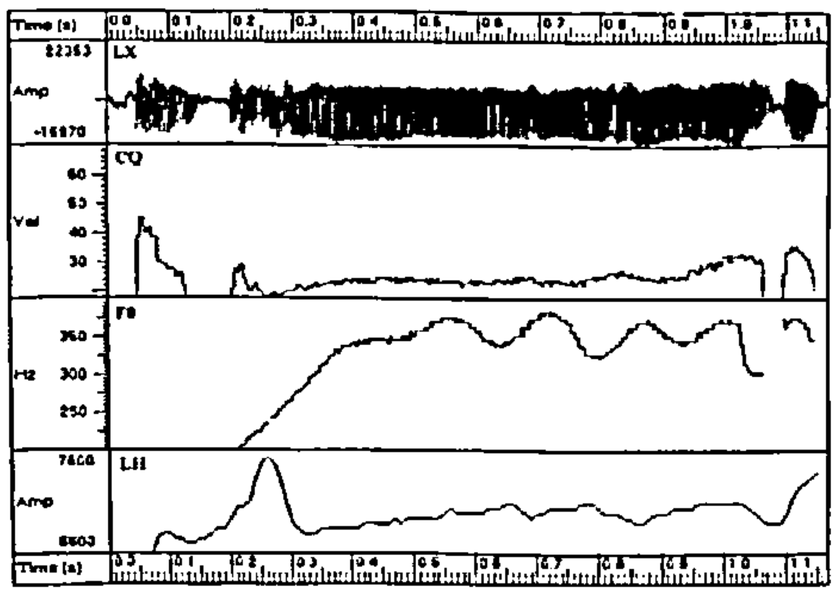
Figure 6.8 (page 1)



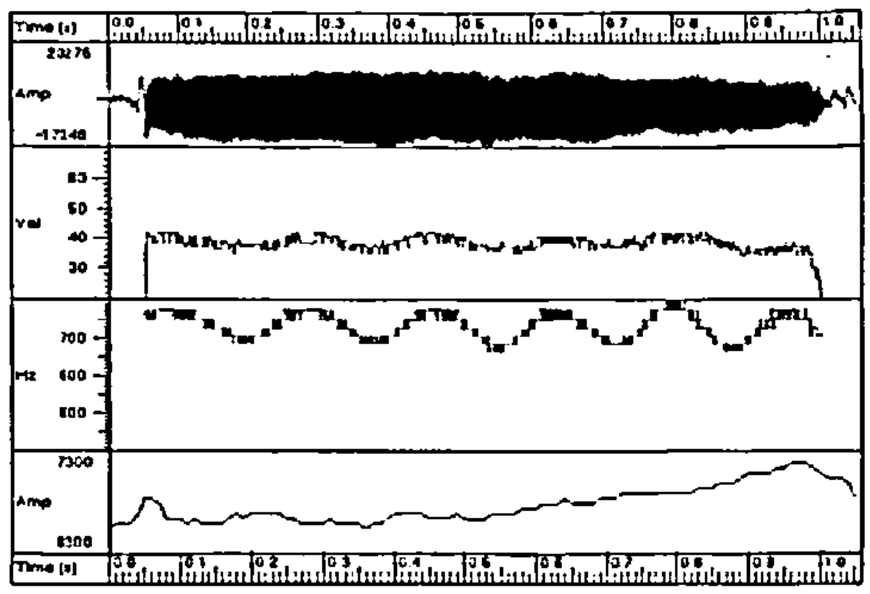
TT(o)E4



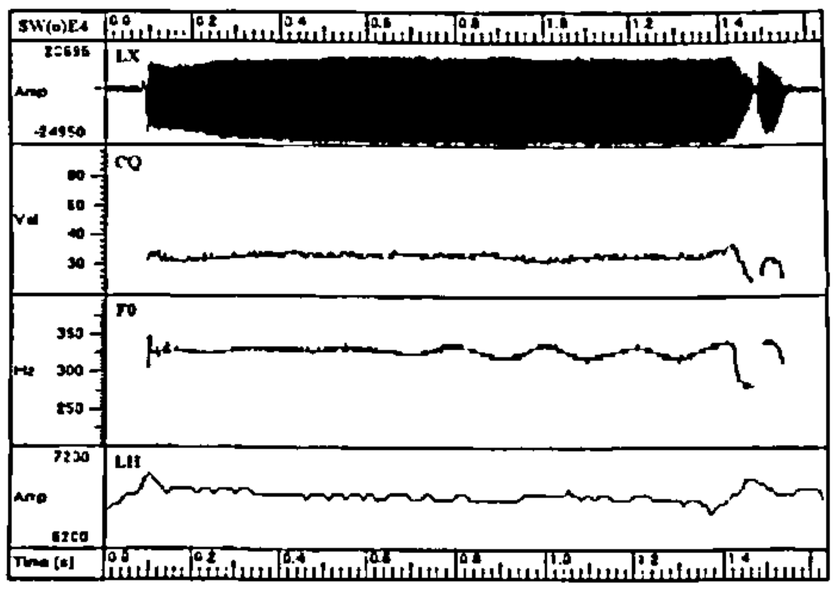
TT(o)E5



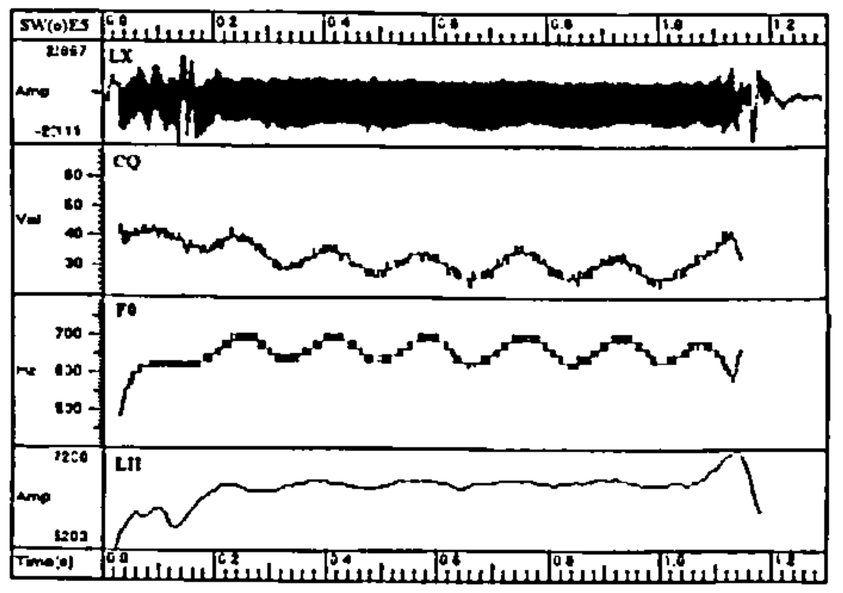
AW(o)E4



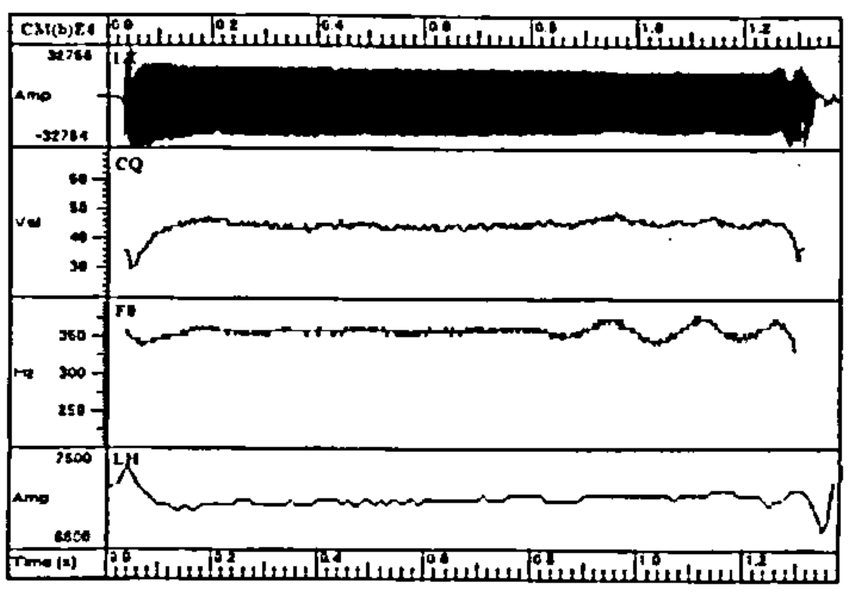
AW(o)E5



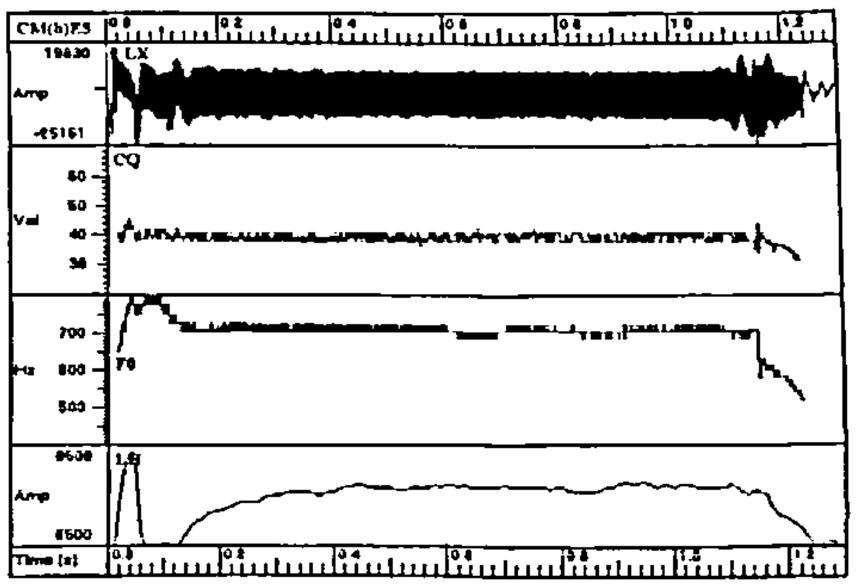
SW(o)E4



SW(o)E5

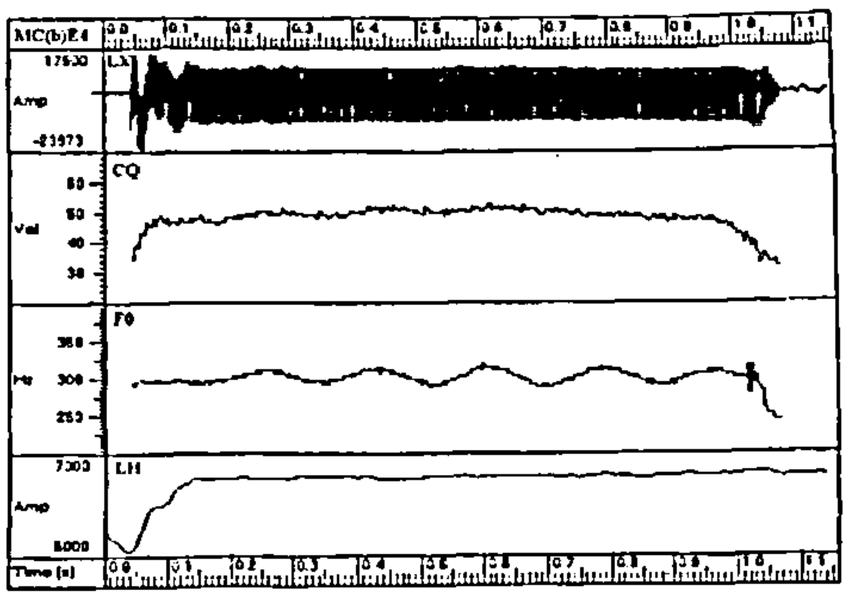


CM(b)E4

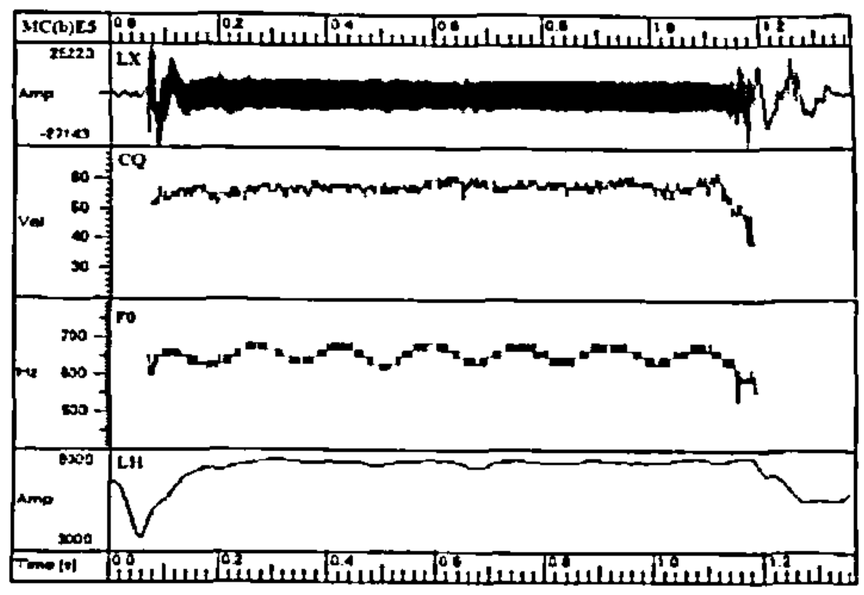


CM(b)E5

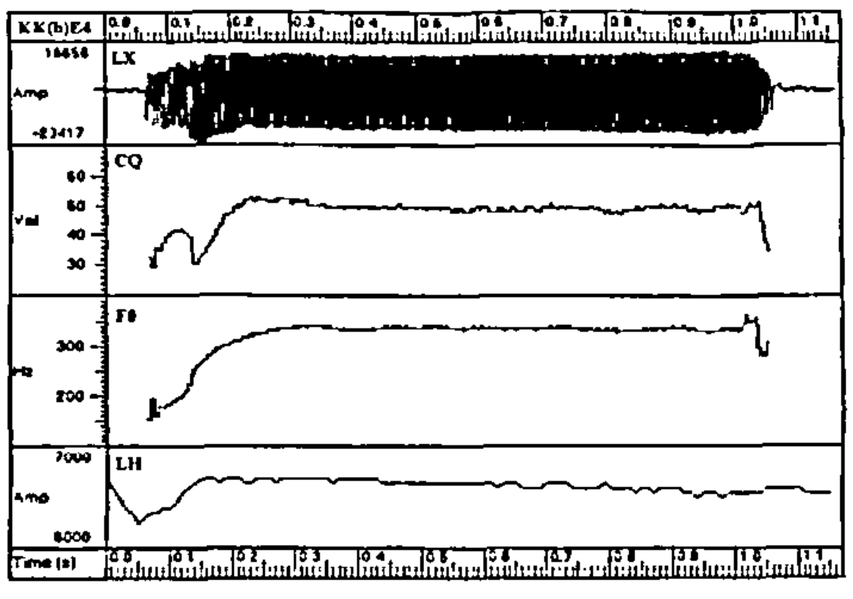
Figure 6.8 (page 2)



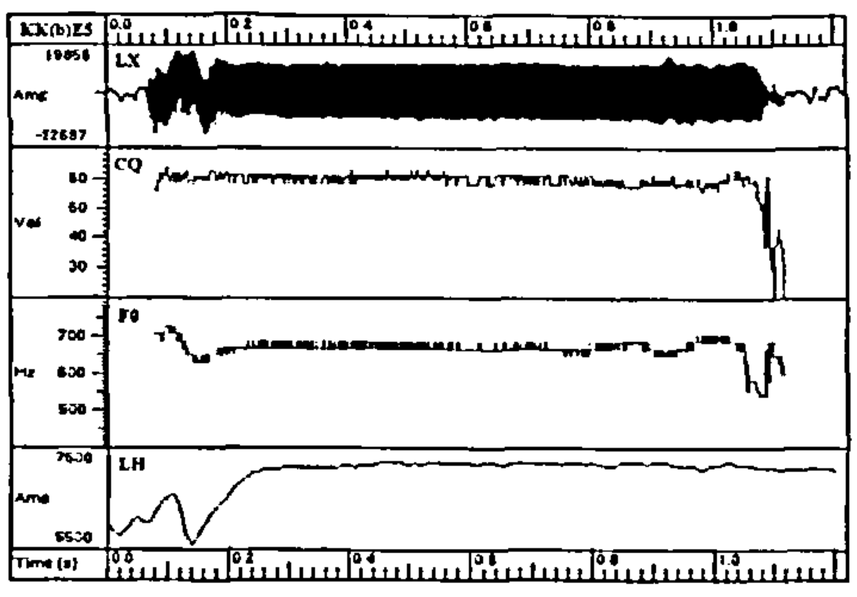
MC(b)E4



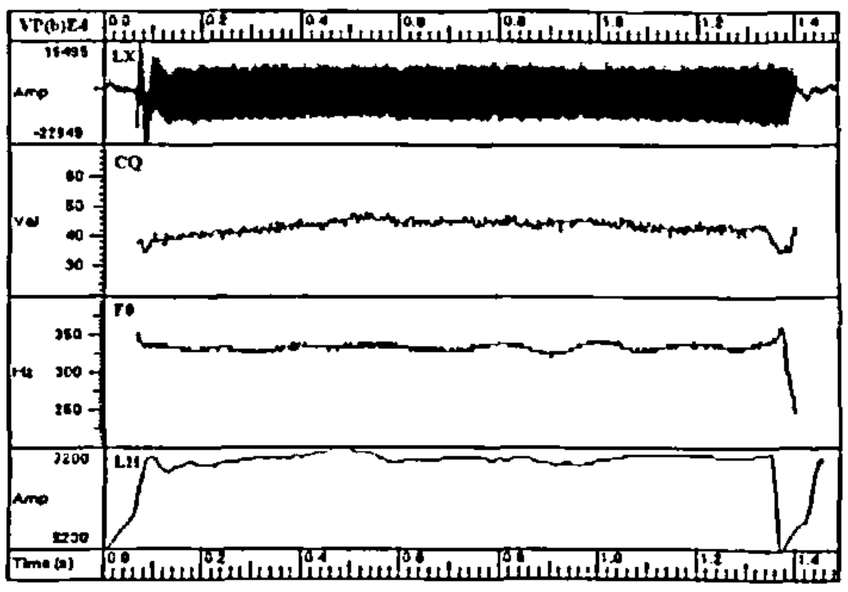
MC(b)E5



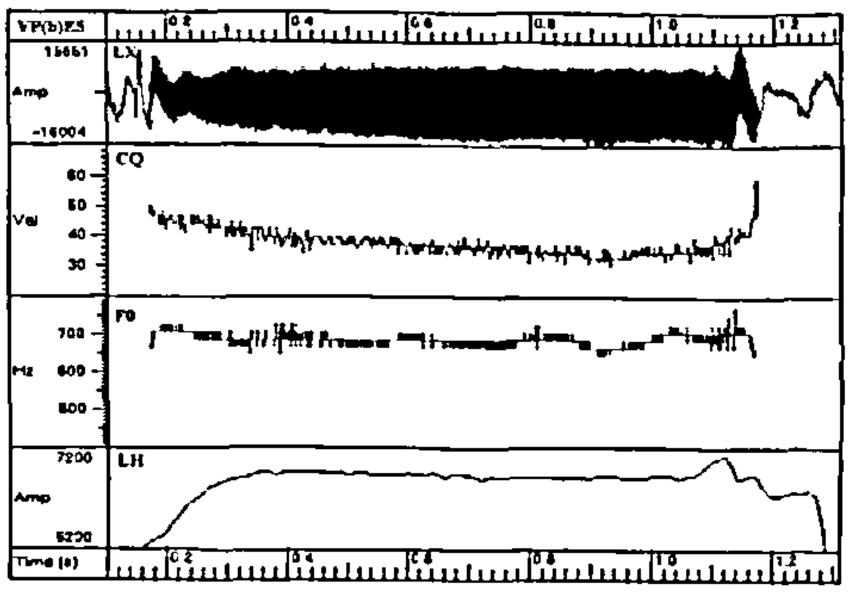
KK(b)E4



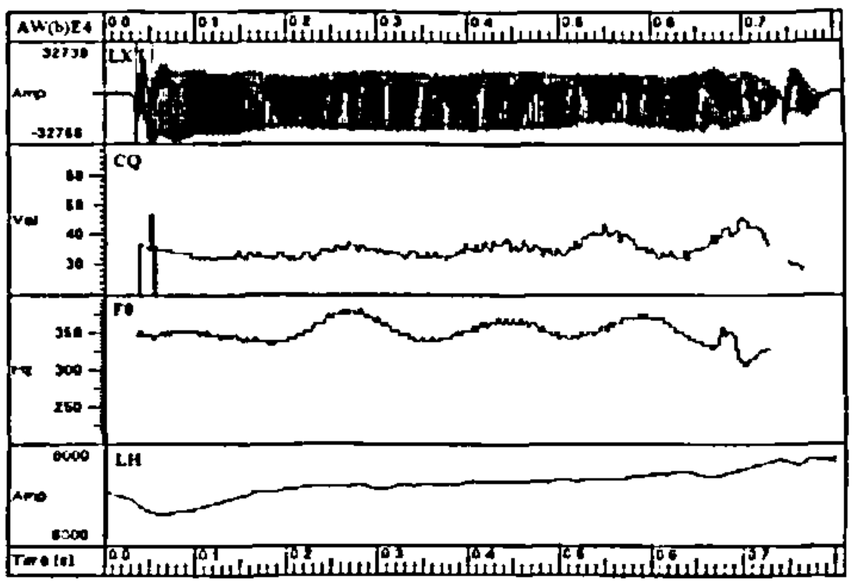
KK(b)E5



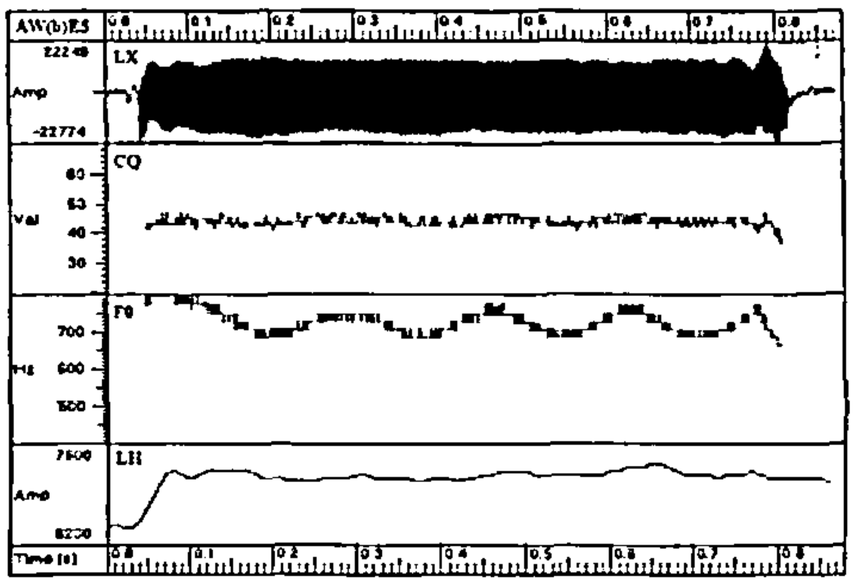
VP(b)E4



VP(b)E5



AW(b)E4



AW(b)E5

Figure 6.8 (page 3)
Figure 6.8 (pages 1-3). Diagrams of the Lx waveform with CQ, F0, and larynx height (LH) analysis results from the production of the word "bard" on E4 and E5 for each singer.

6.4 Spectral Comparison of Opera and Belting

The average spectra of the three vowels /i:/, /3:/, and /a:/ extracted from the words bead, bird, and bard were obtained from the sung exercises from the sample of nine singers.

The average spectra for the vowel /3:/ (pronounced "ir") was chosen as the analysis data. This is a middle vowel where formant positions should in theory be reasonably equally spaced for the spoken voice. This should help facilitate discrimination of formant positions for singing voice qualities. The average spectra results for the spoken vowel are shown in figure 6.4.

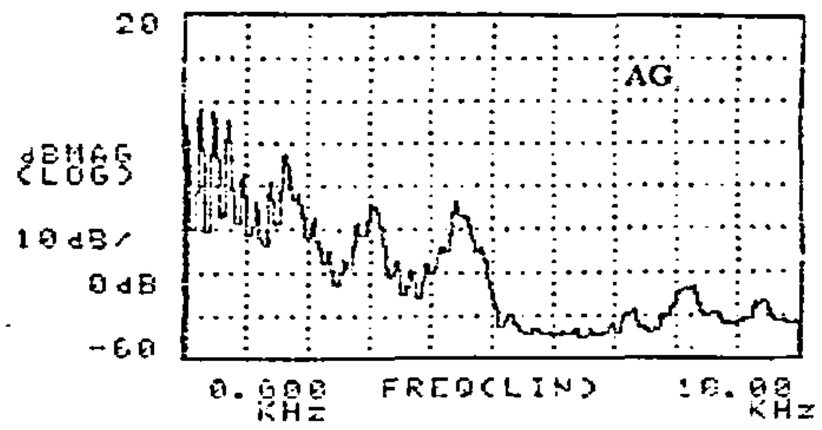
There are two ways of analyzing the data. One can look for relationships and patterns within each subject (inter-subject) or between subjects (extra-subject). Inter-subject analysis can show individual acoustic cues and pitch varying acoustic changes whilst extra-subject analysis allows comparison of these cues for acoustic modelling purposes.

