SOUTHAMP

(Unless

UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND APPLIED SCIENCES

INSTITUTE OF SOUND AND VIBRATION RESEARCH

AN INVESTIGATION OF SPEECH SYNTHESIS PARAMETERS

by Richard Douglas Wright

ACKNOWLEDGEMENTS

This research was supported by a grant from the UK Science and Engineering Research Council, and a Cooperative Award in Science and Engineering from the IBM (UK) Science Centre. Also I was supported throughout by six hours per week of employment at the Royal National Institute for the Deaf.

Theoretical work was performed at the Institute of Sound and Vibration Research (ISVR), University of Southampton; the experimental work was performed at IBM. I am grateful to all the staff at both institutions for the use of their facilities.

This research was made possible by my two supervisors. Dr S J Elliott of the ISVR arranged the SERC-CASE grant, gave unstintingly of his time, set a model for efficiency under stress, and is chiefly responsible for keeping the pace of the work on schedule. Dr D A Sinclair of IBM also was very generous with his time, guided my use of the IBM systems and equipment, and gave up his own PC for my use during the final year of the research.

I should like to thank my principal daytime colleagues at the Abbey Parent and Toddler group, and the staff at E P Collier Nursery School who gave me two extra daytime thesis hours.

The most important thanks are to Annabelle and Ben for giving up family evenings and weekends for more than three years.

Table of Contents page

Appendices:

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCES
INSTITUTE OF SOUND AND VIBRATION RESEARCH

Doctor of Philosophy

AN INVESTIGATION OF SPEECH SYNTHESIS PARAMETERS

by Richard Douglas Wright

The model of speech production generally used in speech synthesis is that of a source modified by a digital filter. The major difference between a number of models is the form of the digital filter. The purpose of this research is to compare the properties of these filters when used for speech synthesis.

Six models were investigated: (1) series resonance; (2) direct form; (3) reflection coefficients; (4) area function; (5) parallel resonance; and (6) a simple articulatory model. Types (2,3,4) are three varieties of linear predictive coding (LPC) parameters.

There are five parts to the investigation: (1) an historical survey of models for speech synthesis and their problems; (2) a formal description of the models and their analytical relationships; (3) an objective assessment of the behavior of the models during interpolation; (4) measurement of intelligibility (using a FAAF test); and (5) measurement of naturalness.

Principal results are: synthesizer types (1) to (4) are all-pole models, formally equivalent in the steady state. But when the parameters of any of the models are interpolated, consequences for motion of vocal tract resonances (formants) differ. These differences exceed the discrimination limen for formant frequency, and make a small but statistically significant difference to intelligibility, but not to naturalness. Simple linear interpolation was found to be as good as cosine or piecewise-linear interpolation. Complete lack of interpolation reduced intelligibility by 30%. Finally, the synthesis studied achieved as few place-of-articulation errors as did LPC speech, indicating that intelligibility was limited not by parameter and transition type, but by other factors such as the excitation signal, phoneme target values, and durations.

This work was supported by the UK Science and Engineering Research Council and the IBM (UK) Science Centre.

LIST OF SYMBOLS

a_i = prediction coefficients, also called direct form polynomial coefficients

A_i = amplitude of resonance i

$\underline{B}$ = back raising tongue factor (vector of loadings)

B_i = bandwidth of formant (resonance) i

BR = back raising tongue parameter, the weight on $\underline{B}$

c = speed of sound

C = consonant

Cn = continuant (consonants other than stops); Cn1=voiced, Cn2 = voiceless

C_i = centre frequency of formant (resonance) i

d = distance along one section of an n-tube model;

$\underline{E}$ = an n-component vector of estimated vocal tract diameters, for an n-tube model;

f(t) = a continuous function of time

Fi = forward volume velocity of section i

$\underline{F}$ = front raising tongue factor (vector of loadings)

FR = front raising tongue parameter, the weight on $\underline{F}$

G = gain term of a system function

h(n) = unit sample response

H(z) = Y(z)/X(z) = system function

i = formant number, vocal tract section number

J = squared error in the estimation of the vocal tract vector $\underline{X}$; also the error function in the estimation of formants using a gradient search

k_i = reflection coefficient out of vocal tract section i

k_{pi} = pressure reflection coefficient

k_{vi} = volume velocity reflection coefficient

L = length of entire vocal tract

LIP = lip opening articulatory parameter

Mx = mixed (random plus periodic) excitation

n = sample number, iteration number

N = total number of formants, total number of vocal tract sections

$\underline{N}$ = a vector giving a neutral tongue postion

r_i = $EXP(-\pi T B_i)$ = magnitude of a root or resonance coordinate in the Z-plane

Q = the ratio of resonance centre frequency to bandwidth

R = sampling rate

Rx = random (voiceless) excitation

R_i = the area ratio between two vocal tract sections

S = stop consonant; S1=voiced, S2=voiceless

S_i = cross-sectional area of a vocal tract section

t = time, propogation time across one vocal tract section

$\underline{T}$ = an nx2 matrix; the first column is $\underline{F}$, the second is $\underline{B}$; each is a vector of loadings

V = vowel

Vx = voiced (periodic) excitation

$\underline{W}$ = a matrix of articulatory position; elements w1 and w2 are weights for tongue factors $\underline{F}$ and $\underline{B}$; w1, w2 are used to estimate FR, BR (respectively).

$x(n)$ = input sequence

$X(z)$ = z-transform of $x(n)$

$\underline{X}$ = vector of vocal tract diameters, estimated by $\underline{E}$

$y(n)$ = output sequence

$Y(z)$ = z-transform of $y(n)$

z = discrete frequency variable of the z-transform

Z_i = specific acoustic impedance of a vocal tract section; a real quantity for progressive plane waves.

$\delta(n)$ = the delta function

ϵ = a small step in a gradient search

μ = rate of adaptation in a gradient search

τ = sampling interval

σ_i = $-\pi B_i$ = resonance factor controlling damping

w_i = $2\pi C_i$ = resonance factor controlling frequency

θ_i = $2\pi\tau C_i$ = angle of a root, or resonance coordinate in the z-plane

Notation: $\underline{V}'$ is the transpose of any vector $\underline{V}$.

$\langle \underline{X}, \underline{Y} \rangle$ is the inner product of vectors $\underline{X}$ and $\underline{Y}$.

Vector and matrix quantities are underlined.

Complex quantities are in **boldface**.

$\delta x/\delta y$ indicates a partial derivative.

∇ is 'del', the vector differential operator.

EXP(x) is the exponential operation.

Phonetic symbols: symbols used in the manuscript are listed together with the symbols of the International Phonetic Alphabet (IPA) and an word to indicate pronunciation.

sound category	symbol	IPA	interpretation
vowels:	i	i	beet
	I	ɪ	bit
	e	ɛ	bet
	æ	æ	bat
	a	ɑ	father
	^	ʌ	but
	O	o	boat (American vowel)
	o	ɔ	bought
	U	ʊ	cook
	u	u	boot
approximants:	w	w	wet
	j	j	yet
	r	r	red
	l	l	led
fricatives:	f	f	fan
	v	v	van
	θ	θ	thin
	D	ð	than
	s	s	Sue
	z	z	zoo
	S	ʃ	shoe
	Z	ʒ	measure
stops:	p	p	pin
	b	b	bin
	t	t	tin
	d	d	din
	k	k	cave
	g	g	gave
nasals:	m	m	man
	n	n	Nan
	N	ŋ	hang

Chapter One: Introduction

1.1 THE PROBLEM

Synthetic speech requires a model of speech generation. The model which is commonly used in the electronic synthesis of speech is that of a source modified by a digital filter. The major difference between a number of models which have been used or proposed is the form of the digital filter. The purpose of this research is to compare the properties of several types of filter, as used for speech synthesis. The question asked in this research is: what are the differences between the available models for electronic speech synthesis?