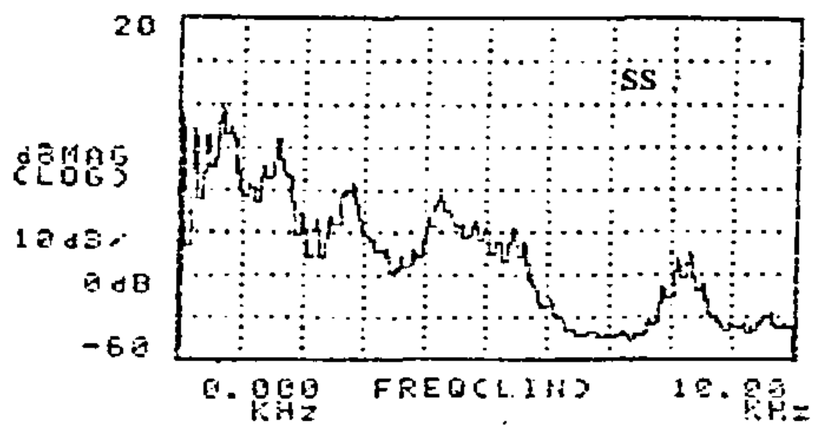
6.4.1 Comparison of Spoken Vowels

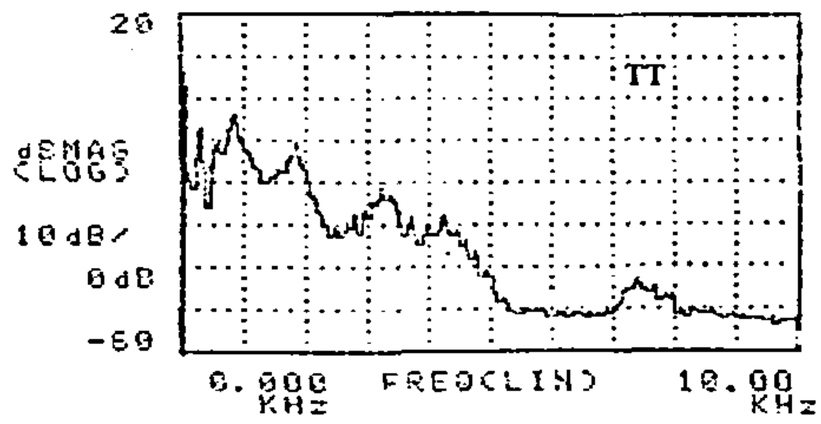
It is useful to first look at a spectrum analysis of the spoken vowels, shown in figure 6.9. The lower formant locations for the vowel /3:/ are clearly visible on the average spectra. They are roughly equidistant for each singer with the CQ values, listed in table 6.2, falling between about 30 -36%, lower than is expected for chest voice quality. Most of the singers are soft spoken with a breathy voice quality. This would account for the generally low CQ values shown below.

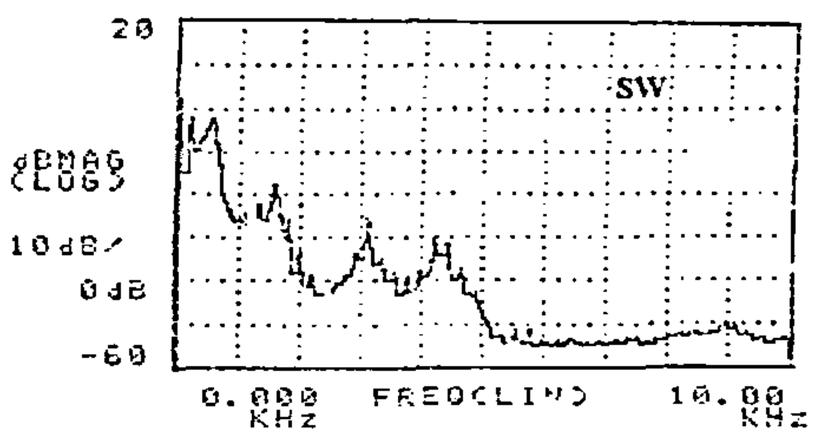
	/3:/	/a:/	/i:/
AG	31.96	30.71	35.57
SS	34.35	36.68	31.01
TT	36.27	37.33	36.51
SW	32.52	33.26	36.33
CM	33.80	33.52	35.57
MC	32.91	35.15	41.26
KK	35.53	37.38	37.14
VP	29.92	26.72	29.38
AW	32.62	32.29	35.77

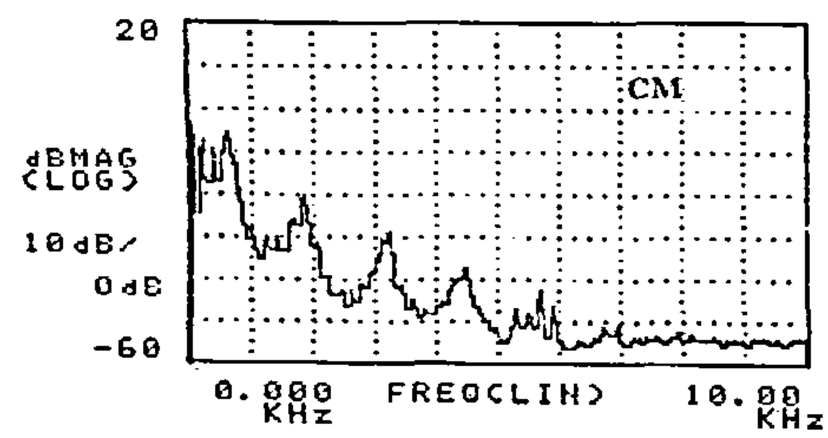
Table 6.2. Average CQ results for spoken /3:/, /a:/, and /i:/ for each singer.

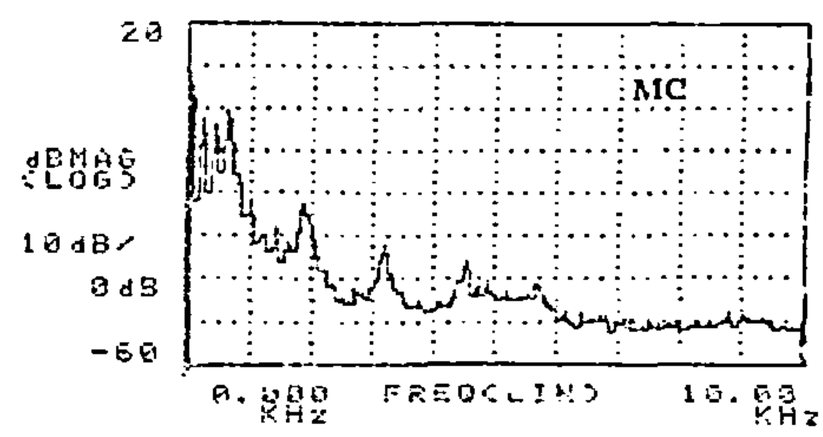


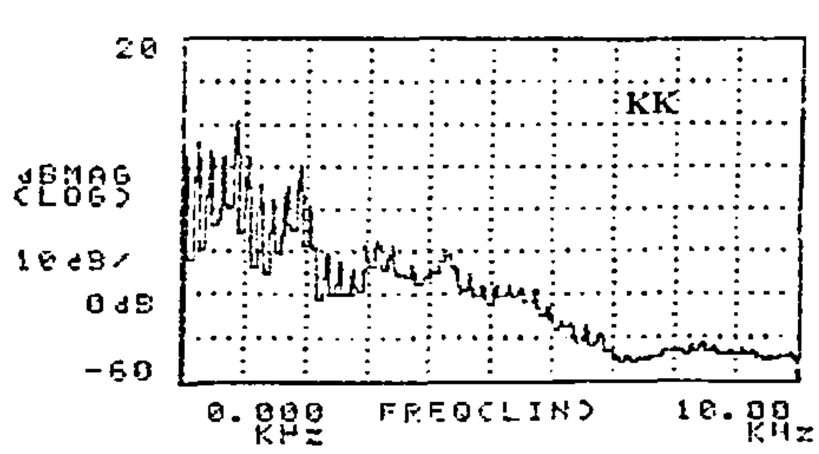


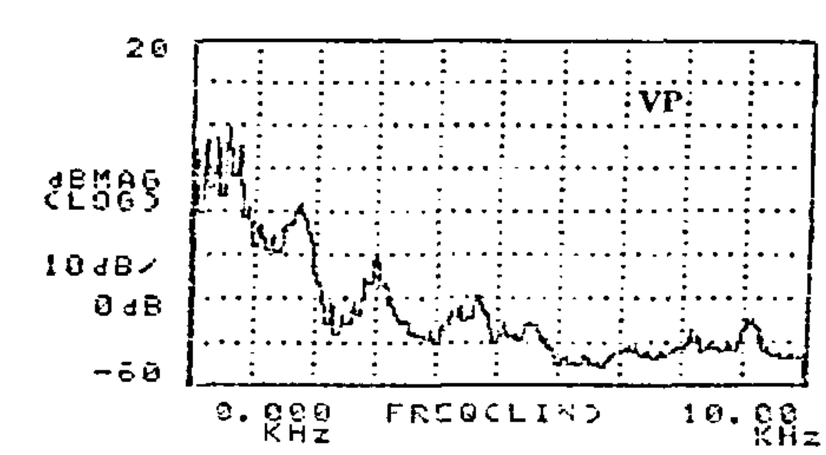












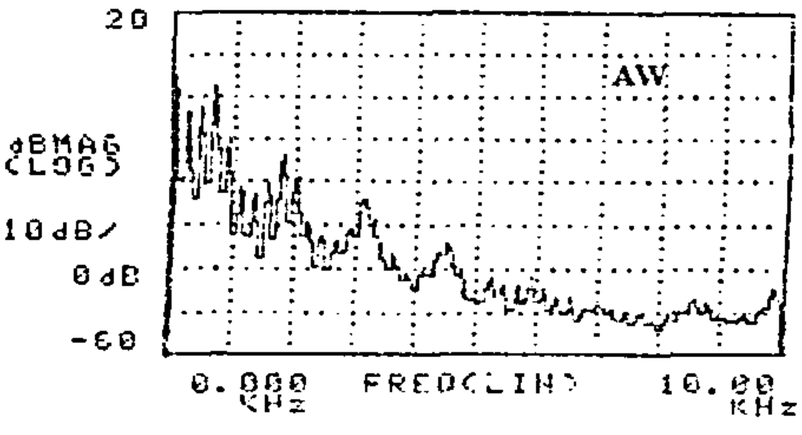


Figure 6.9. Average spectra for the spoken /3:/ vowel for each singer.

6.4.2 Analysis of the Opera G4/ɜ:/ Spectra

The spectra for the opera vowels (AG(o), SS(o), TT(o), SW(o), and CM(o)) from figure 6.10 share several characteristics:

1. the first 3 partials have the highest amplitude (excluding CM(o));

2. the first 2 partials are separated by at most 3 dB;

3. a characteristic pattern emerges - high energy for the first 3 or 4 partials with a dip in energy at about 2 kHz and a hump shaped amplitude curve from 2.5 kHz to about 4.5 kHz. The apex of this hump is at about 3 kHz, corresponding to the spoken 3rd formant position, and is around 12 dB on average lower than the highest partial, which compares to 21 dB lower for the spoken vowel; an increase in gain of 9 dB at about 3 kHz over the spoken vowel.

It is suggested that the first and second formants are shifted down in frequency from that of the spoken vowel, thus accounting for the high energy peak in the lower partials, whilst the third and fourth formants are grouped together, to give the energy hump peaking at around 3 kHz.

4. there is little energy above 4.5 kHz to 5 kHz. Although energy does show up in the spectrums above 6 kHz, rules governing masking suggest that it is not perceived.

6.4.3 Analysis of the Belted G4/ɜ:/ Spectra

The spectra for the first 4 belted vowels from figure 6.10 share several characteristics. However, the differences between these spectra and AW(b) suggest that AW(b) is not indicative of a belted tone, though perceptually it is similar. It is possible that this tone is produced with a "mixed" quality which sounds like belting.

For the first 4 belted vowels (CM(b), MC(b), KK(b), VP(b)) :

1. the 1st partial is much lower in amplitude than the 2nd partial (as observed by Miller & Schutte, 1993)

2. there is less amplitude decrease at 2 kHz compared with the spoken vowel, the high energy level is maintained;

3. there is more energy in the higher partials - extending upwards of 4.5 kHz for MC(b), and KK(b).

4. the average amplitude of the highest partial in the region of 3 kHz is only 7 dB lower than that of the highest partial overall (the 2nd partial) apart from in KK(b) where the highest partial is the one at about 3 kHz.

6.4.4 Analysis of the Opera E5/ɜ:/ Spectra

The spectra for the opera vowels (AG(o), SS(o), TT(o), SW(o), and CM(o)) from figure 6.11 share several characteristics:

1. there is little spectral energy above 4 kHz, that is, above the 6th partial (similar to the opera G4 tokens);

2. the fundamental has the highest energy (apart from AG(o));

3. there are similarities between the spectral envelopes. CM(o), TT(o) and SW(o) have very similar spectral envelopes; a peak at the fundamental with energy decreasing down to the 4th partial then a slight rise, the fundamental dominating the spectra, whilst AG(o) and SS(o) have the 1st and 2nd partials sharing dominance with a large drop in energy in the 3rd partial.

6.4.5 Analysis of the Belted E5/3:/ Spectra

From figure 6.11, VP(b) appears not to be a belted tone- the fundamental has spectral dominance and there is minimal energy above 5 kHz - this could be a "mixed" quality tone. CM(b) is also probably from a "mixed" production; even though high energy is maintained up to 2 kHz, it does not extend further than this.

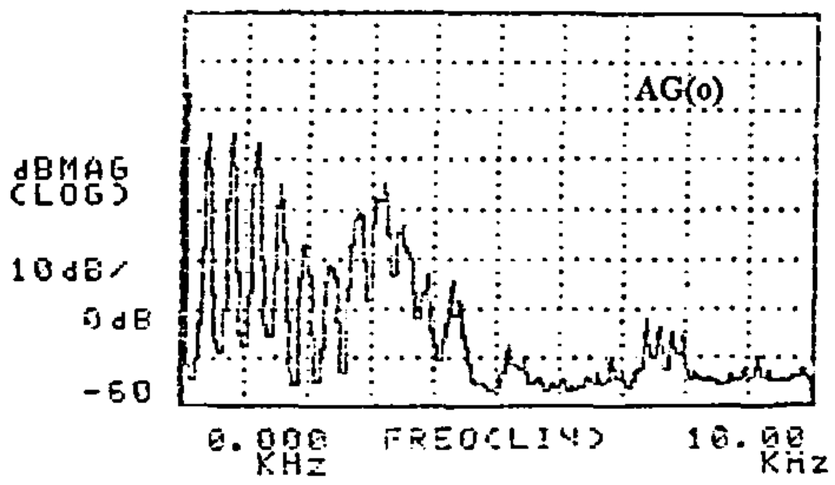
MC(b), KK(b), and AW(b) share several characteristics - spectral energy above 5 kHz (extending upwards of partial 8), and a quite consistent spread of energy in the first 7 partials, with the fundamental having less or equal dominance as the 2nd partial.

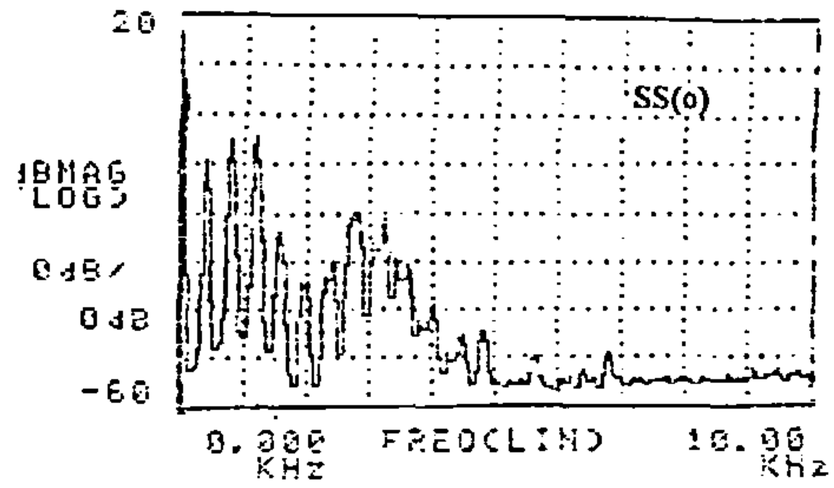
6.4.6 Discussion

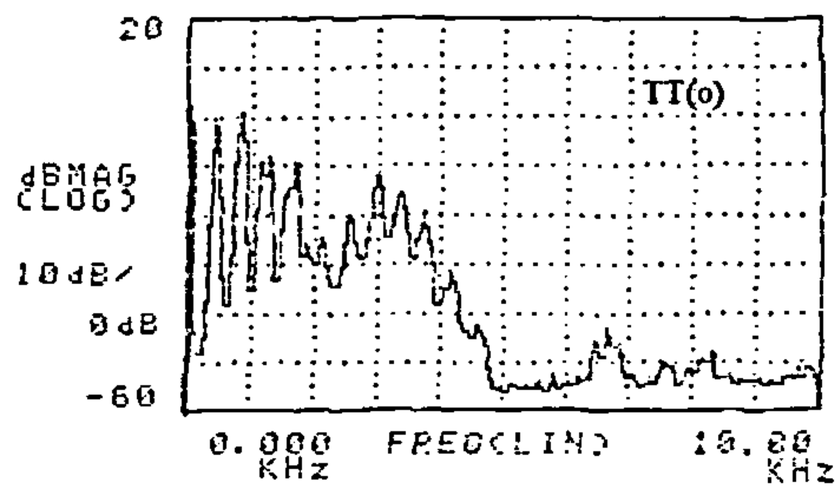
In opera quality, spectral dominance appears to be concentrated in two areas; the highest energy peak is in the lowest partial or partials below about 1.2 kHz, and there is a lower energy hump between about 2.8 kHz and 4 kHz.

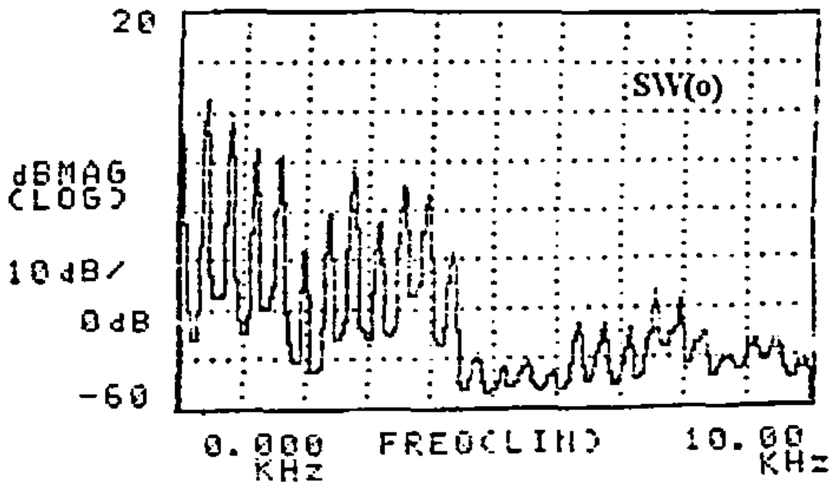
In belting, spectral dominance is concentrated especially in the 2nd partial and across a wide frequency range extending from the 2nd partial upwards, and there is little spectral tilt in the 3 - 4.5 kHz region.

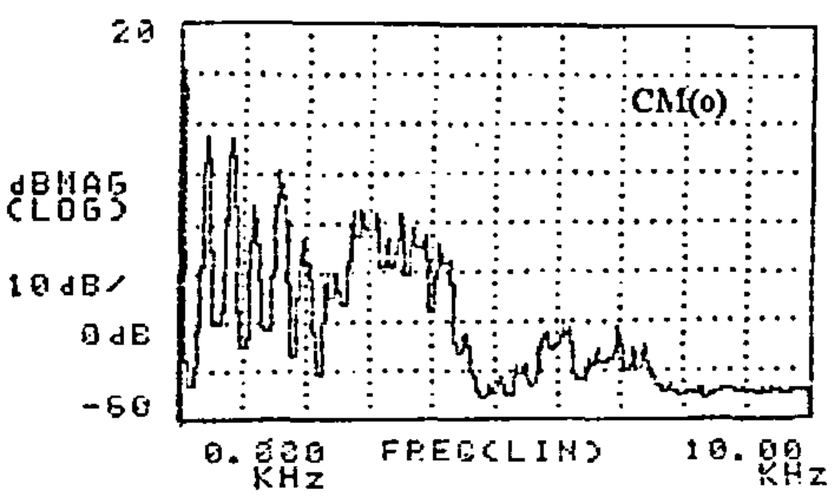
There is a little more variation in spectral patterning for belting - possibly this is because opera quality is a very specifically defined quality in terms of pedagogy and tradition - whereas for belt, there is not such a firm tradition, and singers are possibly allowed to incorporate a little of their naturalness in the belting sound (also suggested by Miller and Schutte, 1993)- in other words, belting may be a little more loosely defined, with there being more room for manoeuvre within the quality, though the high level of energy within the region 2 kHz to 4.5 kHz must be maintained, and also the frequency of the first formant must be raised to boost the energy of 2nd partial (Estill, 1988; Schutte & Miller, 1993).