Specifically considered are resonance synthesizers, both serial and parallel; linear prediction coefficients of three types (prediction, reflection, and area); and derivation of area function coefficients from an articulatory model.

1.1.1 Possible Answers

There are many ways in which approaches to synthesis may differ, ranging from the purely theoretical to the very practical. An example of the first is consideration of the properties of polynomials, when the synthesizers are represented as rational functions in the frequency domain. Examples of the second are practical considerations such as cost and complexity of a hardware implementation.

This research will consider four areas of potential difference:

1) Theory - Formal description of the synthesizers as digital filters, including analytical relations between the descriptions. Specific attention is given to:
   - a- differences of system order;
   - b- all-pole vs poles-plus-zeroes models;
   - c- stability.

2) Interpolation - Systems which are formally identical in the steady state are shown to have different interpolation paths. For example, the linear interpolation of reflection coefficients does not give linear formant transitions. This thesis will investigate interpolation paths in detail.

3) Intelligibility - Formal differences or disparities with regard to interpolation paths may or may not affect human perception of the resultant synthetic speech. Intelligibility tests results are given for all the synthesizers studied.

4) Naturalness - Intelligibility is not the only important subjective dimension. The issue of naturalness was also investigated using listening tests.

1.1.2 Methodology

The methods used to determine synthesizer differences within the four areas just named are as follows:

1) Analysis - All the sythesizers may be formally represented as rational functions H(z) in the discrete frequency domain, as a common basis for theoretical comparison.

2) Objective investigation of interpolation - Parameter trajectories were computed as linear paths in each of the six types of synthesizer control parameter space. These paths were then converted to formant frequency and bandwidth equivalents. Graphical and numerical results were obtained. The data were eight nonsense words covering the range of formant motion and speech sound contrasts. Additional data were spectra, and location of zeroes for the parallel case.

3) Intelligibility - The standard Institute of Hearing Research Four Alternative Auditory Feature (FAAF) test was used, an 80-item closed response wordlist. The principal result is an intelligibility score, a simple percentage of words correctly responded to. Additional analysis of error types in terms of speech contrasts was also performed.

4) Naturalness - There is no standard definition of naturalness, and no commonly-used procedure. A novel test was developed using single word material. The test allows item-by-item rating, and mixing of speech types within a test. Ten varieties of real and synthetic speech were evaluated in one 80-item test.

1.2 ORGANISATION OF THE THESIS

The description of the work divides into two parts. First there are four chapters covering all the issues that needed investigation before the synthesizers could actually be implemented. These preliminary chapters cover:

Ch. 2) the background: literature review and history of speech synthesis, and a review and discussion of general problems in electronic speech synthesis;

Ch. 3) theory: the formal description of the six synthesizers as digital filters, both as z-domain system functions and as recurrence relations; formal properties of these representations; analytical relationships for converting from one parameteric representation to another.

Ch. 4) details of a parallel resonance model, with particular reference to setting and normalisation of gain.

Ch. 5) methods for obtaining articulatory parameter values, and the results of an iterative search method for obtaining those articulatory configurations whose resonant frequencies were a best fit to a standard table of formant values for a full set of English phonemes.

Next there are five chapters describing five experimental investigations of the synthesizers:

Ch. 6) objective examination of interpolation paths; formant transitions for all six synthesizers are studied (using linear interpolation) for a set of nonsense words spanning the range of formant variation, and vowel and consonant types; graphical and numerical results are presented;

Ch's 7-9) intelligibility measurements, as a function of:

a- synthesizer type (except articulatory);

b- interpolation method;

c- articulatory parameters;

Results of the FAAF procedure are presented and analysed; various types of processed natural speech are also studied, as an upper bound for the expected intelligibility of the synthetic speech;

Chapter 7 contains all the detail of the synthesis procedure, experimental method and analysis of results.

Chapter 8 is much shorter, covering just results on linear, piecewise linear, cosine and discontinuous methods of interpolation.

Chapter 9 covers results on the articulatory approach to synthesis, including application of one very simple temporal constraint on parameter variation. This chapter also presents results on

natural speech, with and without a carrier phrase. This chapter ends with a detailed analysis of error types, for the data in all three intelligibility tests (Chapters 7, 8 and 9).

Ch. 10) the fifth chapter on experiments describes the naturalness test, and gives results for all the speech types studied in the intelligibility tests.

The final chapter (11) lists the conclusions. This is followed by several appendices giving detailed results and analysis. The last appendix is a reprint of the author's chapter on 'Basic Properties of Speech'; the approach and material in that chapter are relevant to both the implementation of the synthesis and the interpretation of the experimental results.

1.3 ORIGINALITY

This study was performed with the support of the IBM (UK) Science Centre, Winchester. However this work was not a part of their large project on speech synthesis. Although the research was supervised jointly by IBM and Southampton University, the work to be described in the following chapters is original. Preparation of the experiments made use of the signal processing language IAX (Jackson, 1984), and the speech manipulation system SAY (Speech Research Group, 1986). Also use was made of the IBM digital-to-analogue conversion hardware and software, and their listening room. I was greatly helped by access to this equipment and to the staff of the IBM speech group, and by the generous committment of time and patience provided by my IBM supervisor; but the research work itself is that of the author.

Chapter Two: History; a review of synthesis and problem areas.

2.1 Review of Models

This chapter reviews the literature of electronic speech synthesis, and ends with a discussion of unsolved problems.

There are two main approaches to the generation of synthetic speech:

I- simulation of the human speech mechanism;

II- simulation of the acoustic output.

The first category of models are mechanistic, in that they involve modelling of human anatomy and physiology. The earliest of these models were actual mechanical contrivances. The first electronic synthesizers, however, were analogues of speech generation (rather than direct models): currents and voltages through the circuitry are the analogue for airflow and pressure through the human vocal tract. This is in contrast to the purely mechanical models of earlier centuries in which airflow and pressure were directly modelled with the various bellows, reeds and tubes of these synthesizers.

The second category of model only considers the acoustic output: there is nothing within the synthesizer corresponding to the shape and operation of the human apparatus; the point of comparison is just the output of the system. The electronic versions of acoustic synthesizers attempt to simulate the waveforms or spectra observed at the output of the human speaker.

Certain terminology has arisen in work on speech synthesis. An electrical analogue of the vocal tract is the transmission line, which leads to the acronym LASS - Line Analogue Speech Synthesizer. An electrical anologue which models only the

output of a system is one which has the same properties at the electrical terminals, which leads to the acronym TASS - Terminal Analogue Speech Synthesizer.

Properly both the vocal tract and the transmission line are distributed systems, described by partial differential equations. In practice the system is approximated by a series of lumped-element sections forming a bilateral network. Such a model is formally described by ordinary differential equations.

A discrete time or sampled data implementation of the model (as used in digital signal processing) is then in terms of difference equations. Alternatively, wave propogation can be simulated, again by difference equation (Kelly and Lochbaum, 1962; Mermelstein, 1971).

Finally, the generality of digital simulation allows the electronic analogy to be dropped altogether, and one can attempt direct simulation of acoustic phenomena within the vocal tract, such as noise generation from turbulent airflow (Scully, 1979; see Kaiser, 1983 for a review). These aerodynamic models are mechanistic in the sense used above, but incorporate acoustic phenomena within the mechanism.

A general review of speech synthesis is Flanagan, 1972. Significant publications specifically on synthesis are Holmes (1972) and Linggard (1985). Many important papers up to 1972 are in the Benchmark Papers in Acoustics volume on speech synthesis, edited by Flanagan and Rabiner (1973).

2.1.1 Mechanistic: Vocal Tract Models

Vocal tract models can be divided into:-

(a) models of the vocal tract itself, usually represented by either area functions or reflection coefficients (section 2.1.1.1).