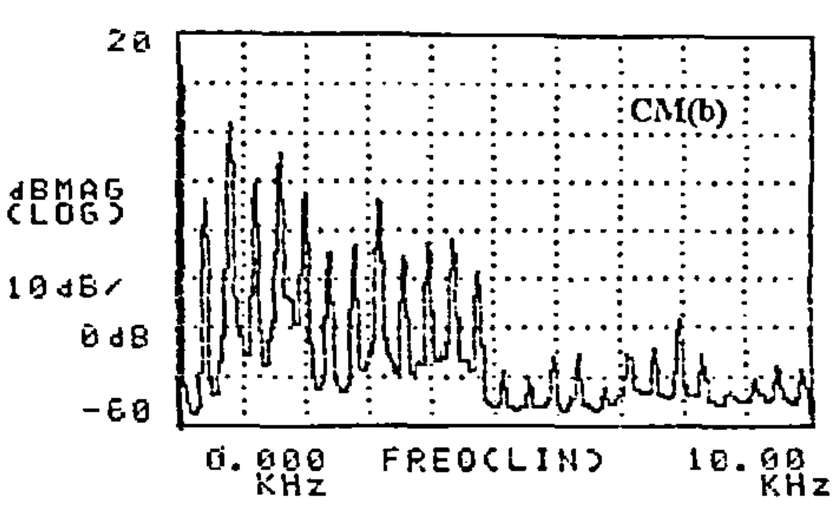


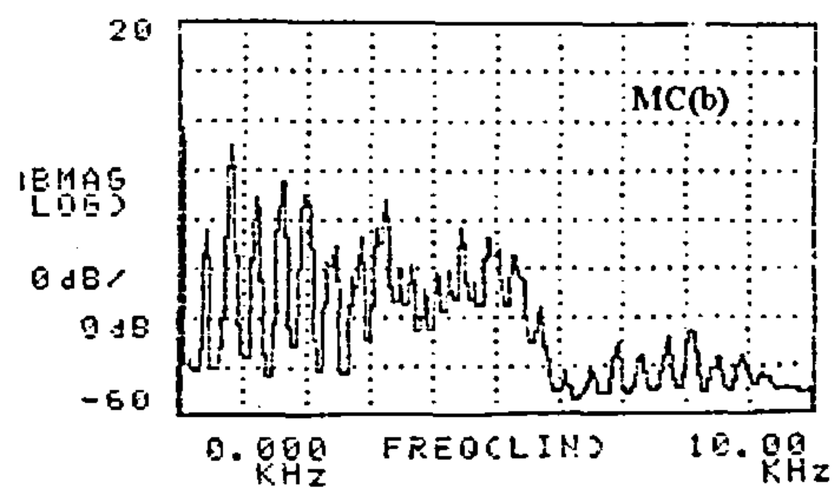


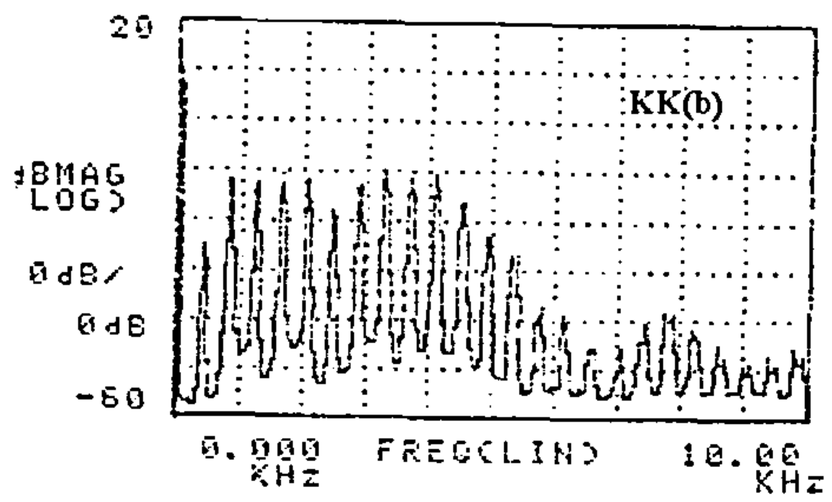


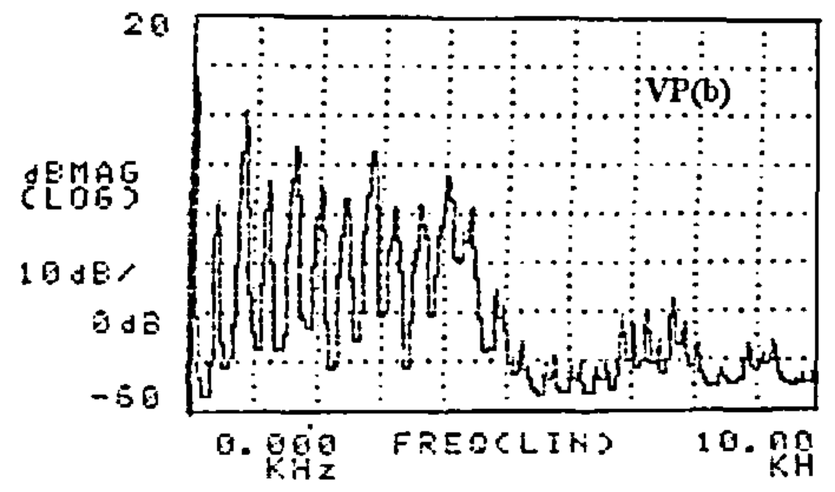












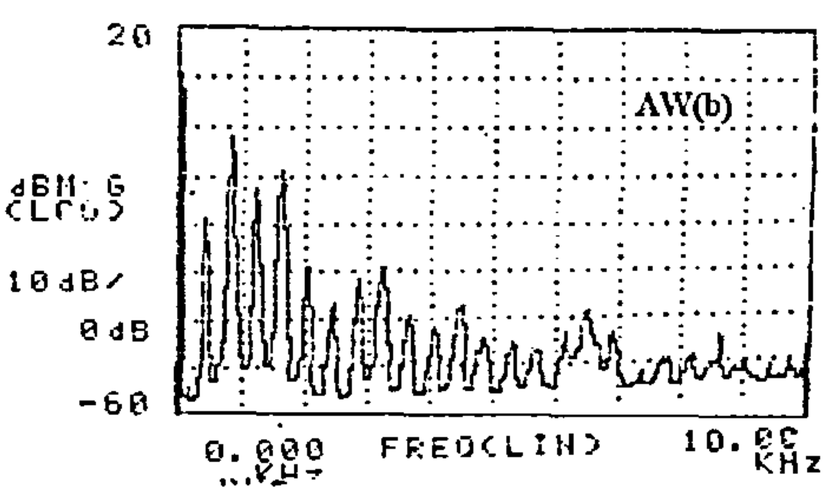
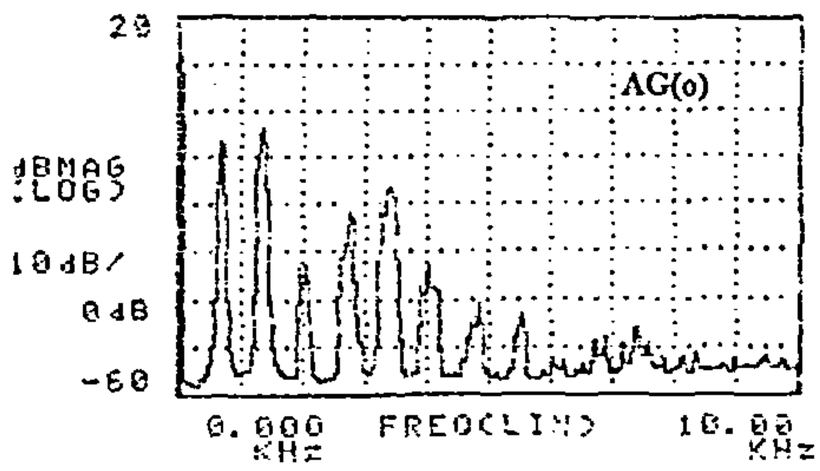
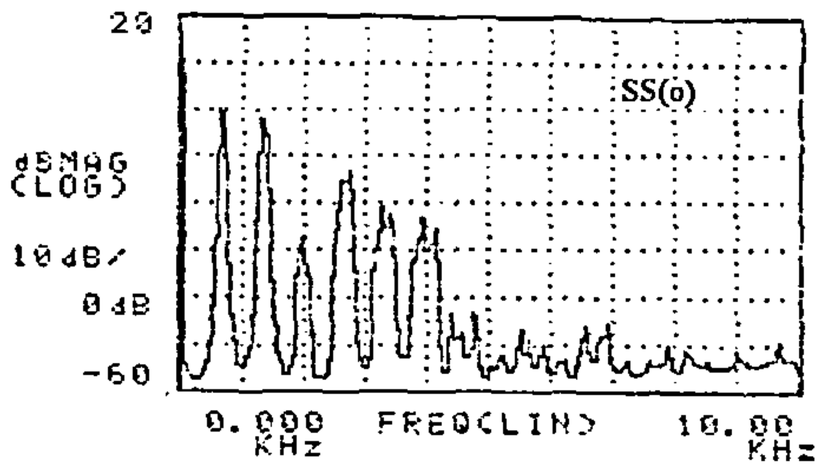
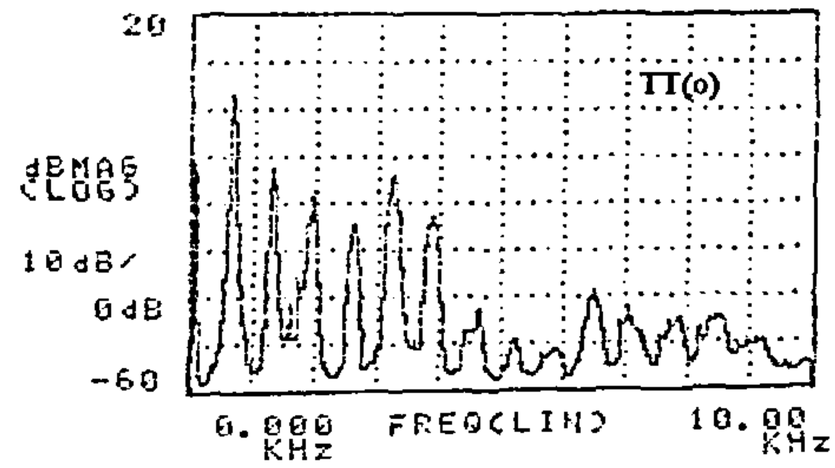
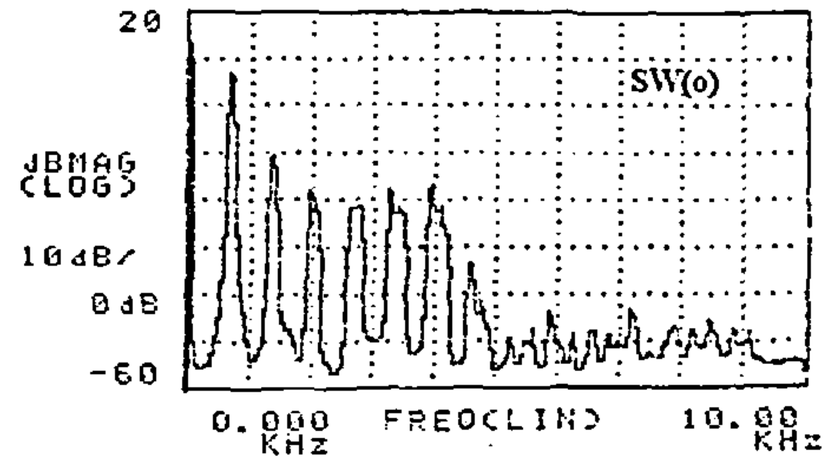


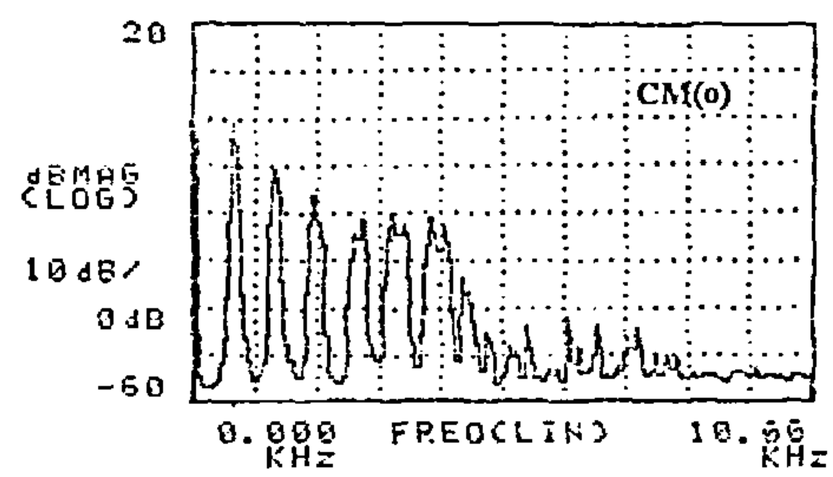
Figure 6.10. Average spectra for sung G4/3:/ tokens.

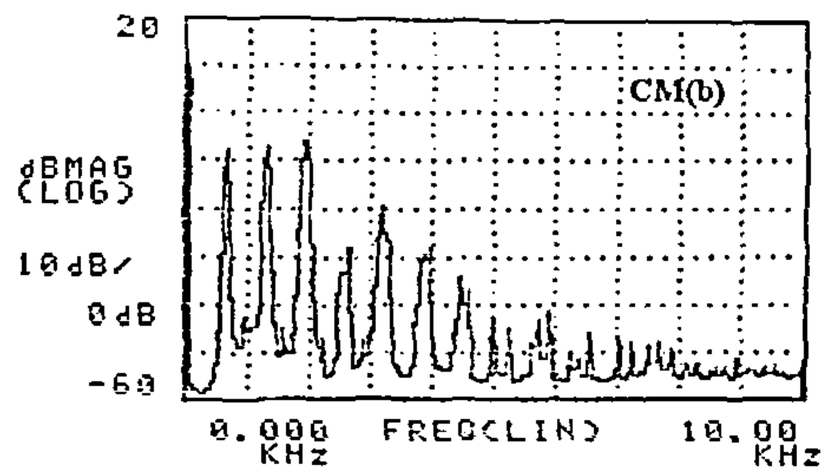


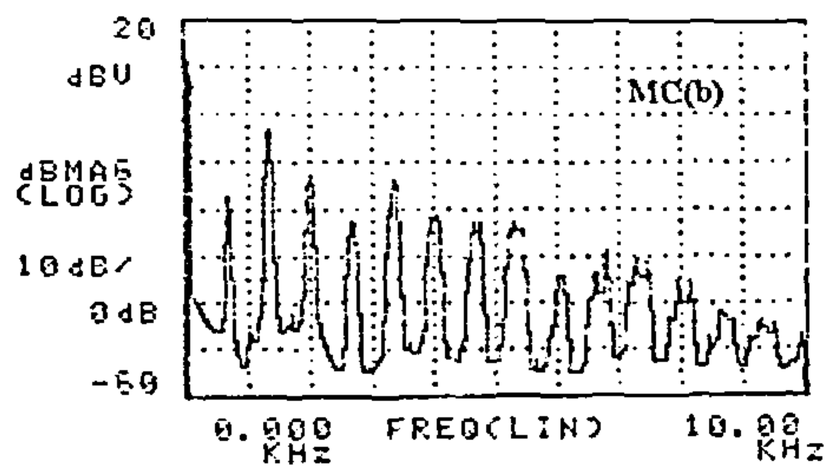


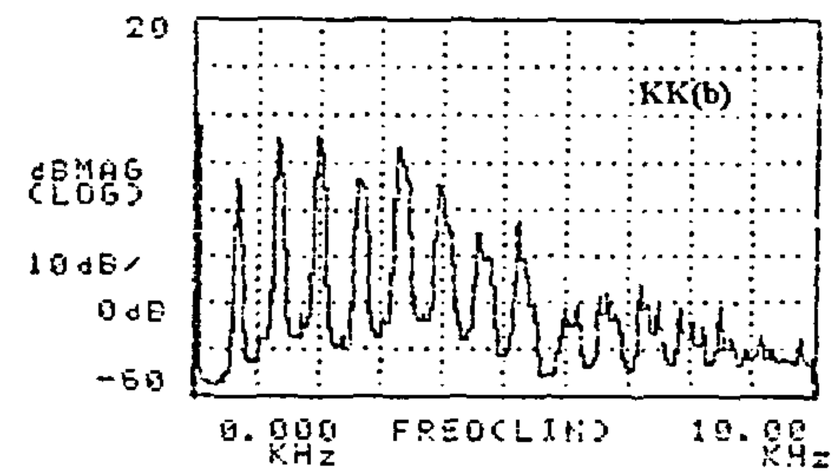


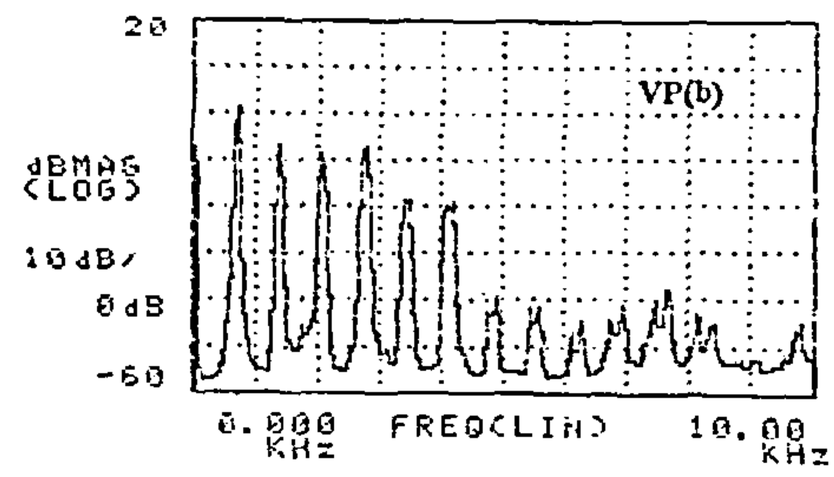












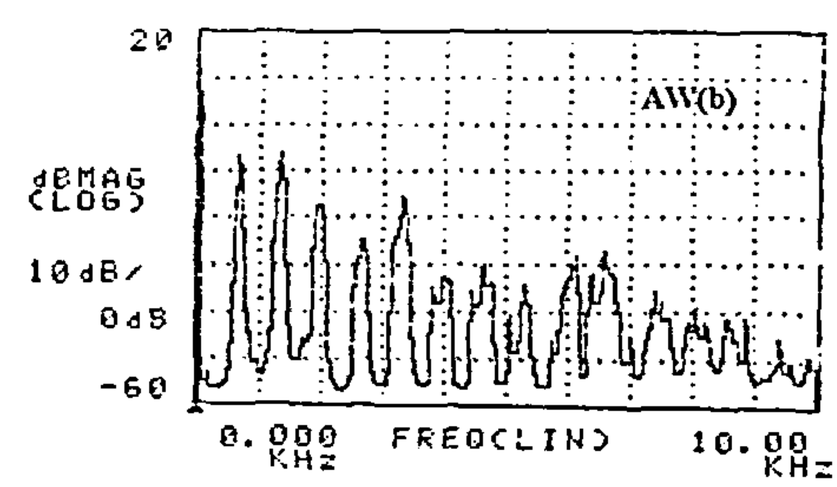


Figure 6.11. Average spectra for sung E5/3:/ tokens.

6.5 Larynx Height Differences Between Opera and Belting

Larynx height data (referred to as Lx-height) was extracted from a sample of singers comprising of four vocalisations on the word "bard". These were at larynx resting position (LRP), spoken, and sung at pitches E4 qualities. The method for extracting the Lx-height data is described in chapter 4.

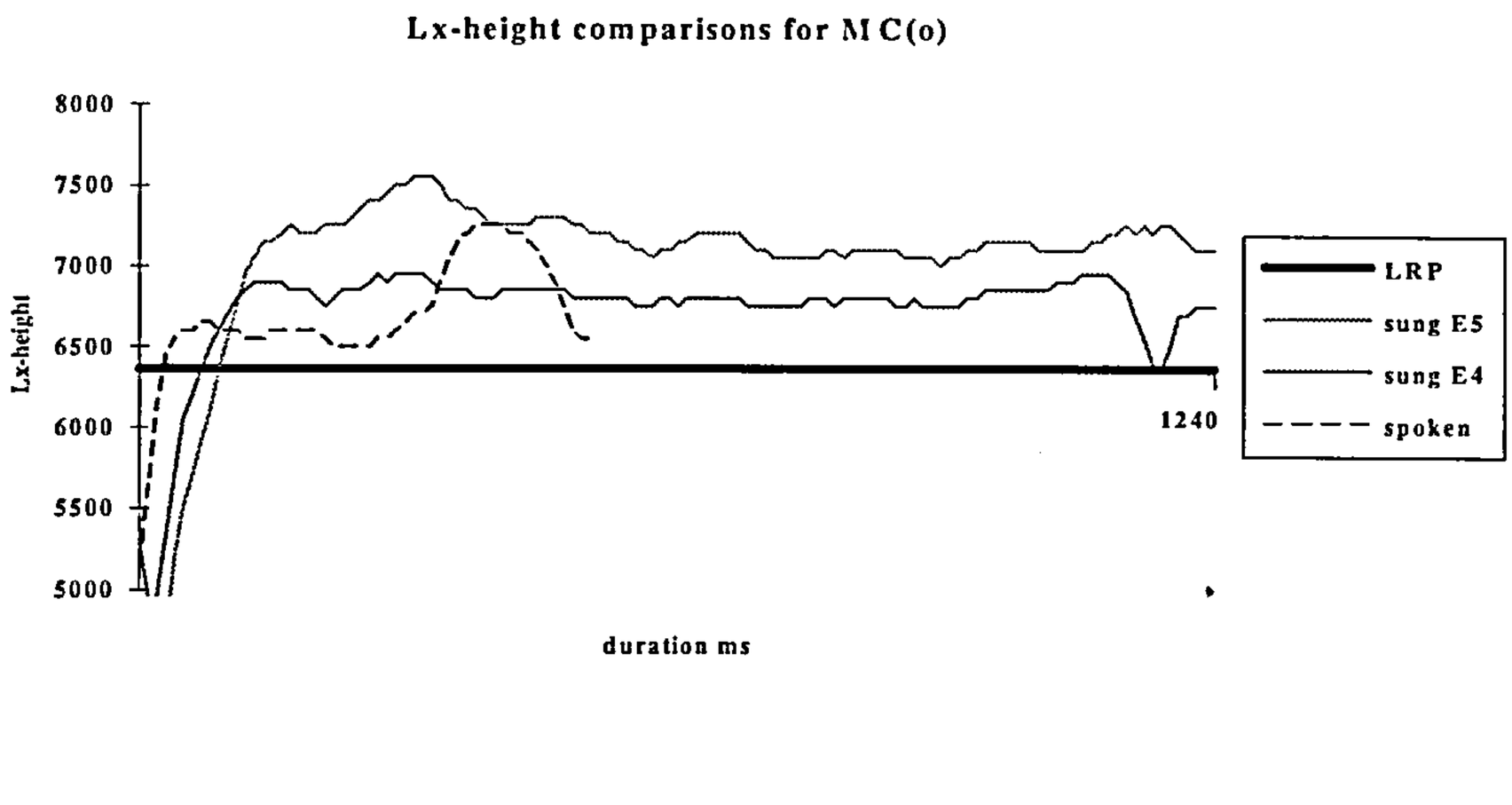
Figure 6.12 presents lx-height data for all the singers. The x-axis of each graph represents the average of the LRP. The two graphs and E5 in either belting or opera representing the opera singers SW(o) and AG(o) share several features; the spoken phonation has the highest Lx-height, at roughly the same distance above the singers' LRPs. The E4 phonation is produced with the larynx just a little above the LRP. This suggests that the singers are stabilising their larynx across the pitch range at a suitably comfortable height.

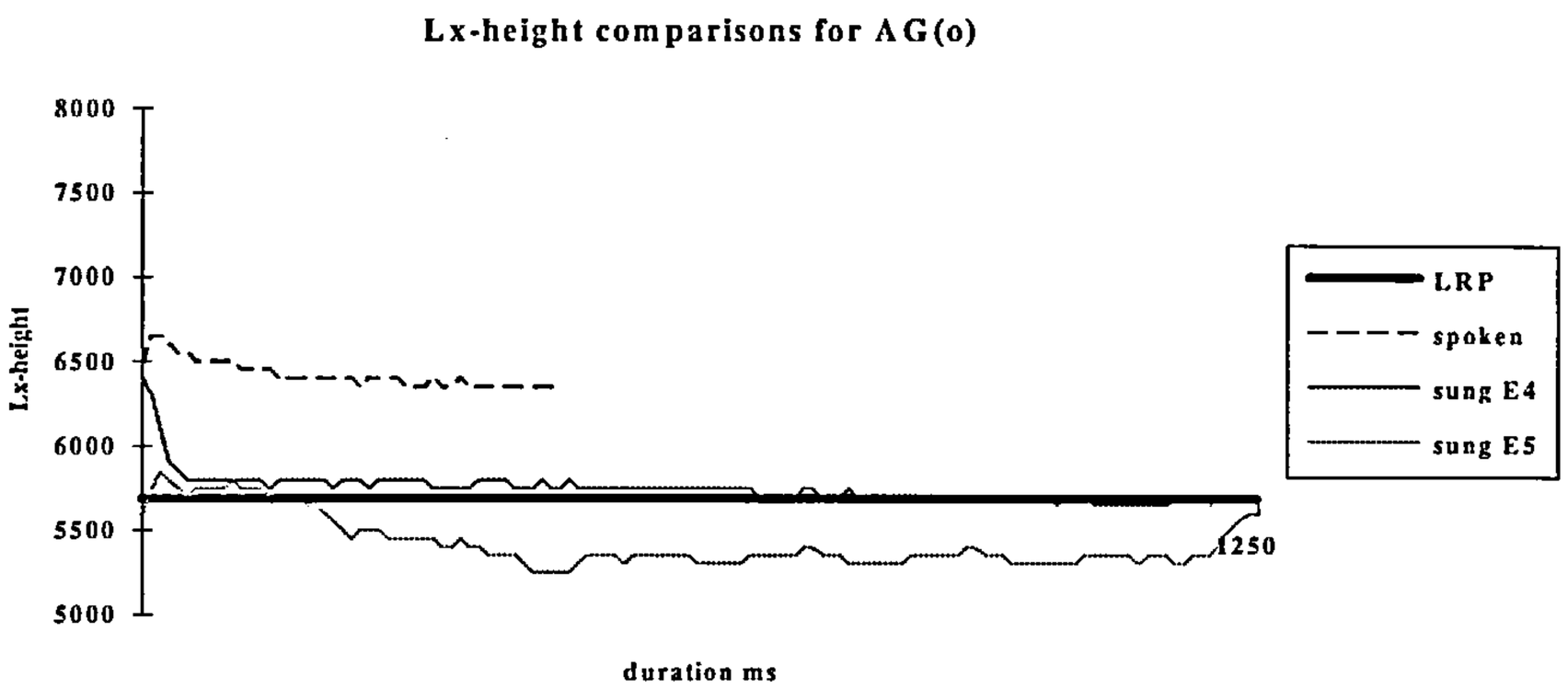
The graph for the KK(b) shows the opposite; all Lx-height positions lie above the LRP. Here the spoken phonation is the lowest of the three phonations, with the E5 phonation lying highest, and occupying a position which is comparable to the spoken phonations of the opera singers. The E4 phonation lies in between the spoken and the sung E5 phonations.

The data for the belters shows that all phonations are produced above the LRP, regardless of whether they are true belt quality sounds (some of the E5 phonations are not belted).

The data for the opera singers is more varied in general, the spoken word is produced with the highest larynx position, with the E4 phonation lying beneath this, but not necessarily below the LRP. The E5 phonation always lies below the LRP. Only in the case of one singer SS does the larynx appear to be anchored down appreciably; even so the E4 phonation has a higher larynx position than at E5.

Of the data for the singers who were asked to sing both operatically and also belt, singer MC shows an interesting feature. All phonations are produced above the LRP, and, as expected, the E5(b) phonation is sung with the highest larynx position (it is also quite unstable). However, lying just below it is the E5(o) phonation. All MC's phonations, whether it be opera or belting are produced with an elevated larynx. At the time of recording she had been singing in a West End show requiring her to belt. Possibly she was applying some physiological aspects associated with belting to her opera singing since that was what she was used to at the time. This is in contrast to singer CM, who had just finished singing an operatic part in a West End musical show. Her operatic phonations are at LRP whilst her E4(b) phonation is roughly at spoken level and her E5(b) phonation is much more elevated. Singer AW, who was also asked to sing in belting and opera, shows another different feature in her attempt at belting. Both her E4(b) and E5(b) phonations do not rise above her spoken level, suggesting that she is not belting at the higher pitch. This together, with a low CQ for this phonation, and an uncharacteristic spectum for the vowel /3:/ (which is taken from the previous word in the exercise) provide good evidence for a production which is neither operatic, nor belt. It is most likely to be "mixed" which is described below.





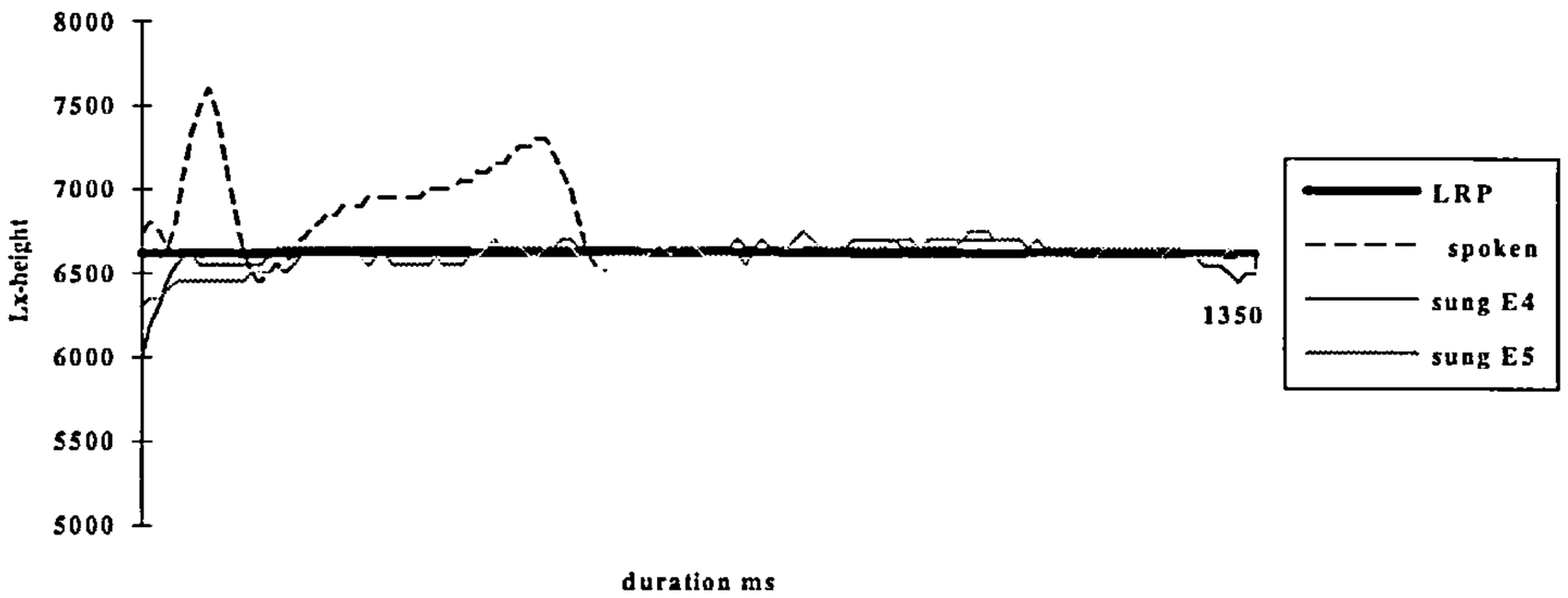
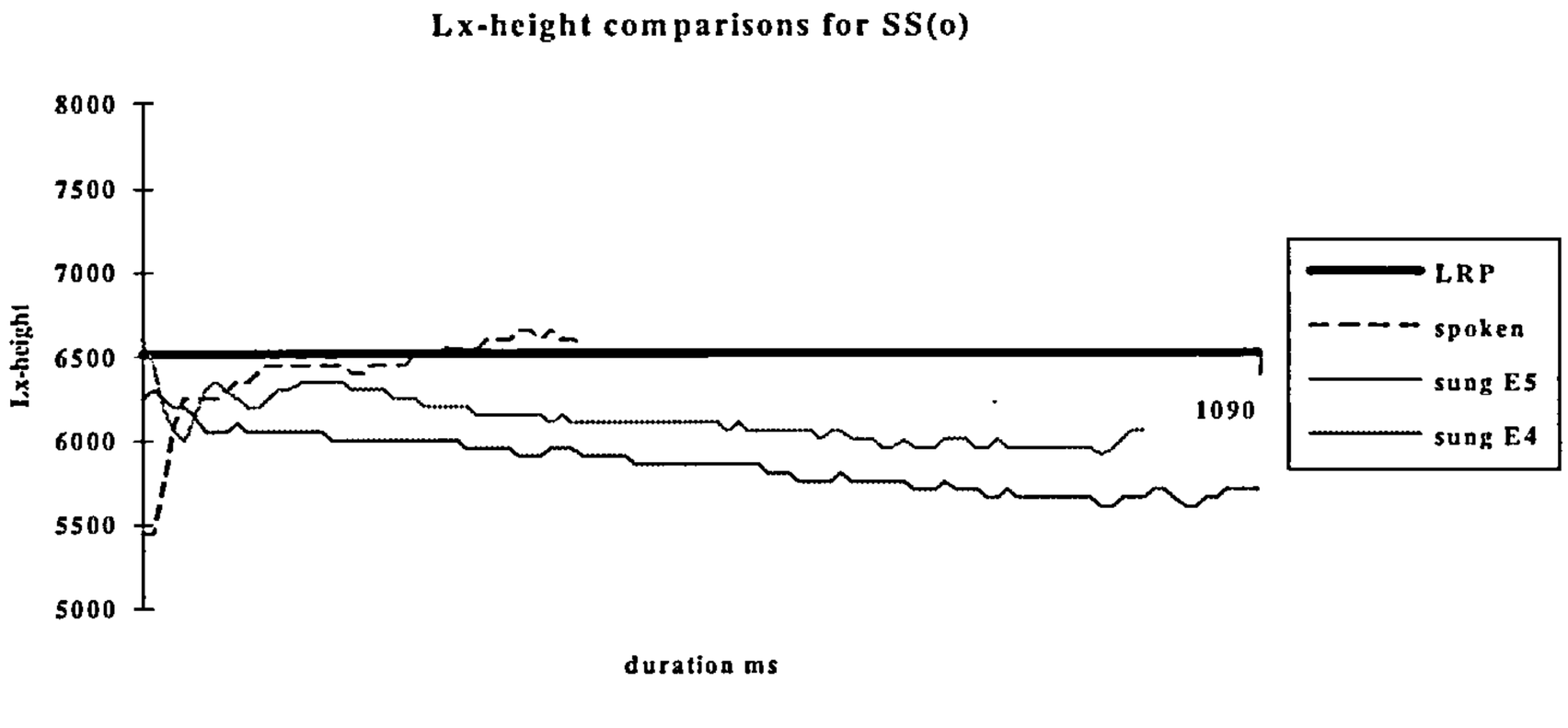
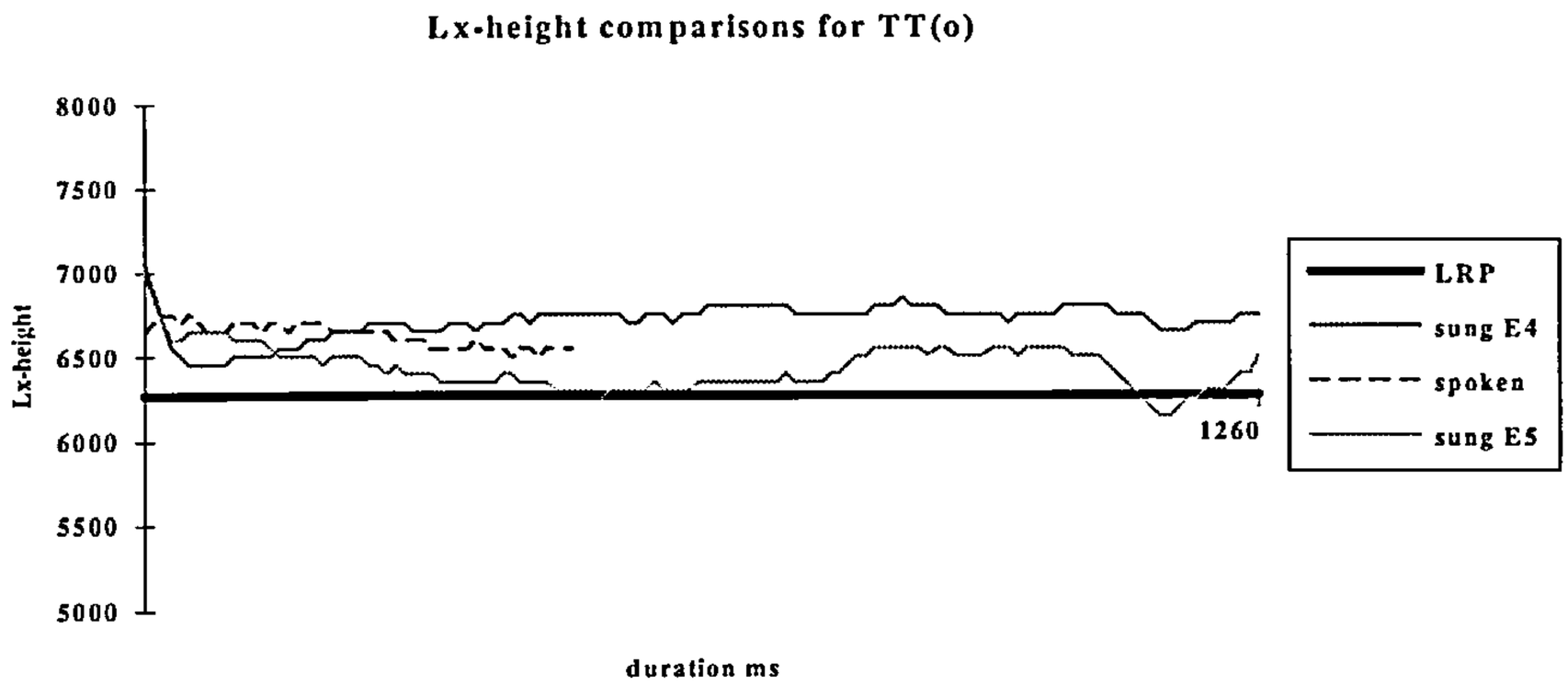


Figure 6.12 (page 1)





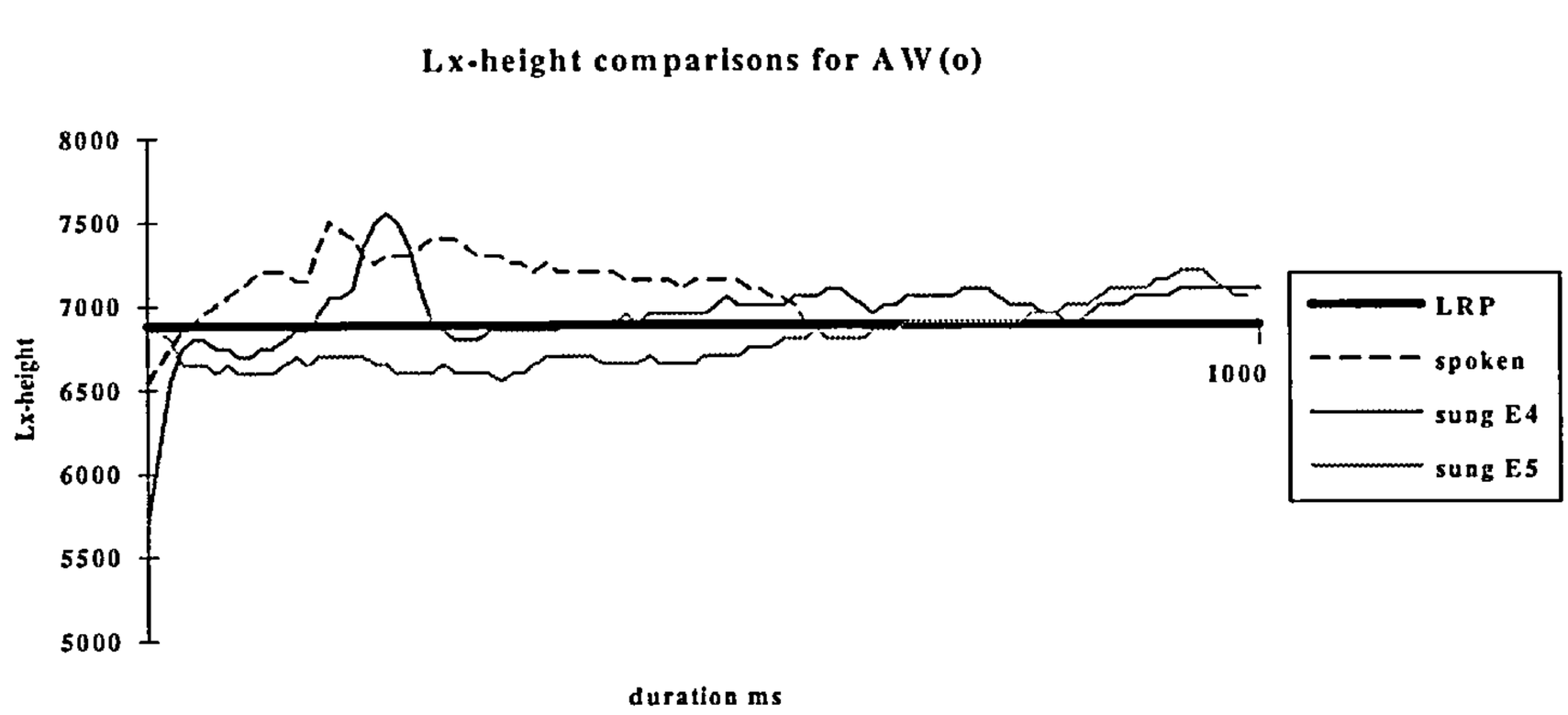


Figure 6.12 (page 2)

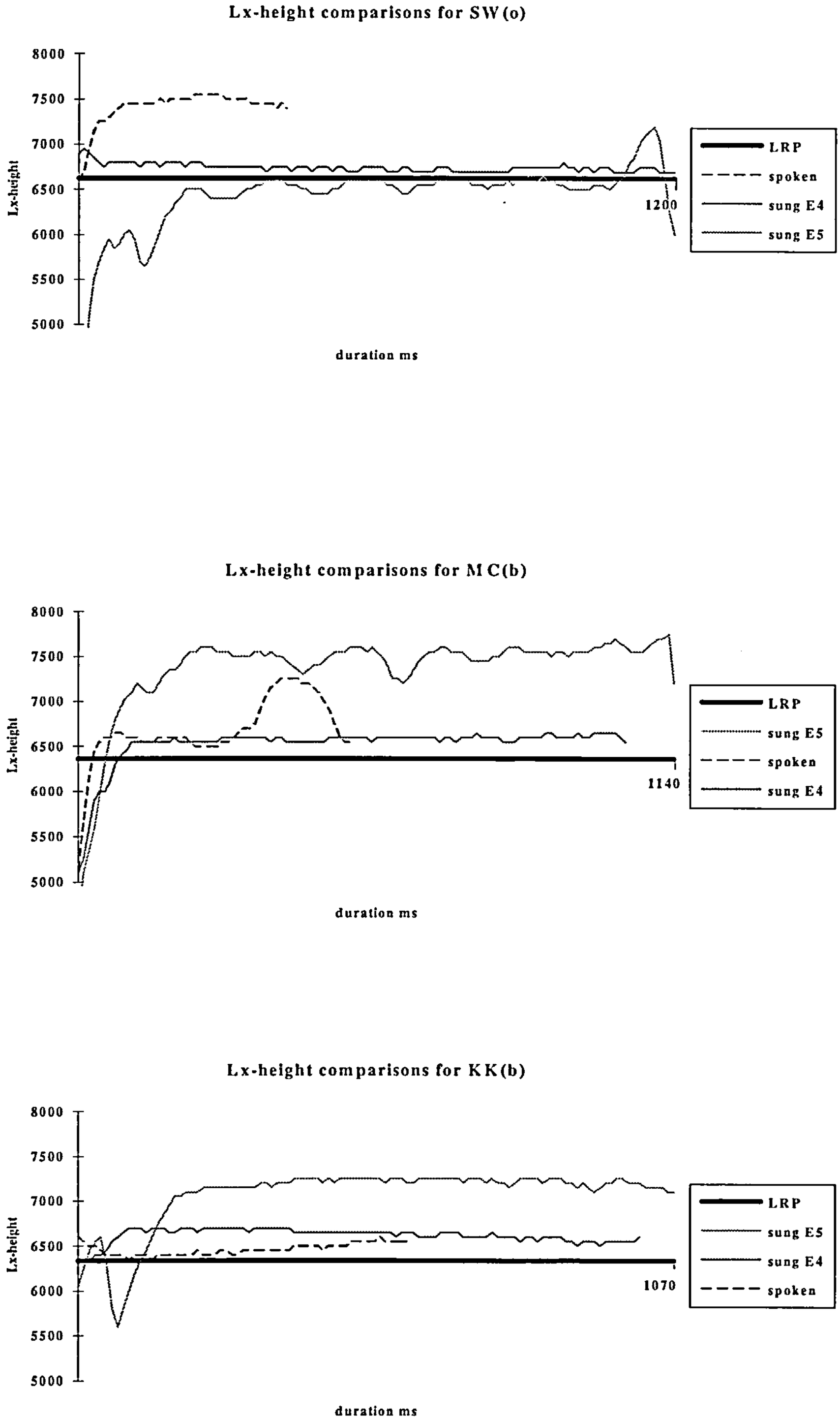


Figure 6.12 (page 3)

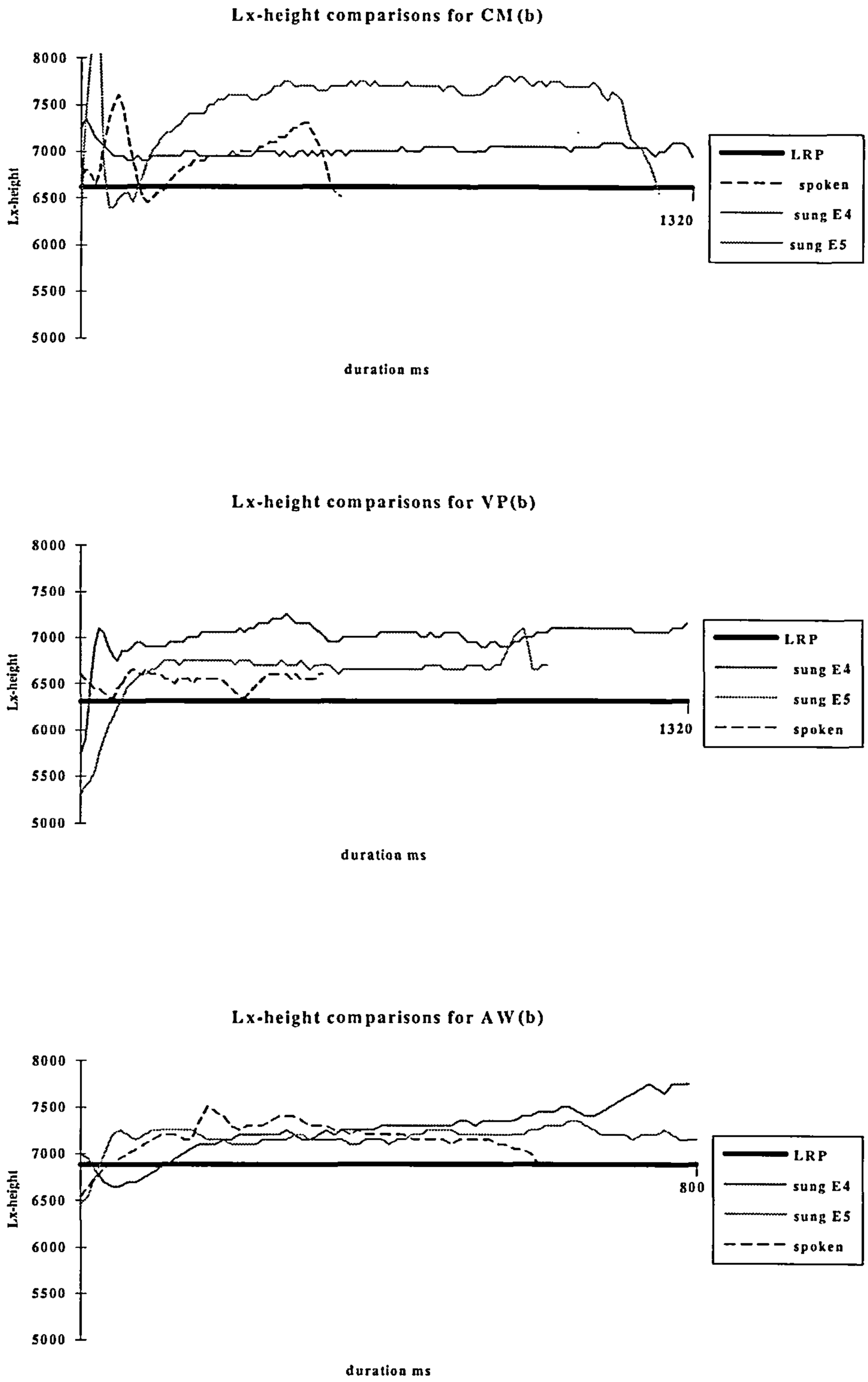


Figure 6.12 (page 4)

Figure 6.12 (pages 1-4). Larynx height comparisons for four opera singers and five West End musical singers comparing the spoken word "bard" with the sung word on pitches E4 and E5.

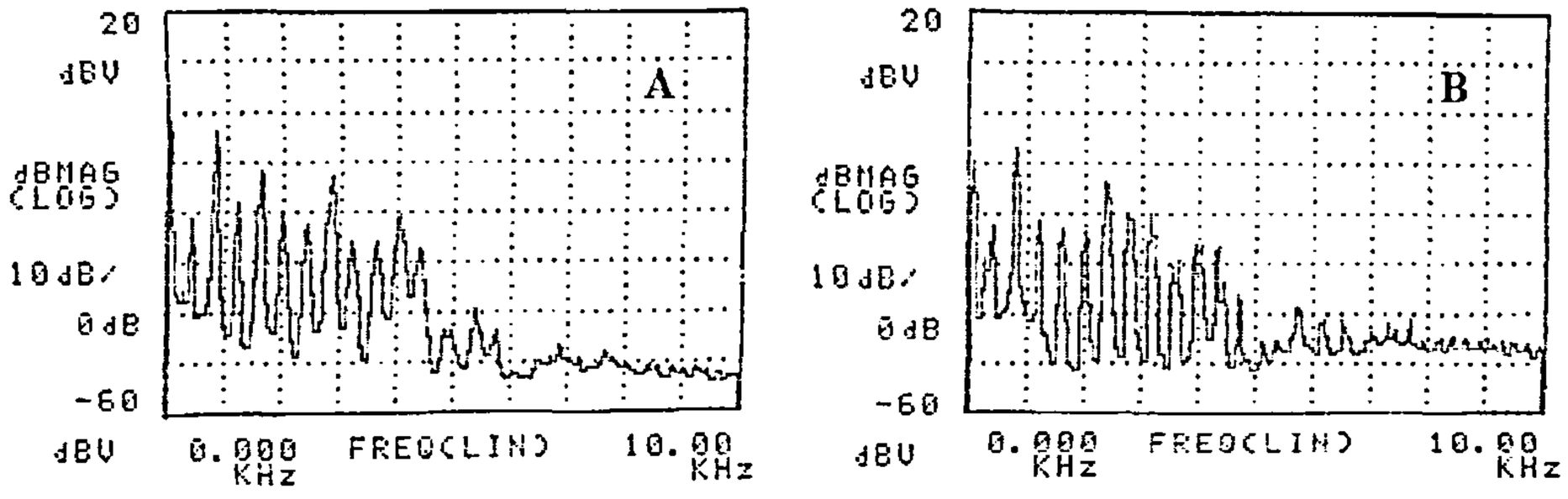


Figure 6.13. Average spectral comparison of vowel-pitch token (belt G4/3:/) from exercises with similar from a song passage (belt G4sharp /a:/) in same singer.

Figure 6.13 presents a spectral comparison of a singer's sung tone from the exercises with a similar sung tone from a song passage. It shows that the spectral envelope is similar in both cases, and so it may be assumed that the tones extracted form the exercises are a reasonable reflection of the singer's vocal quality when performing a song.

6.6 Mixed Quality

The singing quality known as "mixed" (also known as "legit") which is an "intermediate" vocal quality between opera and belting and is perceptually similar to belting in the higher range is in vogue with music-theatre critics and trainers since it reduces the amount of strain on the larynx which can occur from trying to belt. Sundberg, Gramming, & Lovetri (1993) have carried out a single-case study on a professionally trained singer (the co-author, JL) who exhibited three differing vocal qualities; opera, belting, and "mixed". The results show that the three qualities can be defined by differences arising from the relative amplitudes of the spectrum partials, subglottal pressure, and formant frequency locations.

Differences existed between the relative amplitudes of the two lowest spectrum partials, and between those in the upper part of the spectrum. It was shown that in operatic quality the fundamental was strong whilst in belting it was so weak that it was almost missing. In the singer's formant region for opera and mixed qualities the partials were lower in frequency and had a much greater amplitudes than for belting.