(b) models, usually with a reduced number of dimensions, which generate a vocal tract from other control parameters (section 2.1.1.2). These parameters may (amongst other methods) be derived from consideration of physiological constraints or other anatomical features (such as tongue position). The terminology 'articulatory model' is usually reserved for this second type of vocal tract synthesis.

2.1.1.1 Vocal tract models

A. Area Functions

The vocal tract may be represented in terms of cross-sectional area vs distance along the tract: an area function. To describe this function mathematically requires either a closed-form representation (such as an exponential horn) or piecewise approximation, where each element (piece; vocal tract section) has a formal representation. The simplest such element is a cylinder: a circular cross-section and constant area. Hence an electrical analogue of the human vocal tract is a set of the cross-sectional areas (or diameters) of a set of short cylinders. The use of a circular cross-section is justified on the grounds that for small cross-sectional dimensions (with respect to wavelength), the propogation of a wave in the tube is planar, and so propogation in a tube of irregular cross-section will be equivalent to propogation in a tube of circular shape. The cylinders must also be small (again, with respect to wavelength) in length as well as diameter, or the piecewise approximation becomes invalid.

The electrical analogue of an acoustic tube uses current for volume velocity and voltage for pressure. Then each cylinder in a piecewise approximation to a vocal tract is represented by a simple lumped-element electrical circuit (T or pi section), and the whole tract is a set of these circuits connected in series.

An early implementation of such an electrical analogue was that of Stevens, Kasowski and Fant (1953) which had 35 sections, each representing 0.5 cm of a notional vocal tract. An important consideration in such models is the number of sections required. Both Stevens, et al and Dunn (see below) used the rule of thumb that a section was short with respect to wavelength if the section length represented no more than 1/16 the wavelength. A 0.5 cm section was considered accurate for work up to approximately 4300 Hz.

The Stevens, et al synthesiser occupied an entire 19-inch rack, six feet high. It was a passive, static model in that the individual sections were manually adjusted. Each section could have one of 11 areas, ranging from 0.17 cm^2 to 17 cm^2 on roughly a log scale. Then with appropriate excitation a steady sound was produced. This model was still available for use at MIT in the early 1970's.

Another classic static electrical analogue was that of Fant (1960), with 16 sections in the electrical analogue, and 20 sections in a digital simulation.

B. Reflection Coefficients

A representation equivalent to an area function is to specify the reflection coefficient (either of pressure or volume velocity) at the boundary between each pair of cylinders. For the case of a lossless tube, reflection coefficients may be used in a ladder or lattice network or in a relatively simple set of equations for digital simulation of such a network. A resistance may be added as an attenuation factor between each stage of the lattice.

One of the earliest all-digital synthesisers (Kelly and Lochbaum, 1962) used reflection coefficients in a tube that was lossless except for a fixed attenuation of 127/128 between sections. Both Fant (1960) and Flanagan (1972) include frequency dependent terms in all of their models of vocal tract losses, however simplified. Kelly and Lochbaum

give no explanation for this attenuation term. However Stevens et al (1953) included terms representing resistance and conductance per unit length, and referred to early work by Fant (1950) for "dissipation in the cavity of the vocal tract ... which is approximately independent of frequency" (Stevens et al, 1953, p736).

An advantage of the all-digital approach was the relatively easy implementation of synthesis of continuous, connected speech. The devices of Stevens and Fant were limited to steady sounds.

The Kelly and Lochbaum simulation used 21 stages and was operated at a sampling rate of 20 KHz. Thus 21 stages were used for a 10Kz bandwidth, rather than the 35 stages for a 4.3 KHz bandwidth in the Stevens device. The 16-to-1 rule of thumb for wavelength to section length has effectively become reduced to 4-to-1. This happens because the sampling interval of the digital simulation is used as twice the propogation delay of one cylindrical section, and the notion of constructing an analogue which is accurate only for dimensions 'small with respect to wavelength' is lost.