For a phrase sung in each quality, the SPL of opera and mixed were similar, but for belting was 10 dB louder. This reflected the finding that subglottal pressure was lower for opera and mixed than for belt. The findings seemed to follow a relationship between SPL and subglottal pressure, where SPL was "a linear function of the log of the subglottal pressure".

The three singing qualities also showed differences in formant frequency locations. The first two formant frequencies were much lower in opera than in mixed or belt, and the second formant frequency was the highest in mixed.

Schutte & Miller (1993) provide an acoustic explanation for why "mixed" quality (what they call "legit") is potentially safer than belt. They state that

> "for open vowels the first formants rise higher than in speech in the middle range to keep F1 in the vicinity of the second harmonic. If vocal-fold function is allowed to relax into a "falsetto" adjustment, F1 can stay below, but close to the second partial, permitting a high but non-extreme larynx position. This is the basis of the so-called "legit" Broadway voice: a pretty, but nonetheless "open" sound in the middle range with text articulation seemingly not far removed from that of speech" (Schutte & Miller,1993).

Larynxes appear to move considerably more from the LRP in belting quality than in opera quality. For the belting group, for E4 phonations which are similar in Lx-height to the spoken phonations, it is suggested that these singers may not be belting fully, and are more probably using a quality closer to speech quality, possibly because it may be physiologically harder belting at this pitch.

6.8 Conclusions

It has been shown that standard (two-channel) speech technology techniques do provide some useful modelling cues for different singing qualities, such as vibrato rate, spectral differences, and closed quotient differences, though the addition of the larynx height parameter adds a significant contribution to defining a singer's voice production. It is suggested, then, that two-channel speech technology is adequate for describing a singer's vocal quality (in terms of the acoustical features of the vocal output) but falls short of adequately defining its production.

Chapter 7

Synthesis and Perception

7.1 Introduction

To test the robustness of acoustic models derived from analysis, sounds are synthesized using parameters obtained as a result of analysis, and then the results are evaluated perceptually. The investigation here is whether the analysis adequately distinguishes between two different vocal qualities. If the correct acoustic features used to describe each quality have been correctly identified, and have been synthesized faithfully, there should be a perceptual difference which should be measurable. A large number of control parameters are required to optimally resynthesize a speech signal. A consideration of the constraints of articulatory and aerodynamic systems upon the sound, and the timing of such processes as observed in human speech is of critical importance in determining the optimal design, usage and control rate of the parameters in the synthesis system. A description of the perceptual tests will follow descriptions of synthesis systems and perception.

7.2 Speech Synthesis

Two main methods are used to derive the parameters needed to drive speech synthesizers; synthesis by rule and synthesis by analysis. Synthesis by rule attempts to generate intelligible speech either by joining phonemes together using grammatical rules, or by analysing text to derive the input parameters for a synthesizer. Natural sounding synthesis is difficult to achieve using this technique since in English, the context in which a phoneme occurs can change the sound of the phoneme (Allen et al., 1987). However, complex programs do exist in order to generate the effects of context and co-articulation resulting in allophonic variation. The rest of this section describes synthesis by analysis since this is the chosen method of synthesis for this research.

7.2.1 Speech Synthesis By Analysis

Speech synthesis by analysis attempts to simulate the perceptual qualities of the original speech. It can be categorised into articulatory synthesis and spectral synthesis. Articulatory synthesis attempts to

simulate the movement of the vocal tract (Scully & Allwood, 1983), whilst spectral synthesis attempts to simulate the speech signal. Direct analysis of the speech signal is easier to achieve than modelling the vocal tract.

7.2.2 Formant Synthesis

Spectral synthesis is based on the source-filter theory of speech production which states that a quasi-periodic source or noise source can be used to drive a separate filter which imposes resonance responses on the source signal. Two main techniques are used; formant synthesis (Klatt, 1980; Klatt & Klatt, 1990), and diphone concatenation (O'Shaughnessy et al., 1988) which is based on linear prediction coding (LPC) (Atal & Hanauer, 1971).

Formant synthesis reconstructs the vocal tract transfer function by simulating the formant characteristics of the vocal tract. This is achieved by connecting a set of resonators (each resonator representing a formant) and anti-resonators (for nasals, fricatives, and plosives) together. These are then excited by a controlled excitation source (Fant, 1960). The excitation source simulates a voiced sound source or an unvoiced source, which is modified by resonators and anti-resonators resulting in a specified speech spectrum. "The advantage of this technique is that its parameters are highly correlated with the production and propagation of sound in the oral tract" (Styger & Keller, 1994).

The formant resonators act as band-pass filters (or pole filters) with resonance frequency and bandwidth control. The anti-resonators have the inverse characteristics; they act as band-stop filters (or zero filters). For speech, generally the first five or six formants are specified.

Typically two configurations of resonators are used in formant synthesis; parallel and cascade, shown schematically in figure 7.1. In the parallel configuration, each resonator is excited at the same time and has its own peak amplitude control. The parallel design successively adds the transfer functions of the individual resonators. In the cascade configuration, the resonators are connected in series which successively multiplies the transfer functions.

The parallel formant synthesizer (Holmes, 1983) attempts to model the acoustic signal directly from its spectral waveform offering a closer approximation to the real speech signal, than is possible for the cascade synthesizer. Its design is more suited to modelling consonants than vowels since it does not attempt to model vocal tract behaviour. Holmes (1983) states, that for [parallel] formant synthesizers,

> "..it is assumed... that the aim is to approximate as closely as possible to those features of speech signals that are perceptually significant, with no intrinsic importance being attached to the relationship with the human speech production mechanism. It seems to be generally accepted that to achieve this aim it is sufficient to reproduce the short-term spectrum of the speech, defined with a frequency and time resolution similar to that of the human auditory system" (Holmes, 1983).

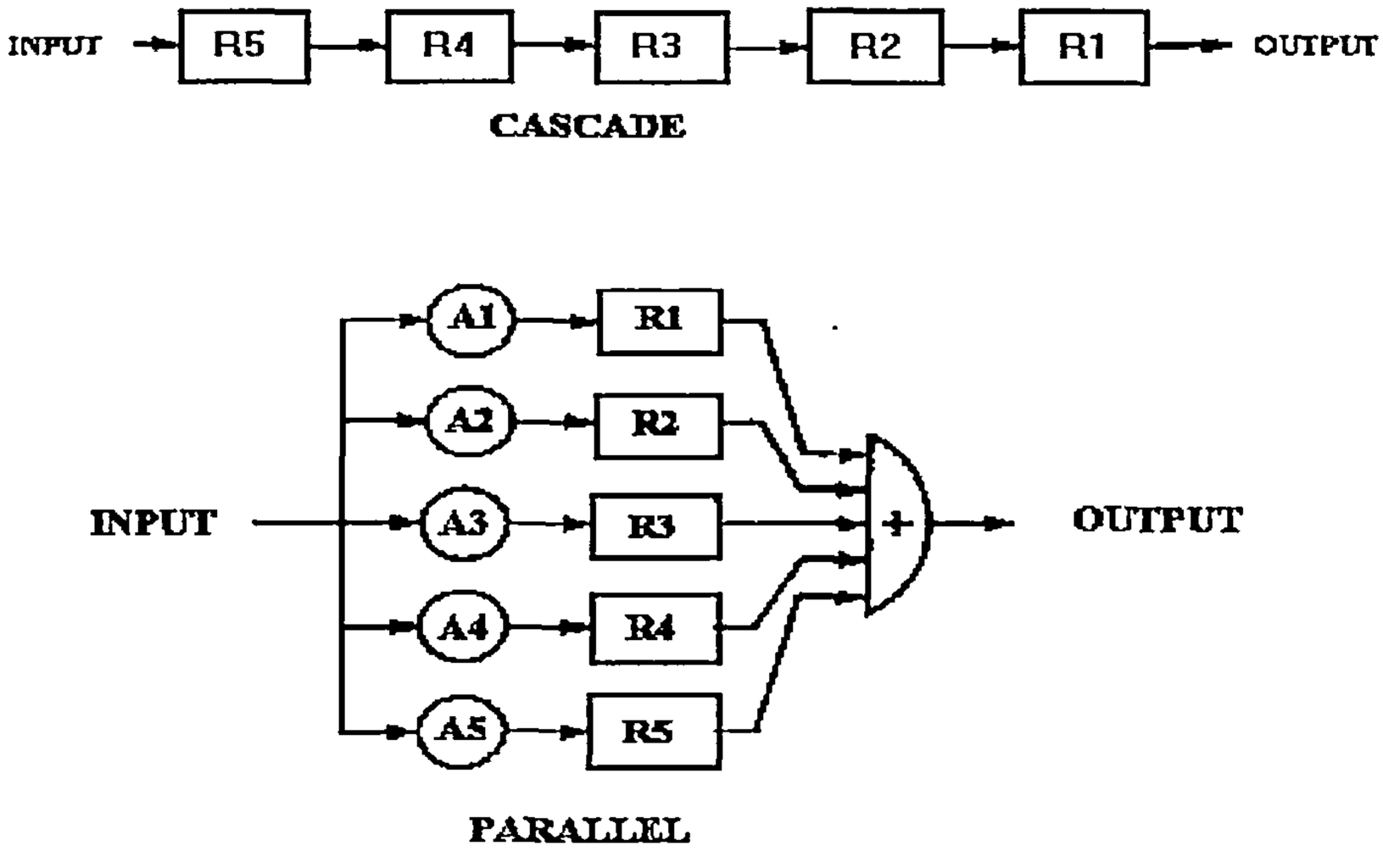


Figure 7.1. Digital resonators can be configured in cascade with the output of one acting as input to the next (top), or in parallel, in which case each receives the same source input, whose gain is determined by an independent amplitude control, and outputs are summed algebraically (bottom) (from Klatt, 1988).

The cascade (series) formant synthesizer does attempt to model the modifications to the voice source signal as it moves through the vocal tract in order to derive the acoustic waveform by adding "the effect of each higher resonance to the final output, and thus produces a direct replica of the total formant energy distribution, which corresponds quite well to the natural resonance mode of the vocal tract. This approach constitutes a fairly faithful imitation of vocal tract behaviour, and as a result, serial synthesizers are particularly good for synthesizing vowel sounds." (Styger & Keller, 1994). It is simple, having only frequency control and bandwidth control of the resonators, the amplitude being implicitly controlled by the all-pole transfer-function of the vocal tract with no nasal coupling. It models appropriately open non-nasal vowel sounds.

7.2.2.1 Vocal Tract Modelling for Formant Synthesizers

Both cascade and parallel designs have advantages and drawbacks. An all-cascade formant synthesizer is better designed to produce non-nasalized vowels than a parallel formant synthesizer, but this is counteracted by the need to use extra control information for consonant production or nasalization, and the difficulty in ensuring correct formant amplitudes (Holmes, 1983). In the case of the parallel formant synthesizer, it has the advantage of being able to effectively model all speech sounds, but requires a more complex implementation in its design for synthesizing vowels. The KLSYN88 (Klatt & Klatt, 1990) is a software based cascade/parallel synthesizer which draws upon the best features from both designs, and is the principle craft for this research.

7.2.3 The KLSYN88 Synthesizer

The information extracted from the analysis procedures outlined in the previous chapter is used as control data to drive a voice synthesizer. Synthesis has been implemented on the KLSYN88 synthesizer (Klatt & Klatt, 1990) modified for PC.

The Klatt synthesizer "KLSYN88" is a greatly improved version of the original KLSYN80 synthesizer (Klatt, 1980). The KLSYN88 has better modelling of female and children speakers with improved voice source and interactions between source and vocal tract (Klatt & Klatt, 1990).

The KLSYN88 models the vocal tract by having two sound sources with the same resonator specifications; one models the larynx and drives the cascade branch, and the other models the points of constriction within the supralaryngeal vocal tract and essentially drives the parallel branch. They can operate separately or together as in real speech. Each resonator is duplicated in both branches so resonance continuity during transitions between consonants and vowels can be maintained. The cascade branch only results in good non-nasalized vowels and aspirants. The parallel branch is used mainly to produce fricatives. Together, nasals and voiced sounds can be produced.

The KLSYN88 is controlled by 60 parameters. These are listed in figure 7.2 which represents the ASCII output file produced by the synthesizer for the default setting. The first 12 control parameters are constants. The other 48 control parameters are time variable, usually in 10 ms frames. Each control parameter is listed with its own symbolised name, a minimum value, a default value, a maximum value, and a description of its function.

7.2.3.1 Voicing Source Models

The constant parameter SS, "source switch" can select one of three voicing source waveforms; an impulse source model, an LF voicing model, and a KLGLOT88 voicing model which is the default model and is used in this research. This is described below: All three sources are controlled by three basic parameters:

1. fundamental frequency (F0). Parameter F0 specifies, in tenths of Hz for increased accuracy, the rate at which the vocal folds vibrate;

2. amplitude of voicing (AV). AV simulates the amplitude of the voicing source which is specified in dB;

3. spectral tilt (TL). TL controls a low-pass filter used to spectrally tilt the voice source. It offers an extra spectral attenuation at 3 kHz. The outputs from all the sources pass through this filter.

	SYM	MIN	VAL	MAX	DESCRIPTION
1.	DU	30	500	5000	Duration of the utterance, in msec
2.	UI	1	5	20	Update interval for parameter reset, in msec
3.	SR	5000	10000	20000	Output sampling rate, in samples/sec
4.	NF	1	5	6	Number of formants in cascade branch
5.	SS	1	2	3	Source switch (1=impulse, 2=natural, 3=LF model)
6.	RS	1	8	8191	Random seed (initial value of random number generator)
7.	SB	0	1	1	Same noise burst, reset RS if AF=0 and AH=0 (0=no,1=yes)
8.	CP	0	0	1	0 implies Cascade, 1 implies parallel tract excitation by AV
9.	OS	0	0	20	Output selector (0=normal,1=voicing source,...)
10.	GV	0	60	80	Overall gain scale factor for AV, in dB
11.	GH	0	60	80	Overall gain scale factor for AH, in dB
12.	GF	0	60	80	Overall gain scale factor for AF, in dB
13.	F0	0	1000	5000	Fundamental frequency, in tenths of a Hz
14.	AV	0	60	80	Amplitude of voicing, in dB
15.	OQ	10	50	99	Open quotient (voicing open-time/period), in %
16.	SQ	100	200	500	Speed quotient (rise/fall time of open period. LF model), in %
17.	TL	0	0	41	Extra tilt of voicing spectrum, dB down @ 3 kHz
18.	FL	0	0	100	Flutter (random fluct in f0), in % of maximum
19.	DI	0	0	100	Diplophonia (pairs of periods migrate together), in % of max
20.	AH	0	0	80	Amplitude of aspiration, in dB
21.	AF	0	0	80	Amplitude of frication, in dB
22.	F1	180	500	1300	Frequency of the 1st formant, in Hz
23.	B1	30	60	1000	Bandwidth of the 1st formant, in Hz
24.	DF1	0	0	100	Change in F1 during open portion of a period, in Hz
25.	DB1	0	0	400	Change in B1 during open portion of a period, in Hz
26.	F2	550	1500	3000	Frequency of the 2nd formant. in Hz
27.	B2	40	90	1000	Bandwidth of the 2nd formant, in Hz
28.	F3	1200	2500	4800	Frequency of the 3rd formant. in Hz
29.	B3	60	150	1000	Bandwidth of the 3rd formant, in Hz
30.	F4	2400	3250	4990	Frequency of the 4th formant. in Hz
31.	B4	100	200	1000	Bandwidth of the 4th formant. in Hz
32	F5	3000	3700	4990	Frequency of the 5th formant. in Hz
33.	B5	100	200	1500	Bandwidth of the 5th formant. in Hz
34.	F6	3000	4990	4990	Frequency of the 6th formant. in Hz (frication or if NF=6)
35.	B6	100	500	4000	Bandwidth of the 6th formant in Hz (only applies if NF=6)
36.	FNP	180	280	500	Frequency of the nasal pole, in Hz
37.	BNP	40	90	1000	Bandwidth of the nasal pole. in Hz
38.	FNZ	180	280	800	Frequency of the nasal zero. in Hz
39.	BNZ	40	90	1000	Bandwidth of the nasal zero. in Hz
40.	FTP	300	2150	3000	Frequency of the tracheal pole. in Hz
41.	BTP	40	180	1000	Bandwidth of the tracheal pole. in Hz
42.	FTZ	300	2150	3000	Frequency of the tracheal zero, in Hz
43.	BTZ	40	180	2000	Bandwidth of the tracheal zero. in Hz
44.	A2F	0	0	80	Amplitude of frication-excited parallel 2nd formant, in dB
45.	A3F	0	0	80	Amplitude of frication-excited parallel 3rd formant, in dB
46.	A4F	0	0	80	Amplitude of frication-excited parallel 4th formant, in dB
47.	A5F	0	0	80	Amplitude of frication-excited parallel 5th formant, in dB
48.	A6F	0	0	80	Amplitude of frication-excited parallel 6th formant, in dB
49.	AB	0	0	80	Amplitude of frication-excited parallel bypass path, in dB
50.	B2F	40	250	1000	Bandwidth of frication-excited parallel 2nd formant, in Hz
51.	B3F	60	320	1000	Bandwidth of frication-excited parallel 3rd formant, in Hz
52.	B4F	100	350	1000	Bandwidth of frication-excited parallel 4th formant, in Hz
53.	B5F	100	500	1500	Bandwidth of frication-excited parallel 5th formant, in Hz
54.	B6F	100	1500	4000	Bandwidth of frication-excited parallel 6th formant, in Hz
55.	ANV	0	0	80	Amplitude of voicing-excited parallel nasal formant, in dB
56.	A1V	0	60	80	Amplitude of voicing-excited parallel 1st formant, in dB
57.	A2V	0	60	80	Amplitude of voicing-excited parallel 2nd formant, in dB
58.	A3V	0	60	80	Amplitude of voicing-excited parallel 3rd formant, in dB
59.	A4V	0	60	80	Amplitude of voicing-excited parallel 4th formant, in dB
60.	ATV	0	0	80	Amplitude of voicing-excited parallel tracheal formant, in dB

Figure 7.2. The parameter listing for the KLSYN88 Synthesizer (Klatt, 1988).

The voicing source 2 model is the default source KLGLOTT88, which approximates a natural waveform. It is the voicing source used in the present synthesis. The characteristics of this source are controlled by a further 4 parameters in addition to F0, AV, and TL:

1. varying the "open quotient", OQ variable changes the relative amplitude of F0 and hence the spectral tilt of the waveform. It simulates the acoustic effects of varying the degree of adduction of the vocal folds prior to the onset of phonation. OQ is defined as the percentage open period of vocal fold vibration at a specified fundamental frequency.

The vocal fold behaviour that is being modelled (i.e. the degree of glottal opening) directly determines the spectral characteristics of the voicing source (Klatt & Klatt, 1990). In synthesis, a breathy phonation would require a high OQ (high relative F0 amplitude with steep spectral tilt) whilst a more pressed phonation should have a low OQ (lower relative F0 amplitude with shallow spectral slope). OQ works in conjunction with the parameter AH.

2. AH controls the amplitude of aspiration. Aspiration is the noise resulting from a constriction at the level of the vocal folds when they are close but not in contact. Aspiration is an important component of breathy phonation.

3. the parameter FL, "flutter" simulates the slow drift of the fundamental frequency found in natural speech. It adds a quasi-random element to each F0 value which is the sum of three slowly changing sine-waves, as opposed to the random component found in jitter.

4. the parameter DI, "diplophonia" increases/decreases alternate F0 pulses. A delayed pulse is attenuated in amplitude.

The excitation pulse spectrum can be varied in a more realistic way due to the inclusion of additional controls. Both the flutter and diplophonia parameters introduce irregularities into the fundamental period cycle contributing to more natural voicing (Klatt, 1988).

The 3rd voicing source is based on a modified Liljencrants-Fant (LF) model (Fant, 1986). It is similar to the natural waveform described above, and is controlled by the same variables; F0, OQ, AV, and TL. However, it uses an additional variable, SQ, "speed quotient" which is the ratio of the duration of glottal opening to closing. The glottal pulse shape is significantly changed at voicing onset and offset, and at vowel-consonant boundaries due to changes in rate of glottal opening and closing (Gobl, 1988). The default source has been used in this synthesis to synthesize the steady-state portion of vowels.

7.2.3.2 Vocal Tract Models

As mentioned above, the synthesizer can be configured to model the resonances of the vocal tract which arise from either a laryngeal source waveform, using the cascade branch, or from frication noise, that is, constriction above the larynx using the parallel branch. The parallel branch can also be used with a laryngeal source waveform for synthesis of some rare pathological types.

Cascade Vocal Tract

The Cascade vocal tract model can produce vowels, liquids and glides using a series of six controllable resonators. Each resonator is controlled by a centre frequency variable and a bandwidth. Varying these two parameters changes the relative amplitudes of the formants and their frequency position for non-nasal sounds. These resonators in cascade approximate the vocal tract transfer function for a decoupled vocal tract, which behaves as an all-pole filter, enhancing those partials corresponding to the transfer function of that particular vocal tract configuration. Consequently, the vocal tract for nonnasalized vowels can be modelled with a set of poles. Spectral peaks in the transfer function correspond to the position of the poles.

However, if the nasal cavities are coupled to the vocal tract, extra damping occurs due to the anti-resonating properties of the nasal cavities. This introduces zeros or troughs into the transfer function, corresponding to dips in the frequency spectrum. Nasalized sounds therefore contain both poles and zeros in their transfer functions. Zeros also occur in fricatives and stops (Styger & Keller, 1994), and when tracheal coupling is present (Klatt & Klatt, 1990).

In the Klatt synthesizer, to account for this, the cascade model has an additional two pairs of resonators to the six controllable ones. The nasal resonator and anti-nasal resonator form the nasal pole-zero pair and are activated to simulate nasals and nasalization. An additional tracheal pole-zero pair can also be used to mimic tracheal coupling which can occur in breathy vowels when the glottal opening is sufficiently large. For non-nasal sounds, the anti-nasal resonator is the exact mirror image of the nasal resonator, and they cancel each other out. If the tracheal pole-zero pair also cancel eachother out, the vocal tract model can be said to have an all-pole transfer function.

The source-tract filter theory assumes that the voice source and the vocal tract are independent of eachother. However, under a number of conditions there exists a non-linear coupling between the voice source and the first few formants of the vocal tract modes (Fant, 1986; Stevens & Bickley, 1986). This is due to the time-varying impedance of the vocal folds with glottal opening which interact with the vocal tract impedances, and various irregular perturbations and constrictions of the vocal folds caused by vocal tract standing-wave pressure changes. The increased glottal impedance when the glottis is open introduces low-frequency zeros into the vocal tract transfer function due to tracheal coupling. Glottal opening also has the effect of increasing the first formant frequency and changing its bandwidth. These are modelled in the KLSYN88 synthesizer.

Parallel Vocal Tract

This branch has 5 resonators and, for bilabials and labiodentals, there is a by-pass path since there are no formants above the point of constriction. A block diagram of the KLSYN88 synthesizer is shown in figure 7.3. A vibrato function has been added to the synthesizer. All synthesis is programmed using an algorithm called "kspandoc" which interpolates between two points set at specific times in either a linear or a logarithmic manner. An example program is shown in Appendix A.

7.3 Perceptual Tests

This section provides a perceptual evaluation of the proposed models described in the previous section on synthesis. It begins with a short introduction to the nature of hearing perception followed by the perceptual experiments.

7.3.1 Introduction

It is possible to predict performance measures from absolute judgement experiment using statistics. It is labelled the "measure of information transmission" (Garner, 1962).