By definition:

$$\tau = 2*t \qquad (2.1)$$

$$R = 1/\tau \qquad (2.2)$$

$$t = d/c \qquad (2.3)$$

$$d = L/n \qquad (2.4)$$

τ = sampling interval
t = propogation time across one section
R = sampling rate
d = Distance along one section
c = speed of sound
L = Length of entire vocal tract
n = Number of sections

Hence:

$$n = 2L / \tau*c = 2LR / c \qquad (2.5)$$

Using l=17cm and c=34000 cm/sec as representative values for the length of an adult vocal tract and for the speed of sound in dry air yields:

$$n = 34R / 34000 = R / 1000 \quad (2.6)$$
$$= \text{Sampling rate in kHz}$$

Kelly and Lochbaum use 20 sections (plus one for the mouth) for their 20KHz sampling rate, and computed speech waveforms with a 10KHz (maximum) bandwidth. The fact that their section represents 17cm/20 = 0.85cm, and the wavelength at 10 KHz is 34000/10000 = 3.4cm (=0.85cm x 4) is not directly considered.

The Kelly and Lochbaum system also demonstrated the ease with which nasals and fricatives could be synthesised using a vocal tract model. Nasals simply required the addition of a branching tract; fricatives had a noise source at the point of maximum constriction. This implementation of fricatives was made simple by virtue of the fact that the model was a digital implementation. Actual hardware with variably-placed sources would have been much more awkward.

Reflection coefficients are of current interest because of their position within linear predictive coding (LPC). This method of analysis-synthesis has become increasingly prevalent in speech processing since 1970 (Markel and Gray, 1976). LPC originated as a method of speech analysis, and the LPC parameters were interpreted as an acoustical model. Use of reflection coefficients was a second stage of LPC development, offering attractive stability and quantisation properties (Witten, 1982; p150). The interpretation of the results of LPC analysis in terms of actual vocal tract area functions has been an area of research and controversy for some years (Wakita, 1973; Crichton & Fallside, 1974; Sondhi, 1977; Brooks et al, 1980).

Other vocal tract parameters include area ratios and log area ratios (of interest mainly for low datarate encoding). An interesting paper by Mermelstein (1967) uses a finite

Fourier series expansion of the log of the area function in order to remove ambiguities when attempting to determine vocal tract shapes from acoustic data.

Other vocal tract synthesizers of historical interest are:

1 Rosen (1958): the first dynamic vocal tract model.
2 Fant (1960): a mathematical analyses of various two, three and four tube models, including horns and Helmholz resonators as well as cylinders.
3 Flanagan and Landgraf (1968): a difference equation simulation of a lumped-parameter network with source/system coupling. This was the first example of excitation within the model, using a self-oscillatory system.
4 Itakura (1968): use of reflection coefficients in a lattice filter.
5 Wakita (1973): derivation of pseudo area function using linear prediction;
6 Maeda (1977): identification of the problem of pressure discontinuities within a lattice filter as reflection coefficients (or areas) are updated.

2.1.1.2 Articulatory models

The terminology articulatory model is used to denote models (usually of reduced dimensionality) which represent an area function in terms of some other set of parameters, incorporating physiological or statistical constraints. These models are of interest for several reasons:
(1) comparison with physiological data (articulatory controls);
(2) restriction of area functions to physiologically valid shapes and motions (articulatory constraints).
(3) reduced datarate representation of speech;

The earliest electrical vocal tract simulation of any sort was an articulatory model, that of Dunn (1950). This was a 25 section electrical model, where each section (as with the Stevens et al 35 section model) represented 0.5 cm. (Dunn simulated only short vocal tracts). The model is

articulatory because it had fixed area but a moveable constriction. The 25 sections each had an area of 6 cm^2 in cross section. The system produced sounds of various qualities through the use of a constriction circuit which could be manually inserted (switched in) between any adjacent pair of fixed sections. Further, the degree of constriction could be varied, and a second constriction representing the mouth opening could also be varied. This gave a 3-parameter articulatory model: 'tongue-hump' position and height, and lip opening. The model was thus quite limited, as determined later in the formal analysis of two tube plus constriction models by Fant (1960). However the spectrograms of vowel sounds produced by the model are quite convincing (reproduced in Flanagan and Rabiner, 1972, p107). This device was subsequently moved to the University of Michigan, where twenty years later it was still in use as a teaching aid.

A considerable refinement was Stevens and House, 1955. Again there were three parameters: position and degree of constriction; mouth opening. In this study the electrical hardware was the same 35-section vocal tract model described above. The conversion from values of the three articulatory dimensions to the 35 areas was done off-line; originally by hand, later on the TX-0 computer. The 35 areas were set on the control knobs, and another 35 resistive settings were re-tuned (as they were linked to effective area). Finally the resonance behavior was measured, yielding information relating formants to articulatory variables. The result of this work was a set of diagrams showing the dependence of the first three formants upon the three articulatory parameters (reproduced in Flanagan and Rabiner, 1972, p120).

This three parameter model was used again twenty years later in a mathematical study of the relationships between resonance frequencies and the vocal tract shapes required for their production (Atal, et al, 1978). The Stevens and House model was used to constrain the search space.

Other notable articulatory models include:

1 Henke (1967): controlled by distinctive features;
2 Haggard (1970): six primary and six secondary parameters;
3 Coker (1976): five tongue, jaw and lip parameters;
4 Mermelstein (1971): the Haskins system;
5 Lindblom and Sundberg (1971): two tongue parameters describing a hump (and corresponding hollow) as displacements from a median position.
6 Ladefoged et al (1978): two tongue parameter dimensions account for 92% of variance for vowels;
7 Terepin (1979): a derivative of the Coker model which included an optimisation procedure for adjusting articulatory parameters to match vowel spectra.

Perkell (1977) gives a review of ten articulatory synthesizers. Orthogonal projections and dimensional analysis of vocal tract data are also relevant (Mermelstein, 1967; Liljenkrants, 1971; Wright, 1973; Sambur, 1975). The general problem of determination of parameters for articulatory synthesis-by-rule is discussed in Chapter Five.

2.1.2 Acoustic Output: Spectrum and Resonance Models

Acoustic synthesizers can be divided into those which model the overall smoothed spectral envelope of speech (vocoders), and those which specifically model resonances of the vocal tract (terminal analogues).

An electrical resonance or terminal analogue synthesizer is a network of lumped elements, representing only the transfer function (input-output relationships) of a vocal tract, not its geometry or distributed properties. Mathematically it is represented by ordinary differential equations. These can be implemented in a sampled-data representation by difference equations, allowing computer simulation of a resonance synthesizer.

2.1.2.1 Vocoders: spectrum construction

Several of the earliest electronic synthesizers constructed or reconstructed the smoothed short-term speech spectrum. This method originated with the vocoder and the Voder (synthesis part of a vocoder) of Homer Dudley (1939). The Voder had 10 bandpass filters (with roughly 1/2-octave spacing) covering the frequency band from 50 to 7500 Hz, and the gain of each band was manually controlled. Other controls were for the excitation signal and a complex control specifically for stops (which evidently produced about a 75 msec voice onset time for /p,t,k/ and about 25 msec for /b,d,g/; see Figure 14 reproduced in Flanagan & Rabiner, 1972, p 209).

There was no method to permanently store control data to allow controlled experimentation. One can only speculate on the progress in synthesis which might have been made had a player-piano mechanism been used. It is also noteworthy that the ten telephonists who trained for a year to operate the Voder represent the first sustained experiment in the use of an artificial speech prosthesis. By all accounts their progress in 'speech acquisition' was slow, and disappointingly linear (Dudley, 1939).

This work also was one of the first demonstrations of the importance of prosodics in speech synthesis: "The technique of the pitch pedal ... is most important to intelligibility" (Dudley, 1939, p25).