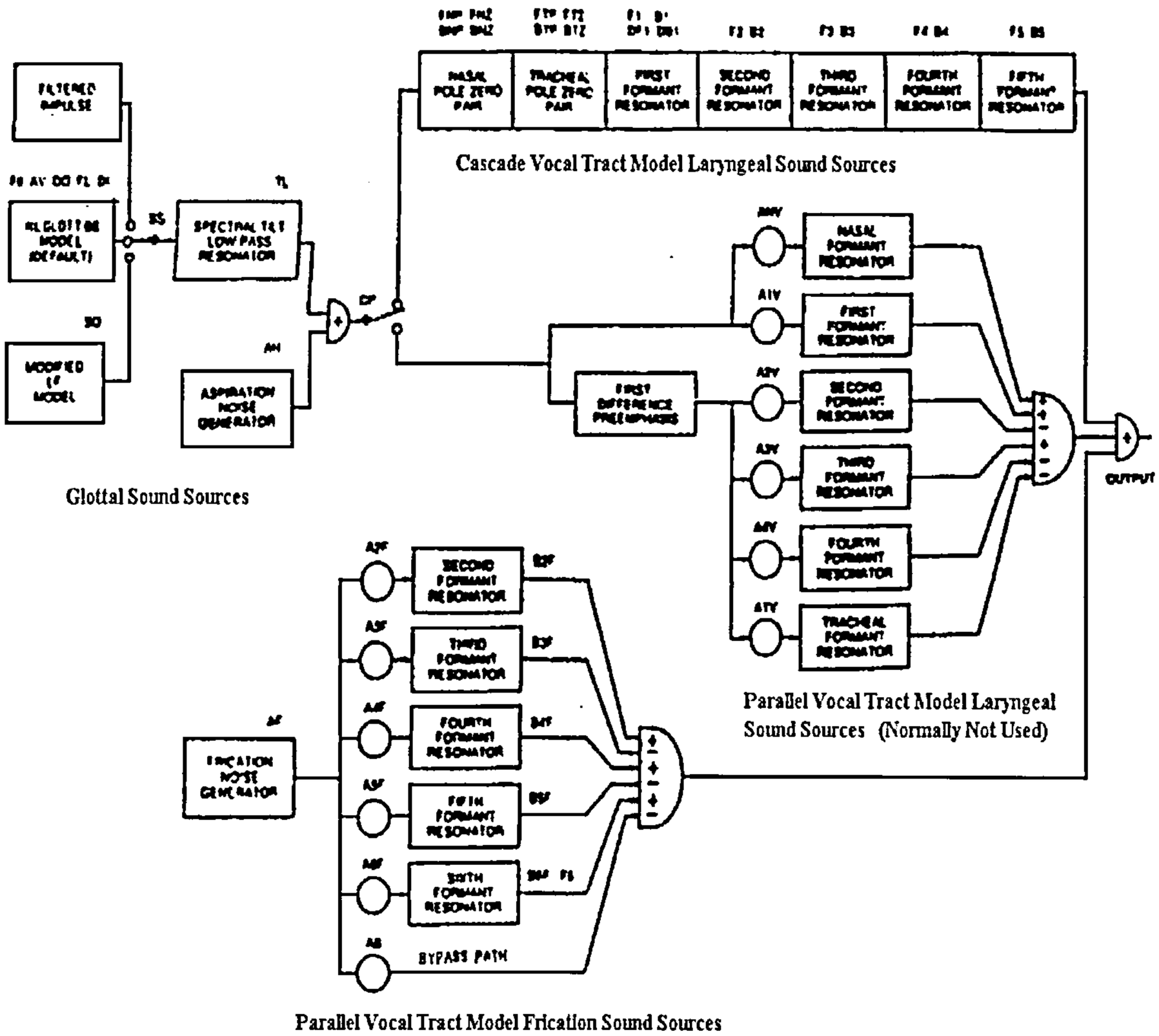


Figure 7.3. Block diagram of the new KLSYN88 formant synthesizer. Three voicing source models are available. Also added are a tracheal pole-zero pair, and control parameters allowing the first formant frequency and bandwidth to vary over a fundamental period (from Klatt, 1988).

Experimental evidence points to the difficulty of identifying stimuli varying along one continuum or attribute. Individuals can only identify up to about 9 different stimuli along the same attribute, depending on that attribute. Garner (1962) has shown that there is little improvement in identification

performance even when the stimulus range is increased greatly. The identification measures are poor when compared to experiments with stimuli varying independently on multi-dimensions. It has been shown that extra dimensions aid in identification as the stimuli vary independently on each dimension, thus increasing perceptual information. Listeners can identify extremely small differences in detail between two stimuli, but are insensitive to stimuli that differ along one attribute. For mulit-dimensional experiments, the larger the physical difference between similar stimuli, the less confusion there is likely to be.

7.3.1.1 Categorical Perception

Complex tones such as speech and music are discriminated differently to pure tones. It is the physical difference which forms the basis to pure tone discrimination, whereas functional labelling of complex tones, that is, the meaning, also aids discrimination between speech events or musical events. The physical difference between tones is related to acoustic perception, and functional labelling is related to categorical perception.

The intensity and frequency of a pure tone can be heard to change gradually. However, for complex tones, the change is non-linear and complicated by the fact that acoustical events such as formant transitions and frequency ratios and are not normally heard. They are either drawn into more "meaningful" segments or are ignored if they are insignificant (Handel, 1989):

> "For categorical perceiving, the event is heard directly; the acoustic properties of the sound are recovered from memory. For auditory perceiving, the acoustic properties are heard directly; the perceptual events are deduced" (Handel, 1989).

Categorical perception is a complex dynamic process which is task dependent and dependent on the listener's performance of memory and judgement. It can be learned. It is possible that categorical perception arises partly out of the inclination of the auditory system to perceive equally continuous changes as discontinuous. For these changes to be significant, they must combine with other acoustic variations resulting from specific production changes.

There is a general psychological theory which relates physical properties to psychological ones. Diehl (1987) describes it as "mutual enhancement". In the auditory sphere, for meaning to be conveyed, speech sounds must be easily discriminated. This is increased by maximising and combining the acoustic contrasts resulting from physical production changes (articulation) with the qualitative perceptual changes related to the hearing system mechanisms. Categorical perception is multi-dimensional and is explained as a combination of both production changes and perceptual discontinuities.

> "All perceiving comes form acoustic patterning, and all perceiving can yield both the categorisation and the auditory detail. Acoustic information is usually ambiguous and supports many possibilities. The best perceptual strategy would be to retain as much acoustic information as possible for as long a time as possible to allow surrounding information to influence the percept. Without the ability to switch among perceptual levels, our perceptual capabilities would be static, and we would be unable to tune to the properties of the stimulation" (Handel, 1989).

7.3.1.2 The Effect of Context

The perceptual meaning of an acoustic segment depends on its context, which requires integration of multiple perceptual information.

Repp (1982) has shown that one acoustic cue can be compensated for by a change in another cue in order to maintain the original perception. The cues can be totally dissimilar, such as the relationship equating a temporal cue (voice onset time) with a spectral cue (first formant onset frequency) (Handel, 1989). This is termed a trading relationship and takes place at specific points in time. It appears that trading relationships are general perceptual phenomena.

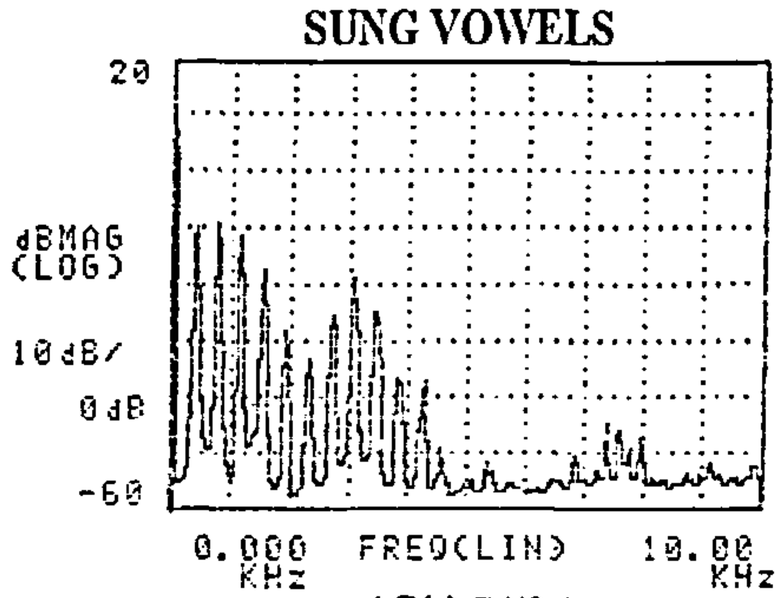
The same speech segment can also give rise to both speech and non-speech perception. Experiments with simulated speech segments where the formants are replaced with sinusoids having the same frequency and amplitude changes have shown that phonetic perception is still possible with a reduction in structure, but it is significantly weakened with an increase in ambiguity due to the inclusion of non-speech elements. This dual perception can be seen in other musical phenomena. A violin tone played badly may result in a set of partials being too loud. This may lead to the perception of the proper timbre plus a separate sound. In summary:

> "The context may decrease the resolution due to interference from other parts of the speech signal, or the context may enhance the resolution, because of comparison with "reference" or unchanging parts of the signal" (Handel, 1989).

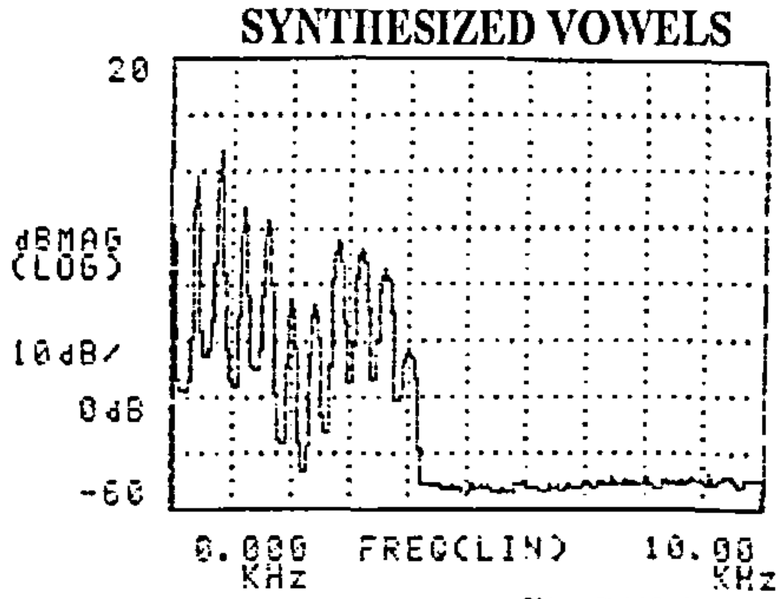
7.3.2 Perceptual Test

Using the Klatt Synthesizer, four sung vowels were synthesized based on results derived from the analysis of the opera and belting sets. The vowel-pitch tokens chosen were G4/3:/ and E5/a:/ in both opera and belting quality. Figure 7.4 provides a spectral comparison of the real sung data (from which the synthesized tones were based on) and the synthesized versions. The synthesized tones used average CQ and average F0 measurements derived from the real data. The synthesized spectrums were based on the spectrums of the real data and visually modified. However, as mentioned above, many more higher formants were needed to model the spectral components above 3 kHz than would be present in the real vocal tract. The introduction of vibrato into the tone, the vibrato rate and the vibrato amplitude was averaged between the opera and belting tokens, and incorporated into both qualities. This was done so that the differences in vibrato would not have served as a major cue for differentiation of the synthesis qualities. It was more important to concentrate on discriminating the spectral qualities (lack of vibrato, as can be found in belting tones, would have been too obvious a cue). Another reason to include vibrato on belting tones, which can be sung straight, is that without some vibrato, it is difficult to achieve "naturalness" on sustained synthesized vowels.

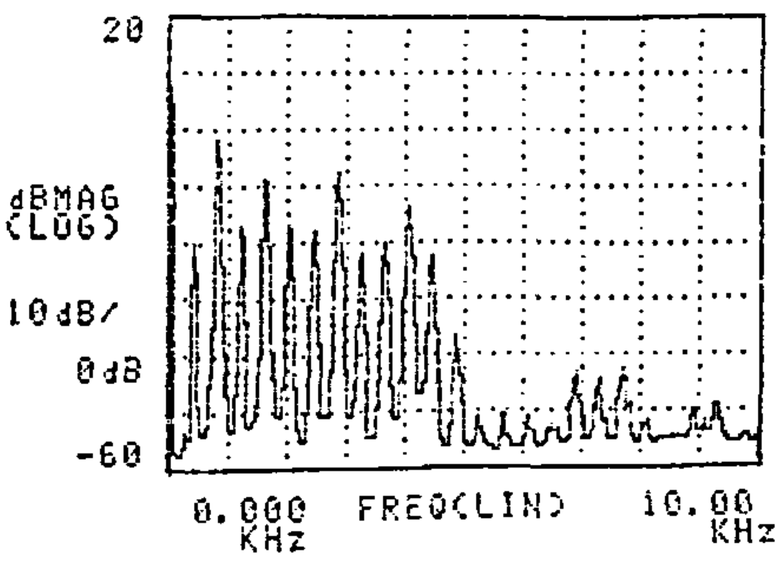
In all, 8 tones (4 real and 4 synthesized) tones were used for the perceptual test. These were played randomly 10 times each. A small passage of real opera and belting was played twice prior to the tests and once through halfway through the experiment (after 40 vowels). Ten York University students



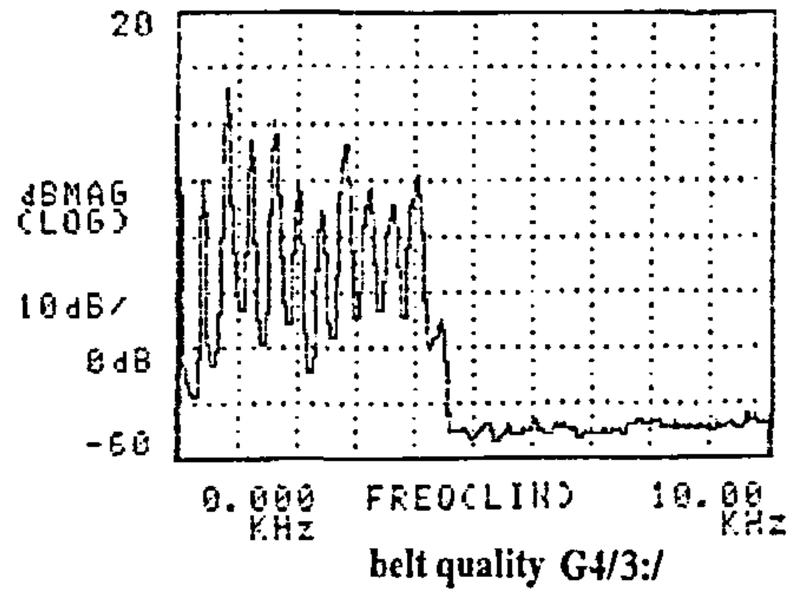
AG(o) G4/3:/



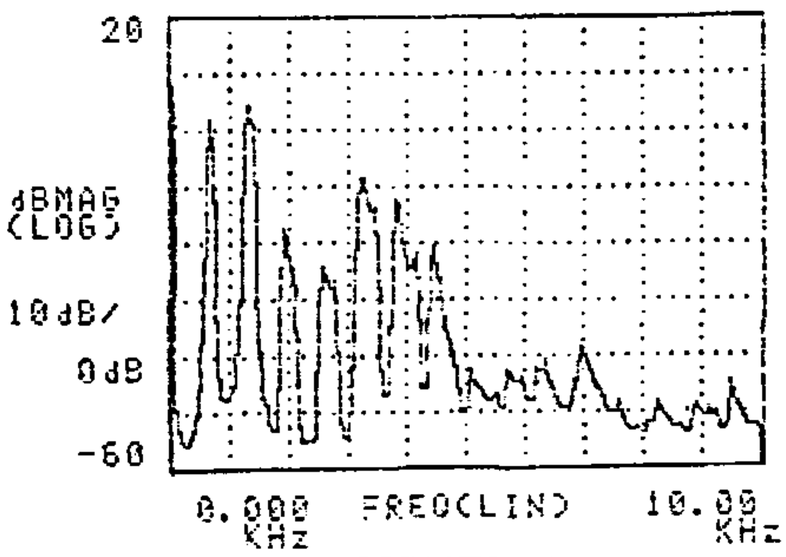
opera quality G4/3:/
synthesized



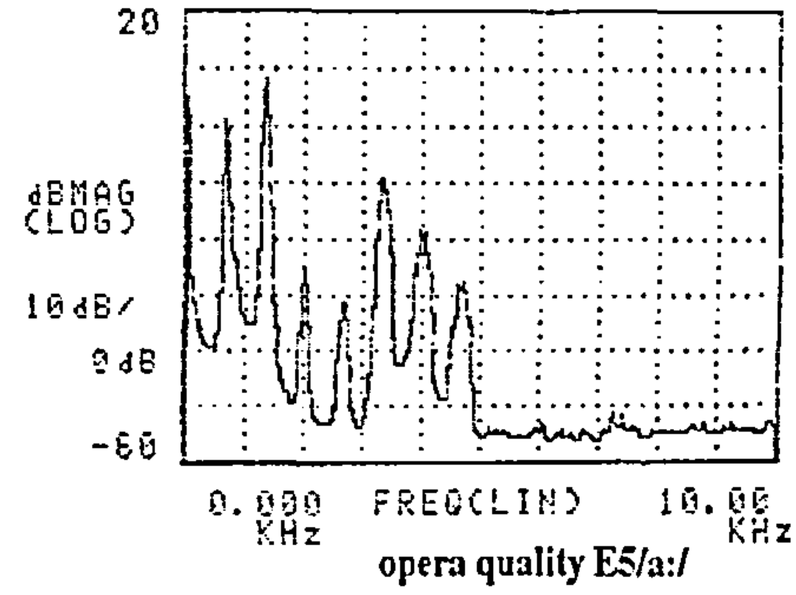
VP(b) G4/3:/



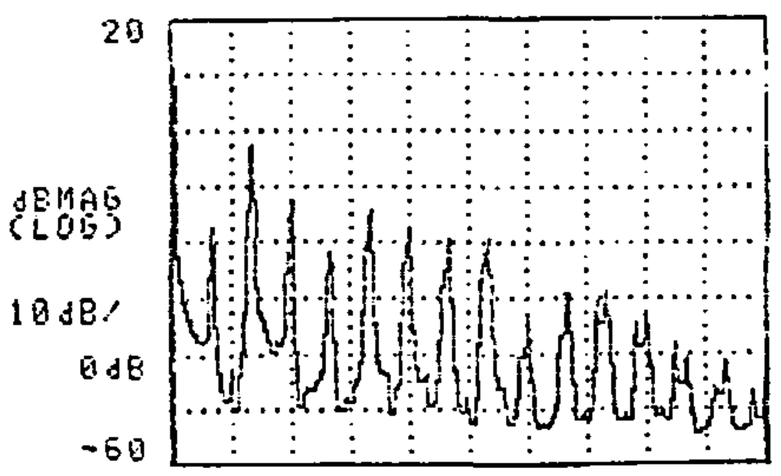
belt quality G4/3:/
synthesized



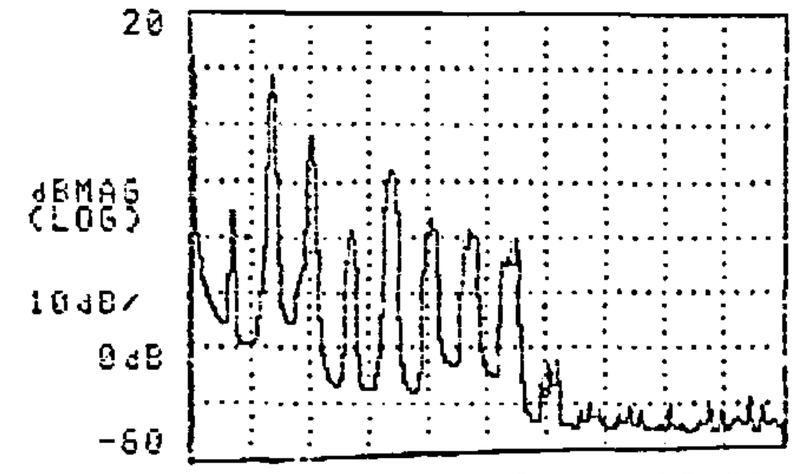
MC(o)E5/a:/



opera quality E5/a:/
synthesized



MC(b)E5/a:/



belt quality E5/a:/
synthesized

Figure 7.4. Average spectra of the synthesized sung vowels (right column) derived from the real sung vowels (left column). All the vowel sounds above were used in the perceptual test.

with varying musical experiences served as judges. The tests were undertaken in a music technology laboratory and played in digital stereo over headphones. The judges were asked to identify whether each tone was produced in belting or opera quality.

The modelling of the vocal tract as a tube open at one end and closed at the other produces a transfer function with an infinite number of poles. In reality, in the KLSYN88 synthesizer only the resonances below 5 kHz can be controlled with accuracy. The resonators above 5 kHz are fixed with respect to centre frequency and bandwidth. Little significance is attached to these higher formants because little voiced energy above 5 kHz is present in the speech signal, due to the low-pass characteristics of the source signal spectrum. and also the hearing system is less sensitive above 5 kHz, so it is unneccessary to have precise control of higher formants above this.

Holmes (1983) gives a number of advantages of the parallel synthesizer over the cascade synthesizer. The cascade model cannot easily reproduce vocal effort changes or changes in the output spectrum due to glottal pulse shape, this would require an more control data since relative formant amplitudes cannot be explicitly set, and glottal pulse shape is modelled as constant. The parallel model has individual control over formant amplitudes so is better suited to modelling these changes. The cascade model only mimics the vocal tract response accurately up to 3 kHz, after which it breaks down leading to large errors in modelling (Holmes, 1983).

Appendix [A] gives a full example of the synthesis algorithm for the opera tone on G4/3:/, labelled here as G4opera.spk. Appendices [B], [C], [D] show the parameters which were varied from the default values in order to produce the other three synthesized tones, the files being called G4belt.spk, E5opera.spk, and E5belt.spk. The programs are all used with an algorithm called "kspandoc" which interpolates between two given values. As can be seen from the input data and the results, the synthesizer does not provide good spectral control over 5 kHz. In order to achieve the correct spectral amplitude over 3 kHz, it was necessary to use synthesizer formants very close together, and those which do not relate to the formant locations of the real tones. A lot of time was spent trying to copy the spectral content. A number of alterations had to be made in order to smooth over perceived vibrato stepping at the high pitches due to the quantisation errors. This was achieved by placing vibrato on the formant frequency locations and bandwidths. This would not necessarily occur in the real voice, though it does point to how important smoothness of vibrato is as a naturalness cue. The open quotient (OQ) values were determined from the real data (the inverse of the CQ values). The only parameters which were varied between synthesis tones were number of formants, formant frequency and bandwidth and OQ.

7.3.3 Results

The results for the tests are shown in figure 7.5. The judges have been grouped simply into three categories depending on musical experience.

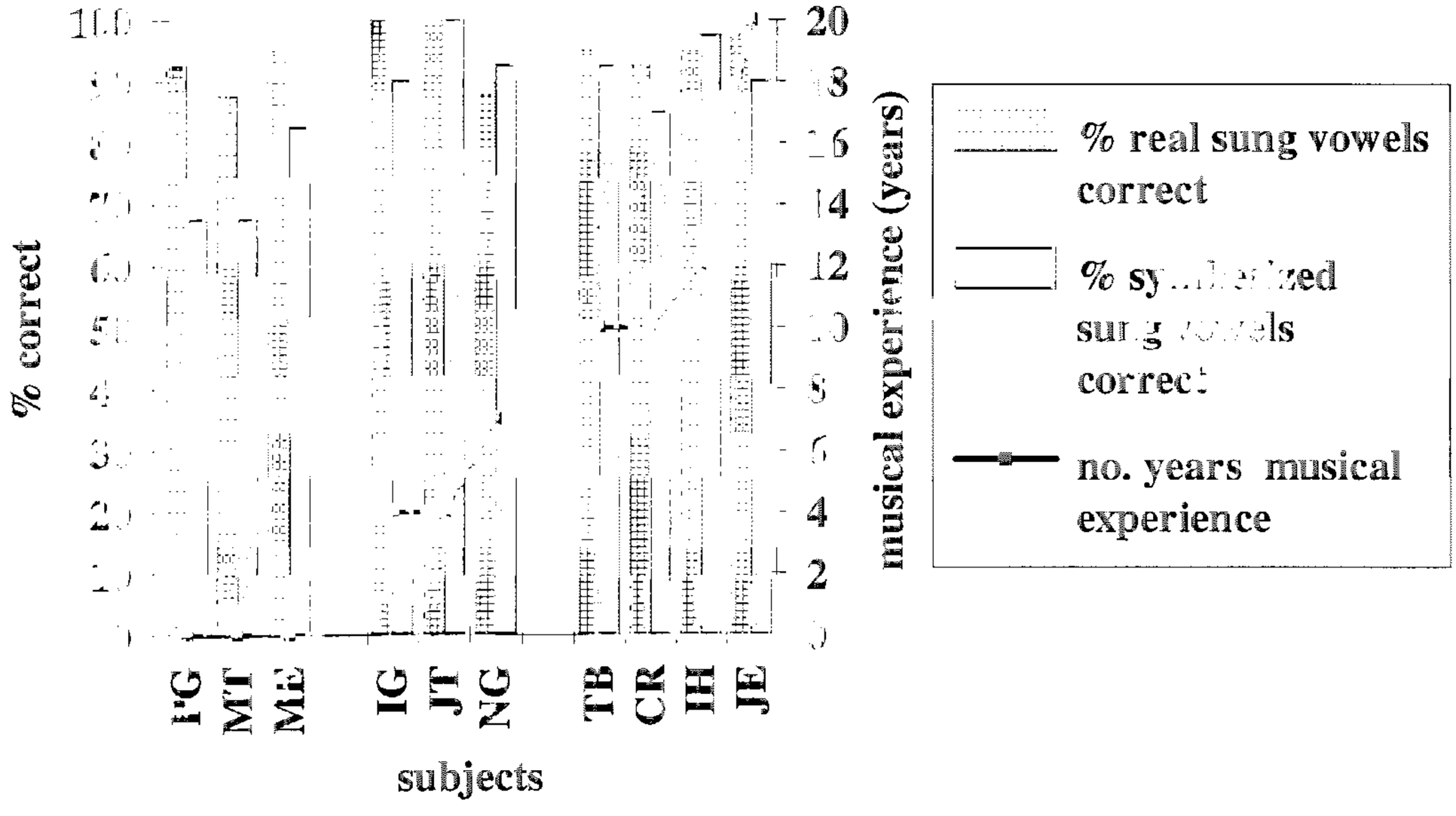


Figure 7.5. Perceptual test results arranged according to the musical experience of each judge.

It can be seen that correct identification of the real data is high for all judges. The performance for the synthesized tones, as expected, is lower than for the real tones. It is interesting to note that the group with no musical experience has the lowest identification performance for the synthesized tones.

Figure 7.6 below shows the average percentage of incorrect judgements. It can be seen that many more synthesized tones were incorrectly judged than sung tones, and overall, there was a higher instance of error in correctly labelling the belting tones than the opera tones.

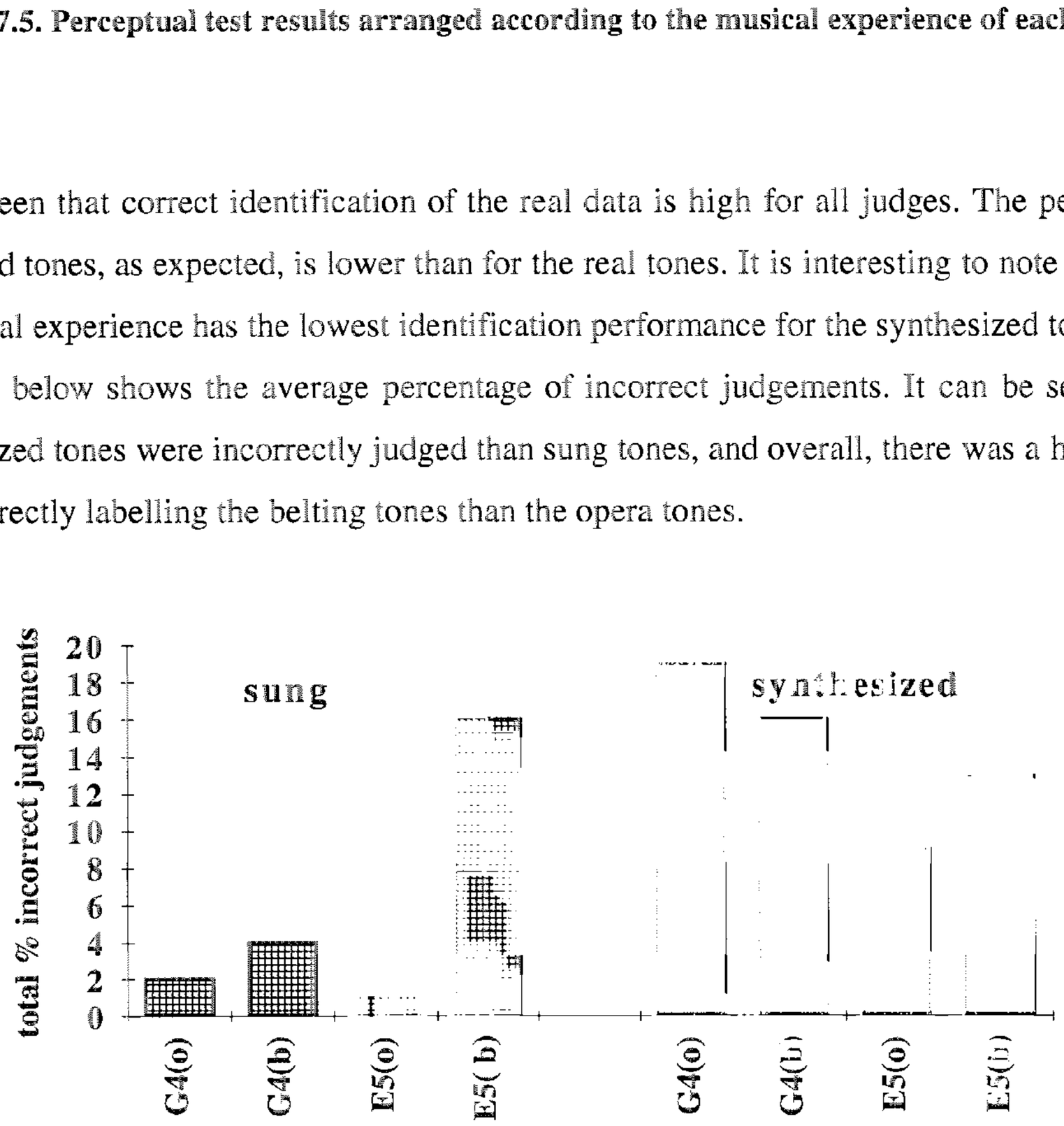


Figure 7.6. Average incorrect judgements (%) for both sung and synthesized tones.

7.4 Discussion and Conclusions

Figures 7.5 and 7.6 show that most judges managed to correctly label the real sung tones, yet the instance of error in the labelling of the synthesized tones was greater in the judges with no musical experience. One may speculate that this is due to auditory familiarity. Most people, regardless of musical background tend to be familiar with the Western female singing voice, hence the high instance of correct judgements across the board. It may be that those judges with less musical background have more difficulty in differentiating between the slightly reduced auditory information associated with the higher pitches in female singing. This is proabably more significant when the tones are synthesized, partly due to the increased unnaturalness of the synthesized tone (the synthesized tone is a repeated waveform pattern with no fluctuations) and also due to the great reduction in the amount of vibrato on the synthesized opera tones. One would suggest that vibrato is an important auditory cue in the identification of differing vocal qualities.

The cascade model of the KLSYN88 Synthesizer cannot account appropriately for vibrato and high energy content in the region of the singers formant and above when synthesizing the female singing voice, hence vibrato was reduced in the synthesiaed tones. However, the results from the perceptual tests have shown that the parameters used to differentiate between belting and opera qualities in the female voice seem to be sufficiently important in the identification of these qualities. These parameters include CQ and spectral envelope.

With assessing the number of formants, formant frequency and bandwidth, it was impossible to exactly relate the values of these synthesis parameters to the real sung tones since the KLSYN88 synthesizer was not designed for singing work and hence, had to be forced in to a mode of operation unrelated to the vocal tract. However, it is apparent that it is possible to cause a fair distinction between vocal qualities in female singing by using very few parameters. The problem with cascade synthesizers is that the real human voice, especially the female singing voice is that there is a vast amount of interaction between various muscles, and structures within the human vocal tract, which cannot yet be defined and accounted for in the synthesizer.

One can speculate that a lot of the perceptual distinction between belting and opera is due to two spectral features: firstly, whether or not there is lowest partial dominance, as is usually the case in opera; and secondly, the amount of energy content of the partials above the 6th or 7th, in the area of 3-5 kHz, which is high and contributes to the rough and cutting sound of belting. It was possible to achieve this "edge" sound using the KLSYN88 synthesizer, even though it was time costly and the number, position, and location of formants used to recreate the spectral envelope shapes did not appear to tally with the real sung tones. Much fine manual adjustments to the synthesized parameters had to be done which were unrelated to the real tones. In one sense, the synthesizer did achieve a purpose in being able to recreate the acoustic differences in female singing vocal quality, though it cannot recreate the differences in production.

Chapter 8

Conclusions and Future Research

8.1 Conclusions

This thesis set out to assess the appropriateness of standard two-channel speech technology for the differentiation of two vocal qualities exhibited by female singers.

The speech analysis techniques used go some way towards differentiating opera and belting qualities in terms of assessing single words or vowels. However, singing requires extended phonations of up to several seconds at a time, and so it would have been useful to look at singing production and vocal quality over these long passages to see how a tone may develop over these phrases which may lead to insights into interpretation. The speech analysis system, SFS, used in this study was unsuited for this purpose since it could only present 2 seconds worth of data at a time, it was very slow, the manipulations were not universal for all items, the spectrographic analysis tools and LPC analysis tools were unsuited for singing analysis (and hence were discounted for this work), and it could only deal with two channels worth of input data. Larynx height analysis (which has been added to the speech analysis techniques) appears to be a good indicator of voice technique (not necessarily good voice technique) and a good addition to the standard two-channel analysis techniques, although a multi-channel recorder is needed plus a multi-channel analysis system. The speech synthesizer is unsuited to proper female singing synthesis, but it does recreate the necessary perceptual changes based on analysis of opera and belting qualities.

To summarize: In order to do this type of work justice, it is proposed that multi-channel analysis and synthesis equipment should be used which can deal with both the demands placed on the equipment of the additional features and extended vocal ranges found in the singing voice plus the extended durational demands of vocal music.

8.2 Future Research

There are a number of areas which need further investigation.

Mixed quality need further investigation since it has been shown that not all the belters were belting when asked to, but rather, sang in mixed quality, which is more commonly used in the West End musicals.

Comparisons in the ease of articulation of different vowels at different pitches in both qualities should be considered. This may provide clues as to the nature of vowel modification. This has been scientifically investigated for opera quality, but not for belting quality. For example, do different vocal tract and larynx settings provide different problems for articulation, and what are the acoustical consequences? Positioning of the lips greatly influences formant positions. How do lip settings differ in belting and opera, and what are the acoustical consequences? Is lip setting an important factor in producing different vocal qualities?

Voice source-vocal tract interaction should also be considered. This may have some affect on the quality of the lower part of the middle register in opera quality which has a large open quotient. The glottis remains open for a large period of the vibratory cycle which potentially couples the subglottal airways with the supraglottal airways. This may lead to extra damping of the voice source since some air pressure can escape back down into the lungs. This may account for the decrease in loudness in this pitch range for female opera singers. An airflow mask (Rothenberg, 1973) is required in order to assess the subglottal pressure in phonation.

Jitter is a random fluctuation in the length of a vibratory cycle providing minute inflections in the pitch of a tone (Klatt and Klatt, 1990). It is present in all spoken voices to some extent. Perceptual experiments have shown that if jitter is not incorporated into a synthesized spoken vowel, it is judged to be "unnatural" and mechanical sounding. However, too much jitter is indicative of a pathological voice, or an aged voice. It seems that jitter is a consequence of the muscular setting of the larynx when speaking. It will be interesting to investigate how jitter in opera quality compares to that in belting quality; how it varies with pitch; and whether it is tension dependent. Shimmer is a cycle-to-cycle variation in amplitude (Klatt & Klatt, 1990).

This could be useful as well as further assessing the real value of CQ in singing analysis. For example, some important areas which require investigation are how subglottal pressure influences CQ, how CQ changes across different levels of intensity, the relationship between CQ and registration, lx signal shapes and patterning for different voice qualities, and fundamental frequency range for different qualities. Interpretational aspects of singing in terms of vocal quality modification during different passages is also a useful area of investigation.

References

Abberton, E., Howard, D., & Fourcin, A. (1989). "Laryngographic Assessment of Normal Voice: A Tutorial", Clinical Linguistics and Phonetics, **3**, (3), 281-296.

Agren, K., & Sundberg, J. (1978). "An Acoustic Comparison of Alto and Tenor Voices", Journal Research in Singing (JRS), **1**, 26-32,

Ainsworth, W.A. (1976). "Mechanisms of Speech Recognition", Oxford: Pergamon Press.

Allen, J., Hunnicutt, M.S., & Klatt, D. (1987). "From Text to Speech: The MITalk System", Cambridge: Cambridge University Press.

Anderson, V.A. (1977). "Training the Speaking Voice", Oxford: Oxford University Press.

Atal, B.S., & Hanauer, S.L. (1971). "Speech Synthesis by Linear Prediction of the Speech Wave", Journal Acoustical Society America (JASA), **50**, 637-655.

Berg., J. van den (1968). "Mechanism of the Larynx and the Laryngeal Vibrations", In: Manual of Phonetics, B. Malmberg, (Ed.), London: North-Holland, 278-308.

Bismarck, G. von (1974). "Timbre of Steady Sounds: a Factorial Investigation of its Verbal Attributes", Acustica, **30**, 146-159.

Bjørklund, A. (1961). "Analysis of Soprano Voices", JASA, **33**, 575-582.

Blauert, J. (1983). "Spatial Hearing: the Psychophysics of Human Sound Localization", J.S. Allen (Tr.), Cambridge, Mass: MIT Press.

Bless, D.M., Biever, D., and Shaikj, A. (1986). "Comparisons of Vibratory Characteristics of Young Adult Males and Females", Proceedings of International Conference on Voice, Kurume, Japan, **2**, 46-54.

Bloothooft, G. (1985). "Spectrum and Timbre of Sung Vowels", PhD diss., Amsterdam: Vrije Universiteit te Amsterdam.

Boer, E. de (1980). "Auditory Physics. Physical Principles in Hearing Theory. I", Physics Reports, **62**, 87-174.

Borden, G.J., & Harris, K.S. (1984). "Speech Science Primer: Physiology, Acoustics, and Perception of Speech", J.P.Butler (Ed.), U.S.A: Waverly Press.

Bouhuys, A., Proctor, D.F., & Mead, J. (1966). "Kinetic aspects of singing", J. Appl. Physiol., **21**, 483-96.

Breen, A. (1990). "An Investigation into Synthesis-by-Analysis", PhD diss, UCL, Unpublished.

Campbell, M. & Greated, C. (1987). "The Musician's Guide to Acoustics", London: J.M.Dent & Sons.

Catellengo, M., Roubeau, B., & Valette, C. (1983). "Study of the Acoustical Phenomena Characteristic of the Transition Between Chest Voice and Falsetto", In: Proc. of Stockholm Music Acoustics Conference 1983 (SMAC 83), (1), A. Askenfelt, S. Felicetti, E. Jansson, & J. Sundberg (Eds.), Stockholm: Royal Swedish Acad. of Music, **46**, (1), 113-124.

Catford, J.C. (1964). "Phonation Types: The Classification of Some Laryngeal Components of Speech Production", In: Elements of General Phonetics, Abercrombie et al. (Eds.), Edinburgh: Edinburgh University Press, 26-37.

Catford, J.C. (1977). "Fundamental Problems in Phonetics", Edinburgh: Edinburgh University Press.

Cleveland, T.F. (1977). "Acoustic Properties of Voice Timbre Types and Their Influence on Voice Classification", JASA, **61**, 1622-29.

Cleveland, T., & Sundberg, J. (1983). "Acoustic Analysis of Three Male Voices of Different Quality", In: SMAC 83, (1), A. Askenfelt, S. Felicetti, E. Jansson, & J. Sundberg (Eds.), Stockholm: Royal Swedish Acad. of Music, **46**, (1), 143-56.

Coleman, R.O. (1976). "A Comparison of the Contributions of Two Voice Quality Characteristics to the Perception of Maleness and Femaleness in the Voice", Journal Speech and Hearing Research (JSHR), **20**, 197-204.

Cooper, F.S., Peterson, E. & Faringer, G.S. (1957). "Some Sources of Characteristic Vocoder Quality", JASA, **29**, 183(A).

Crowder, R. (1993). "Auditory memory", In: Thinking in Sound: the Cognitive Psychology of Human Audition, S. McAdams & E. Bigand (Eds.), Oxford: Clarendon Press.

Curtis, J. & Schultz, M. (1986). "Basic Laboratory Instrumentation for Speech and Hearing", Boston: Little, Brown.

Davies, P., Lindsey, G.A., Fuller, M., & Fourcin, A.J. (1986), "Variation in Glottal Open and Closed Phase for Speakers of English", Proceedings of the Institute of Acoustics (Proc. IOA), **8**, 539-546.

Doehring, D.G. (1974). "Pitch", In: Introductory Hearing Science: Physical and Psychological Concepts , S.E. Gerber (Ed.), Philadelphia: Saunders, 128-150.

Dowling, W.J. & Harwood, D.L. (1986). "Music Cognition", Orlando: Academic Press.

Edgington, M., Barnes, C., Stringer, P., & Howard, D. (1992). "The Speech Filing System: A Tool for Cooperative Speech Research", Proc. IOA, **14**, 79-86.

Erickson, R. (1977). "The Structure of Music: A Listener's Guide", Conneticut: Greenwood Press.

Estill, J., Baer T., Honda, K., & Harris, K. (1983). "The Control of Pitch and Quality, Part 1: An EMG study of supralaryngeal activity in six voice qualities", In: Transcripts, Twelfth Symposium: Care of the Professional Voice, The Juillard School, New York, 1983, V. Lawrence (Ed.), New York: The Voice Foundation, 86-91.

Estill, J., Baer T., Honda, K., & Harris, K. (1984). "The Control of Pitch and Quality, part 2: An EMG study of Infrahyoid Muscles", Transcripts, Thirteenth Symposium: Care of the Professional Voice, New York: The Juillard School, New York, 1984, New York: The Voice Foundation, 65-69.

Estill, J., Baer T., Harris, K., and Honda, K. (1983). "Supralaryngeal Activity in a study of six voice qualities", In: SMAC 83, A. Askenfelt, S. Felicetti, E. Jansson, & J. Sundberg (Eds.), Stockholm: Royal Swedish Acad. of Music, **46**, (1), 157-174.

Estill, J. (1988). "Belting and Classic Voice Quality: Some Physiological Differences", Medical Problems of Performing Artists, Philadelphia: Hanley and Belfus, March, 37-43.

Evans, E.F. (1975). "Cochlear Nerve and Cochlear Nucleus", In: Handbook of Sensory Physiology, W.D. Keidl & W.D. Neff (Eds.), Berlin: Springer, **2**, ch. 1..

Evans, M. & Howard, D.M. (1993). "Larynx Closed Quotient in Female Belt and Opera Qualities: a Case Study", Voice, **2**, (1), 7-14.

Fant, G. (1960). "Acoustic Theory of Speech Production", The Hague: Mouton.

Fant, G. (1975). "Nonuniform Vowel Normalization", Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Stockholm: Royal Institute of Technology, **2**, (3), 1-19.

Fant, G. (1986). "Glottal Flow: Models and Interaction", Journal of Phonetics, **14**, 393-399.

Fletcher, H. (1940). "Auditory Patterns", Rev. Modern Physics, **12**, 47-65.

Fletcher, H. & Munson, W.A. (1933). "Loudness, its Definition, Measurement, and Calculation", JASA, **5**, 82-108.

Fletcher, H. & Munson, W.A. (1937). "Relation Between Loudness and Masking", JASA, **9**, 1-10.

Fourcin, A. (1987). "Electrolaryngographic Assessment of Phonatory Function", J. Phonetics, **14**, 435-442.

Fourcin, A. & Abberton, E. (1971). "First Applications of a New Laryngograph", Med. and Biol. Ill., **21**, 172-182.

Fung, Y.C. (1981). "Biomechanics", New York: Springer.

Gerber, S.E., & Bauer, B.B. (1974). "Loudness", In: Introductory Hearing Science: Physical and Psychological Concepts, S.E. Gerber (Ed.), Philadephia: Saunders, 151-171.

Gelfand, S.A. (1981). "Hearing: An Introduction to Psychological and Physiological Acoustics", New York: Marcel Dekker.

Gobl, C. (1988). "Voice Source Dynamics in Connected Speech", STL-QPSR, **21**, 123-159.

Goldstein, J.L. (1973). "An Optimum Processor Theory for the Central Formation of the Pitch of Ccomplex Tones", JASA, **54**, 1496-1516.

Gould, W.J. (1977). "The Effect of Voice Training on Lung Volumes in Singers and the Possible Relationship to the Damping Factor of Pressman", JRS, **1**, 3-15.

Hall, D.E. (1980). "Musical Acoustics: an Introduction", Belmont: Wadsworth.

Handel, S. (1989). "Listening", Cambridge, Mass.: MIT Press.

Heffner, R. (1950). "General Phonetics", Madison: University of Wisconsin Press.

Helmholtz, H.L.F. von (1954). "On the Sensation of Tone as a Physiologucal Basis for the Theory of Music", repr. New York: Dover (originally 1863).

Henton, C.G., & Bladon, A.W. (1985). "Breathiness in Normal Female Speech: Inefficiency versus Desirability", Language and Communication **5,** (3), 221-227.

Hertegard, S., Gauffin, J., & Sundberg, J. (1990). "Open and Covered Singing as Studied by Means of Fiberoptics, Inverse Filtering, and Spectral Analysis", J. Voice, **4**, (3), 220-230.

Hess, D.A. (1959). "Pitch, Intensity, and Cleft Palate Voice Quality", JSHR, **2**, 113-25.

Hirano, M. (1983). "The Structure of the Vocal Folds", Vocal Fold Physiology, K. Stevens and M. Hirano (Eds.), Tokyo: University of Tokyo, 33-34.

Hirano, M. (1975). "Phonosurgery: Basic and Clinical Investigations", Otol. Fukuoka, **21**, Suppl 1.

Hirano, M., Hibi, S., & Sanada, T. (1989). "Falsetto, Head/Chest, and Speech Mode: An Acoustic Study with Three Tenors", J. Voice, **3**, (2), 99-103.

Hixon, T. (1987). "Respiratory Function in Speech", In: Respiratory Function in Speech and Song, T. Hixon (Ed.), Mass.: College Hill Press, 1-54.

Hixon, T., & Hoffman, C. (1978). "Chest Wall Shape in Singing", In: Transcr. of the 7th Syposium Care of the Professional Voice, L. van Lawrence (Ed.), New York: Voice Foundation, 9-10.

Hollien, H. (1960). "Vocal Pitch Variation Related to Changes in Vocal Fold Length", JSHR, **3**, (2), 150-156.

Hollien, H. (1971). "Three Major Vocal Registers: A Proposal", Proceedings of the 7th International Congress of Phonetic Sciences, Montreal, 320-31.

Hollien, H. (1974). "On Vocal Registers", J. Phonetics, **2**, 125-143.

Hollien, H. (1983). "The Puzzle of the Singer's Formant", In: Vocal Fold Physiology: Contemporary Research and Clinical Issues, D.M. Bless & J. H. Abbs (Eds.), San Diego: College-Hill, 368-378.

Hollien, H., & Michel, J.F. (1968). "Vocal Fry as a Phonational Register", JSHR, **11**, 600-604.

Hollien, H., & Moore, G.P. (1960). "Measurements of the Vocal Folds During Changes in Pitch", JSHR, **3**, (2), 157-165.

Hollien, H., Moore, P., Wendahl, R.W., & Michel, J.F. (1966). "On the Nature of Vocal Fry", JSHR, **9**, 245-247.

Hollien, H., & Schoenhard, C. (1983). "The Riddle of the "Middle" Register", In: Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control, I. Titze and R. Scherer (Eds.), Denver: The Denver Centre for the Performing Arts, 256-272.

Holmberg, E.B., Hillman, R.E., & Perkell, J.S. (1988). "Glottal Airflow and Transglottal Air Pressure Measurements for Male and Female Speakers in Soft, Normal, and Loud Voice", JASA, **84**, (2), 511-529.

Holmes, J.N. (1983). "Research Report. Formant Synthesizers: Cascade or Parallel?", Speech Communication, **2**, 251-273.

Howard, D.M. (1995). "Variation of Electrolaryngographically Derived Closed Quotient for Trained and Untrained Adult Female Singers", J. Voice, **9**, (2), 163-172.

Howell, E. (1978). "Chest Voice-Belting", Equity News, April, 14.

Ishizaka, K., & Flanagan, J.L. (1972). "Synthesis of voiced sounds from a two-mass model of the vocal cords", Bell Syst. Tech. J., **51**, 1233-1268.

Isshiki, N. (1964). "Regulatory Mechanism of Voice Intensity Variation", JSHR, **7**, 17-29.

Johansson, C., Sundberg, J., & Willbrand, H. (1983). "X-ray Study of Articulation and Formant Frequencies in Two Female Singers". In: SMAC 83, A. Askenfelt, S. Felicetti, E. Jansson, & J. Sundberg (Eds.), Stockholm: Royal Swedish Acad. of Music, **46**, (1), 203-218.

Johnson, A., Sundberg, J., & Wilbrand, H. (1983) " "Kölning". Study of Phonation and Articulation in a Type of Swedish Herding Song." In: SMAC 83, A. Askenfelt, S. Felicetti, E. Jansson, & J. Sundberg (Eds.), Stockholm: Royal Swedish Acad. of Music, **46**, (1), 187-202.

Kahane, J. (1978). "A Morphological Study of the Human Prepubertal and Pubertal Larynx", American. Journal Anatomy, **151**, 11-120.

Karlsson, I. (1986). "Glottal Waveforms for Normal Female Speakers", J. Phonetics, **14**, 415-419.

Kent, R.D. & Read, C. (1992). "The Acoustic Analysis of Speech", London: Whurr.

Kitzing, P. (1982). "Photo- and Electroglottographical Recording of the Laryngeal Vibratory Pattern During Different Registers", Folia Phoniatrica, **34**, 234-241.

Klatt, D.H. (1980). "Software for a Cascade/Parallel Formant Synthesizer", JASA, **67**, 971-995.

Klatt, D.H. (1988).KLSYN88 Syntheiszer Manual.

Klatt, D., & Klatt, L. (1990). "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers", JASA, **87**, (2), 820-857.

Kmucha, S., Yanagisawa, E., & Estill, J. (1990). "Endolaryngeal Changes During High-Intensity Phonation Videolaryngoscopic Observations", J. Voice, **4**, (4), 346-354.

Kunze, L.H. (1964). "Evaluation of Methods of Estimating Sub-Glottal Air Pressure", JSHR, **7**, 151-164.

Large J.W. (1968). "An Acoustical Study of Isoparametric Tones in the Female Chest and Middle Registers in Singing", NATS Bulletin, **0**, 12-15.

Large, J.W. (1973). "Acoustical Study of Register Equalization in Singing", Folia Phoniatrica, **25**, 39-61.

Larsson, B. (1977). "MUSSE - Music and Singing Synthesis Equipment", Master's Thesis, Dept. Speech Comm. & Music Acoustics, Royal Institue of Technology, Stockholm, 1975.

Lawrence, V. (1979). "Laryngological Observations on Belt", JRS, **2**, 26-28.

Laver, J. (1980). "The Phonetic Description of Voice Quality", Cambridge: Cambridge University Press.

Leanderson, R., Sundberg, J, & von Euler, C. (1987). "The Role of Diaphragmattic Activity During Singing", J. Applied. Physiology, **62**, (1), 259-270.

Lieberman, P. (1977). "Speech Physiology and Acoustic Phonetics", New York: MacMillan.

Lieberman, P. & Blumstein, S.E. (1988). "Speech Physiology, Speech Perception, and Acoustic Phonetics", Cambridge: Cambridge University Press, 3-15.

Makhoul, J (1975). "Linear Prediction: A Tutorial Review", Proc. of the IEEE, **63**, (4), 561-580.

Markel, J.D., & Gray, A.H. (1976). "Linear Prediction of Speech", Springer-Verlag.

Mayer, A.M. (1876). "Researches in Acoustics", Philos. Mag., **2**, 500-507, Repr. in Schubert (1979), 193-200.

McAdams, S. (1993). "Recognition of Sound Sources and Events", In: Thinking in Sound: The Cognitive Psychology of Human Audition, S. McAdams & E. Bigand (Eds.), Oxford: Clarendon Press.

Meyer, J. (1978). "Acoustics and the Performance of Music", Frankfurt: Verlag Das Musikinstrument.

Miller, D., & Schutte, H. (1990). "Formant Tuning in a Professional Baritone", J. Voice, 4, (3), 231-237.

Miller, D., & Schutte, H. (1990). "Feedback From Spectrum Analysis Applied to the Singing Voice", J. Voice, 4, (4), 329-334.

Miller, D. & Schutte, H. (1991). "Toward a Definition of Male "Head" Register, Passagio, and "Cover" in Western Operatic Singing", Paper presented at the XXth Annual; Symposium: Care of the Professional Voice. Philadelphia, Pennsylavania.

Monsen, R.B., & Engebretson, A.M. (1977). "Study of Variations in the Male and Female Glottal Wave", JASA, **62**, (4), 981-993.

Moore, B.C.J. (1986). "Parallels Between Frequency Selectivity Measured Psychophysically and in Cochlear Implants", Scand. Audiol. Suppl., **25**, 139-152.

Moore, B. C. J. (1989). "An Introduction to the Psychology of Hearing", London: Academic Press.

Moore, G.P. (1971). "Organic Voice Disorders", Englewood Cliffs, N.J.: Prentice-Hall.

Nordström, P.E. (1977). "Female and Infant Vocal Tracts Simulated From Male Area Functions", J. Phonetics, **5**, 81-92.

Osborne, C. (1979). "The Broadway Voice, Part 1: Just Singin' in the Pain", Hi-Fidelity, **29**, (1), 57-65.

O'Shaughnessy, D., Barbeau, L., Bernardi, D., & Archambault, D. (1988). "Diphone Speech Synthesis", Speech Communication, **7**, 55-65.

Pabst, F., & Sundberg, J. (1992). "Tracking Multi-Channel Electroglottograph Measurement of Larynx Height in Singers". In: STL-QPSR **2-3**, 67-78.

Pappenheimer, J., et al. (1950) "Standardization of Definitions and Symbols in Respiratory Physiology", Federation Proceedings, **9**, 602-605.

Peretz, I. (1993) "Auditory Agnosia: a Functional Analysis", In: Thinking in Sound: The Cognitive Psychology of Human Audition, S. McAdams & E. Bigand (Eds.), Oxford: Clarendon Press.

Plomp, R. (1976). "Aspects of Tone Sensation", London: Academic.

Pollard, H.F., & Jansson, E.V. (1982). "A Tristimulus Method for the Specification of Musical Timbre," Acustica, **51**, 162.

Proctor, D.F. (1980). "Breathing, Speech, and Song". New York: Springer-Verlag, 16-43.

Rabiner, L.R., and Schafer, R.W. (1978). "Digital Processing of Speech Signals", New Jersey: Prentice Hall:

Rhode, W.S. (1978). "Some Ovservations on Cochlear Mechanics", JASA, **64**, 158-176.

Risset, J.C. & Mathews, M.V. (1969). "Analysis of Musical Instrument Tones", Physics Today, **22**, (2), 23-30.

Rossing, T.D. (1990). "The Science of Sound", Dekalb: Addison Wesley.

Rothenberg, M. (1973). "A New Inverse-Filtering Technique for Deriving the Glottal Air FLow Waveform During Voicing", JASA, **53**, (6), 1632-1644.

Rothenberg, M. (1985). "Cosi Fan Tutte and What it Means or Nonlinear Source-Tract Acoustic Interaction in the Soprano Voice and Some Implications for the Definition of Vocal Efficiency", In: Vocal Fold Physiology: Laryngeal Function in Phonation and Respiration, T.Baer, C.Sasaki, & K.Harris (Eds.), Boston: College-Hill Press, 254-270.

Rothenberg, M. (1992). "A Multichannel Electroglottograph", J.Voice, **6**, 36-43.

Rubin, H.J., LeCover, M., & Vennard, W. (1967). "Vocal intensity, subglottic pressure, and airflow relationships in singers", Folia Phoniatrica, **19**, 393-413.

Ruhl, J. (1986). "Is Singing a Dying Art?", The NATS Journal, **42**, 30-35.

Russell, G.O. (1936). "Etiology of Follicular Pharyngitis, Catarrhal Laryngitis, So-called Clergyman's Throat; and Singer's Nodes", Journal of Speech Disorders, **1**, 113-122.

Saldanha, E.L., & Corso, J.F. (1964). "Timbre Cues and the Identification of Musical Instruments", JASA, **36**, 2021-2026.

Sawashima, M., & Hirose, H. (1983). "Laryngeal Gestures in Speech Production", In: The Production of Speech, P.F. MacNeilage, (Ed.), New York: Springer-Verlag, 11-38.

Schubert, E.D. (Ed.) (1979). "Benchmark Papers in Acoustics Vol. 13: Psychological Acoustics", Pennsylvania: Dowden, Hutchinson & Ross.

Schutte, H.K. (1980). "The Efficiency of Voice Production", Netherlands: Groningen Druk.

Schutte, H., & Miller, D. (1993). "Belt and Pop, Nonclassical Approaches to the Female Middle Voice: Some Preliminary Considerations", J. Voice , **7**, (2), 142-150.

Scully, C. and Allwood, E. (1983). "Simulation of Singing with a Composite Model of Speech Production", In: SMAC 83, A. Askenfelt, S. Felicetti, E. Jansson, & J. Sundberg (Eds.), Stockholm: Royal Swedish Acad. of Music, **46**, (1), 247-260.

Seidner, W., Schutte, H., Wendler, J., and Rauhut, A. (1983). "Dependence of the High Singing Formant on Pitch and Vowel in Different Voice Types", In: SMAC 83, A. Askenfelt, S. Felicetti, E. Jansson, & J. Sundberg (Eds.), Stockholm: Royal Swedish Acad. of Music, **46**, (1), 261-268.

Sellick, P.M., Patuzzi, R., & Johnstone, B.M. (1982). "Measurement of Basilar Membrane Motion in the Guinea Pig Using the Mössbauer Technique", JASA, **72**, 131-141.

Sodersten, M., and Lindestad, P. (1990). "Glottal Closure and Perceived Breathiness during Phonation in Normally Speaking Subjects", JSHR, **33**, 601-611.

Sonninen, A. (1956). "The Role of the External Laryngeal Muscles in Length Adjustment of the Vocal Cords in Singing", Acta Otolaryngol, Suppl 2, 130.

Sonninen, A. (1968). "The External Frame Function in the Control of Pitch in the Human Voice", In: Sound Production in Man, New York: New York Academy of Science, 68-90.

Sunaga, Y. (1971) "A Study on the Singing Voice. A Physiological Experiment on an Opera Singer", Japanese Journal Logopedics Phoniatrics, **12**, 53-61.

Sundberg, J. (1974). "Articulatory Interpretation of the "Singing Formant"", JASA, **55**, 838-844.

Sundberg, J. (1978). "Synthesis of Singing", Swedish Journal of Musicology, **60**, 107-112.

Sundberg, J. (1987). "The Science of the Singing Voice", Dekalb: Northern Illinois University Press.

Sundberg, J. (1991). "The Science of Musical Sounds", San Diego: Academic Press.

Sundberg, J., and Askenfelt, A. (1983). "Larynx Height and Voice Source: A Relationship?", In: Voice Physiology, D.M. Bless & J.H. Abbs (Eds.), San Diego: College-Hill, 307-316.

Sundberg, J., Gramming, P., & Lovetri J. (1993) "Comparisons of Pharynx, Source, Formant, and Pressure Characteristics in Operatic and Musical Theatre Singing", Journal of Voice, **7**, (4), 301-309.

Sundberg, J., and Nordström, P.-E. (1983). "Raised and Lowered Larynx: The Effect on Vowel Formant Frequencies", JRS, **6**, 7-15.

Terhardt, E. (1974). "On the Perception of Periodic Sound Fluctuations (Roughness)", Acustica, **30**, 201-13.

Terhardt, E.(1980). "Toward understanding pitch perception: problems, concepts and solutions", In: Psychophysical, Physiologcial and Behavioural Studies in Hearing: Proceedings of the 5th International Synposium on Hearing, The Netherlands: Delft University Press, 353-360.

Terhardt, E. & Fastl, H. (1971). "Zum Einfluss von Störtönen und Störgerauschen auf die Tonhöhe von Sinustönen", Acustica, **25**, 53-61.

Titze, I.R. (1989). "Physiologic and Acoustic Differences Between Male and Female Voices", JASA, **85**, 1699-1707.

Van Riper, C., & Irwin, J.V. (1958). "Voice and Articulation", Englewood Cliffs, N.J.: Prentice-Hall.

Vennard, W. (1967). "Singing, the Mechanism and the Technique", New York: Fischer.

Vennard W., and Hirano, M. (1973). "The Physiological Basis for Vocal Registers", In: Vocal Registers in Singing, J.W. Large (Ed.), The Netherlands: Mouton & Co., 45-58.

Wang, S. (1983). "Singing Voice: Bright Timbre, Singer's Formants and Larynx Positions." In: SMAC 83, A. Askenfelt, S. Felicetti, E. Jansson, and J. Sundberg (Eds.), Stockholm: Royal Swedish Acad. of Music, **46**, (1), 313-322.

Watson, P., J., & Hixon, T.J. (1985). "Respiratory Kinematics in Classical (Opera) Singing", JSHR, **28**, 104-22.

Wawezynek, J. (1989) "VLSI Models for Sound Synthesis", Current Directions in Computer Music Research, M. V. Mathews & J.R. Pierce (Eds.), Cambridge,Mass.: MIT Press, 113-148.

Wegel, R.L., & Lane, C.E. (1924). "The Auditory Masking of one Pure Tone by Another and its Probable Relation to the Dynamics of the Inner Ear", Physics Review, **23**, 266-276.

Wendahl, R.W., Moore, P., & Hollien, H. (1963). "Comments on Vocal Fry", Folia Phoniatrica, **15**, 251-255.

Wever, E.G. (1949). "Theory of Hearing", New York: Wiley.

Whifield, I.C. (1979a). "Periodicity, Pulse Interval and Pitch", Audiology **18**, 507-512.

Whitfield, I.C. (1979b). "The Object of the Sensory Cortex", Brain Behav. Evol., **16**, 129-154.

Whitfield, I.C. (1980) "The Relation Between Pitch and Frequency in Complex Tones", In: "Psychophysical, Physiologcial and Behavioural Studies in Hearing: Proceedings of the 5th International Synposium on Hearing", The Netherlands: Delft University Press, 361-366.

Wightman, F.L. (1973). "The Pattern-Transformation Model of Pitch", JASA, **54**, 407-416.

Winkel, F. (1967). "Music, Sound and Sensation", New York: Dover.

Yanagisawa, E., Estill, J., Kmucha, S., and Leder, S. (1989). "The Contribution of Aryepiglottic Constriction to "Ringing" Voice Quality - A Videolaryngoscopic Study with Acoustic Analysis", J. Voice, 3, (4), 342-350.

Yanagisawa, E., Kmucha, S., and Estill, J. (1990). "Role of the Soft Palate in Laryngeal Functions and Selected Voice Qualitites", Ann. Otol. Rhinol. Laryngol., 99, 18-28.

Zemlin, W.R. (1964). "Speech and Hearing Science", Englewood Cliffs, N.J.: Prentice Hall.

Zwislocki, J.J. (1978). "Masking: Experimental and Theoretical Aspects of Simultaneous, Forward, Backward, and Central Masking", In: Handbook of Perception, E.C. Carterette & M.P. Friedman (Eds.), New York: Academic, 4, ch.8.

Appendices

Appendix [A]

```
/****************/
/* G4OPERA.SPK */
/* ************************************************************************** */
/* D.M. Howard April 1993 - for use with kspan using kspandoc        */
/* based on Mark Huckvale's example .spk file from kspan manual pages  */
/* In order for the .doc version (kspandoc) to work, all parameters    */
/* MUST be set up to some value at the start                          */
/* ************************************************************************** */
/* NOTES: F0 (or FX) values are in Hz and NOT in *10Hz as for klsyn88a */
/*       Call SET_DEFAULTS at the start to initialise default values  */
/* ************************************************************************** */

/* Set up the values of klsyn88a constant parameters */
/* these can be edited if desired but format MUST be preserved */
/* NOTE: kspan sets DU and UI at present */
extern struct klpars { char    *name;
            char   type;
            short  min;
            short  val;
            short  max;
            char   *desc;
};
/* set the synthesiser update interval in ms */
int    update_interval = 5;

struct klpars consts[] = {
/* DU is set by kspan as determined by your synthesis length */
/* UI takes the value set for 'update-interval' defined above */
/* ****** THEREFORE ALTERING THESE TWO LINES MAKES NO DIFFERENCE ****** */
{"DU", 'C',  30,  300,  5000, "Duration of synthesis utterance, ms"},
{"UI", 'C',   1,    5,   20, "Update interval for parameter reset, ms"},
/* ************************************************************************** */
{"SR", 'C', 5000, 29000, 20000, "Output sampling rate, in samples/sec"},
{"NF", 'C',   1,    6,    6, "Number of formants in cascade branch"},
{"SS", 'C',   1,    2,    3, "Source (1=impulse, 2=natural, 3=LF model)"},
{"RS", 'C',   1,    8, 8191, "Random seed"},
{"SB", 'C',   0,    1,    1, "Same noise burst, (0=no, 1=yes)"},
{"CP", 'C',   0,    0,    1, "Excitation by AV, 0 cascade, 1 parallel"},
{"OS", 'C',   0,    0,   20, "Output selector (0 = normal)"},
{"GV", 'C',   0,   60,   80, "Overall gain for AV, dB"},
{"GH", 'C',   0,   60,   80, "Overall gain for AH, dB"},
{"GF", 'C',   0,   60,   80, "Overall gain for AF, dB"}
};

#define SILENCE { \
            AV(0,LOG); AH(0,LOG); AF(0,LOG); \
            A2F(0,LOG); A3F(0,LOG); A4F(0,LOG); \
            A5F(0,LOG); A6F(0,LOG); AB(0,LOG); \
            ANV(0,LOG); A1V(0,LOG); A2V(0,LOG); \
            A3V(0,LOG); A4V(0,LOG); ATV(0,LOG); \
            }

#define SET_DEFAULTS { \
            SILENCE \
            F0(100,FIX); OQ(50,FIX); SQ(200,FIX); \
            TL(0,FIX); FL(0,FIX); DI(0,FIX);  \
            F1(500,FIX); B1(60,FIX); DF1(0,FIX); DB1(0,FIX); \
            F2(1500,FIX); B2(90,FIX); F3(2500,FIX); B3(150,FIX); \
            F4(3250,FIX); B4(400,FIX); F5(4700,FIX); B5(400,FIX); \
            F6(4990,FIX); B6(500,FIX); FNP(280,FIX); BNP(90,FIX); \
            FNZ(280,FIX); BNZ(90,FIX); FTP(2150,FIX); BTP(180,FIX); \
            FTZ(2150,FIX); BTZ(180,FIX); B2F(250,FIX); B3F(300,FIX); \
            B4F(320,FIX); B5F(360,FIX); B6F(1500,FIX); ANV(0,FIX); \
            }

/* ********************** */
/* SYNTHESIS STARTS HERE */
/* ********************** */
```

```
SYNTH()
{
/* synthesis 0 */
AT(0);
      /* initialise parameters */
      SET_DEFAULTS;

/* ***********************************    */
/* YOUR ALTERATIONS SHOULD START HERE .. */
/* ***********************************    */
AT(0);

      FX(390,VIB);
      F1(700,VIB); F2(1450,VIB); F3(2800,VIB); F4(3250,VIB); F5(3600,VIB); F6(4000,VIB);
      B1(60,VIB);  B2(90,VIB);  B3(200,VIB);  B4(200,VIB);  B5(200,VIB);  B6(150,VIB);
      AV(0,LIN);
      OQ(74,FIX);

AT(15);
      AV(55,VIB);
AT(925);
      AV(55,VIB);
AT(1000);
      FX(390,FIX);
      OQ(74,FIX);
      F1(700,FIX); F2(1450,FIX); F3(2800,FIX); F4(3250,FIX); F5(3600,FIX); F6(4000,FIX);
      B1(60,FIX);  B2(90,FIX);  B3(200,FIX);  B4(200,FIX);  B5(200,FIX);  B6(150,FIX);
      AV(50,FIX);

WAIT(15);
      /* make this synthesis */
      FLUSH;

}
```

****

Appendix [B]

```
/* G4BELT.SPK */

{"NF", 'C',   1,   8,   6, "Number of formants in cascade branch"},

AT(0);
      FX(390,VIB);
      F1(880,VIB); F2(1650,VIB); F3(2738,VIB); F4(3341,VIB); F5(3890,VIB); F6(4140,VIB);
      B1(60,VIB);  B2(60,VIB);  B3(175,VIB);  B4(250,VIB);  B5(150,VIB);  B6(200,VIB);
      AV(0,LIN);
      OQ(44,FIX);

AT(15);
      AV(55,VIB);
AT(925);
      AV(55,VIB);
AT(1000);
      FX(390,FIX);
      OQ(44,FIX);
      F1(880,FIX); F2(1650,FIX); F3(2738,FIX); F4(3341,FIX); F5(3890,FIX); F6(4140,FIX);
      B1(60,FIX);  B2(60,FIX);  B3(175,FIX);  B4(250,FIX);  B5(150,FIX);  B6(200,FIX);
      AV(50,FIX);
```

****

Appendix [C]

```
/* E5OPERA.SPK */

{"NF", 'C',   1,   8,   6, "Number of formants in cascade branch"},

AT(0);
      FX(645,VIB);
      F1(800,VIB); F2(1300,VIB); F3(3250,VIB); F4(3700,VIB); F5(4200,VIB); F6(4600,VIB);
      B1(60,VIB);  B2(80,VIB);  B3(150,VIB);  B4(150,VIB);  B5(250,VIB);  B6(150,VIB);
      OQ(65,FIX);
```

```
      AV(0,LIN);
AT(20);
      AV(55,LIN);
AT(925);
      AV(55,LIN);
AT(1000);
      FX(645,FIX);
      F1(800,FIX); F2(1300,FIX); F3(3250,FIX); F4(3700,FIX); F5(4200,FIX); F6(4600,FIX);
      B1(60,FIX);  B2(80,FIX);  B3(150,FIX);  B4(150,FIX);  B5(250,FIX);  B6(250,FIX);
      AV(50,FIX);

*************************************************************************************
```

Appendix [D]

```
/* E5BELT.SPK */

{"NF", 'C',   1,   10,    6, "Number of formants in cascade branch"},

AT(0);
      FX(651,VIB);
      F1(1225,VIB); F2(1800,VIB); F3(3250,VIB); F4(4250,VIB); F5(4900,VIB); F6(5400,VIB);
      B1(90,VIB);  B2(120,VIB);  B3(180,VIB);  B4(250,VIB);  B5(250,VIB);  B6(200,VIB);
      OQ(40,VIB);
      AV(0,LIN);

AT(20);
      AV(55,VIB);
AT(980);
      AV(55,VIB);
AT(1000);
      FX(651,FIX);
      F1(1225,FIX); F2(1800,FIX); F3(3250,FIX); F4(4250,FIX); F5(4900,FIX); F6(5400,FIX);
      B1(90,FIX);  B2(120,FIX);  B3(180,FIX);  B4(250,FIX);  B5(250,FIX);  B6(200,FIX);
      OQ(40,FIX);
      AV(50,FIX);
```

Appendix [E]

Lists the answers given by the judges, arranged in order of musical experience, to questions asked after the perceptual tests.

1. Were you aware that half of the vowels were real and half were synthesized?

PG - Yes

MT - Yes

ME - Yes

IG - No. I could recognise a proportion were synthesized.

JT - Vaguely, I knew some were synthesized

NG - Yes

TB - Yes - well I suspected anyway.

CR - No, but I thought some of the higher ones might have been transposed versions of the lower ones

IH - Some sounded synthesized.

JE - Yes

2. Could you tell which ones were real and which ones were synthesized?

PG - Some

MT - Some seemed to lack "fullness" of the sound.

ME - Yes

IG - Some appeared synthesized.

JT - Most of the time

NG -Yes

TB - I thought I could, mostly

CR - There was a subtle difference. You could tell.

IH - Mainly high requency ones and mainly belt

JE - Yes. The synthesized ones sounded very restrained - they lacked "roundness" and "fullness".

3. Did the synthesized vowels sound realistic?

IG - Most sounded real. The higher registers sounded less realistic.

JT - The ones that I noticed didn't

NG - No

PG - Most

MT - Yes. Some seemed to lack "fullness" of the sound.

ME - Low pitch - yes. High pitch - not so realistic

TB - Sounded a bit loopy, especially the high opera ones

CR - Yes. But maybe not so much compared to the real ones.

IH - Some of the lower frequencies

JE - No

4. Did the real vowels sound realistic?

PG - Yes

MT - Yes.

ME - Yes - significantly more than the synthesized

IG - Most sounded real. The higher registers sounded less realistic.

JT - Mostly

NG - Yes

TB - Apart from being chopped, yes

CR - Yes

IH - Mainly the lower opera

JE - Yes

5. Could you make out the vowel quality?

IG - Yes, especially in opera.

JT - No

NG - Sometimes

PG - Yes

MT - Yes

ME - Yes, particularly with opera and low-pitch.

TB - What's that?

CR - Yes, but not so well with the higher notes.

IH - Sometimes

JE - Not really

6. How did you make your judgements on whether it was an opera sound or belt sound?

IG - Opera - clearer, more defined vowel sound. Belt - rougher quality, less defined vowel sounds, more gritty quality.

JT - "Harshness" of sound - belt seemed harsher than the opera

NG - Belt sounds rougher and less rounded

PG - Belt - sounded harder, sharp attack ,more higher freuqencies. Belt - had vibrato, more lower frequencies.

MT - I expected the synthesized vowels to be gritty, or not as rich as the real vowels.

ME - The opera vowels tend to have a high frequency vibrato on them. The pitch variations on the Belt vowels tended to be slower - and less "precise" than the opera.

TB - The belt ones seemed to have more presence (and top harmonics, to a lesser extent).

CR - By how the sound was being voiced. The belt sounds tend to be more "natural" or "organic", the opera voices sound more controlled.

IH - The opera sounds were smoother less comlex sounding more like a pure sinewave, whereas the belt has more of an edge.

JE - Instinct. Opera had a particular sound quality so if it wasn't opera it became "belt".

7. Did you have to guess at all?

IG - With maybe 5%

JT - Once or twice.

NG - Yes

PG - No

MT - Yes.

ME - Yes - for some of the synthesized vowels.

TB - Yes, can't remember why though.

CR - Yes, for a few of the lower ones.

IH - No

JE - No

8. Any other comments?

MT - Some of the vowels did seem absolutely identical, to the extent I had to guess completely.

ME - It seemed easier to make a judgement with low-pitch vowels than with higher ones. The main difference between opera and belt was high-frequency vibrato on the opera vs. a lower and more variable variations on the belt. Belt vowels often also seemed to make abrupt pitch/register variation wheras the opera flowed more continuously.

IG - A good time - I recommend this test.

NG - Cheers for the beer.

TB - Changed my mind about a few when I heard the next ones. Also found it more difficult at first while I was getting used to the test.

CR - The synthesized vowels threw me slighty. I thought they were processed versions of the "real" ones. This was where I had to guess on a few occasions.

IH - The sounds tended to cut off very abruptly thus creating a very off-putting click.

JE - There were some sounds which whilst not sounding "opera", did appear to be an attempt at synthesizing opera so I chose opera. The lower pitches were harder to distinguish. It sounded as thought the lower pitch was too low for both the "belt" and "opera" so both sounded unnatural and constrained. The higher pitches sounded much more natural for both so were much easier to distinguish between [possibly related to her personal experience - she thinks she finds it difficult to sing at lower pitches because she has no formal training].