Speech reconstruction based on permanent patterns began with spectrographic playback devices (Schott, 1948; Cooper, Pattern Playback, 1950). The sound spectrograph was a device invented in the early 1940's for sonar work, and was applied to speech beginning in 1946 (Potter, et al, 1947; Joos, 1948). This device produced a frequency vs amplitude vs time display, originally on a rotating phospherescent tube (which could be photographed, though the polar coordinates were awkward to interpret), then a phosperescent belt (which at least gave a rectangular display) and finally

on electrosensitive paper. This was the tool that launched acoustic phonetics and the spectrogram has become the standard visual representation for short samples of continuous speech. Development of a synthesis counterpart using a stylised spectrographic representation was a logical next step. The use of a spectrographic representation allowed synthesis to be related to analysis. The Pattern Playback allowed researchers at Haskins Labs and elsewhere to begin systematic experimentation into the relations between acoustic signals and human perception; this research was one of the first uses of synthetic speech as anything other than an end in itself. Synthetic speech became a tool of scientific investigation in the 1950's.

Vocoders are devices which analyse and resynthesise speech using a parametric representation. Examples are: channel; autocorrelation; phase; cepstrum; residue; homomorphic; orthogonal; cf Flanagan (1972). All vocoders involve an aspect which can be called synthesis, though the parameters are generally obtained solely from measurement and not arranged in tables for synthesis by rule. Reasons for not using vocoders for synthesis include the reduced naturalness of these systems, and the large number and arbitrary nature of the parameters. Vocoders typically use more than twice as many parameters as do resonance synthesizers.

Other devices capable of generation of arbitrary spectra or waveforms could be used for speech synthesis, such as music synthesizers though they could have the same problems as vocoders and would not in general have a corresponding analyser.

Time domain representations of speech are only a transform away from spectrum construction, though speech generated from stored time waveforms is not usually considered synthetic. However speech construction from concatenation of individual glottal cycles amounts to a representation in terms of a finite impulse response (approximately ten msec duration is usual), and is a form of synthesis. It is capable of use for synthesis by rule of utterances completely unconnected with

the utterances used to derive the data. This method is the time-domain equivalent of synthesis from stored spectra, as in vocoders. Such a method is a model of the acoustic output, it does not specifically model resonances, and its basic unit is the time-domain equivalent of a spectrum; thus it falls closest to the category of spectrum construction techniques (Wright, 1972; Beddoes, 1982).

2.1.2.2 Resonance Synthesizers

A principal acoustical property of the vocal tract is resonance: frequencies of the appropriate wavelength add constructively as they rebound through the tract; other wavelengths are subject to cancellation to some degree. This knowledge leads to synthesis models which do not specify the overall spectrum (as in the vocoder) but directly model the individual resonances.

The basic parameters of such a synthesizer are the frequencies and bandwidths of the vocal tract resonances (formants). The first electronic synthesizer (Stewart, 1922) consisted of two simple RLC resonance circuits. The resonant frequencies and bandwidths were manually varied and the device was connected to a buzzer or motor-driven switch to produce a steady sound. Further manual adjustment allowed desired vowel sounds to be obtained by trial and error. The results were in good accord with modern theory for the 'close front' vowels /i/ and /ɛ/, where the first two formants are widely separated. For the back vowels such as /u,o,a/ a single resonance was used. These back vowels have a spacing between the first and second formants which is less than 1/3 of an octave, which is approximately the resolving power or critical bandwidth of the ear.

Stewart also recognised the importance of prosodics. He noted that the "really difficult problem involved in the artificial production of speech-sounds is not the making of a device which shall produce sounds which, in their fundamental physical basis, resemble those of speech, but in

the manipulation of the apparatus to imitate the manifold variations in tone which are so important in securing naturalness."

He also comments on the use of a resonance model vs harmonic analysis, which was a matter in dispute between Helmholz and Scripture. He commends the "steady-state" theory (meaning harmonic analysis, because Fourier analysis assumes a periodic waveform with no beginning or end) for transmission purposes, but states that such a representation is "less compact and definite" for describing vowels than is a resonance description.

The question which arises as soon an one considers more than a single resonance is how to connect them. The simplest method conceptually is a series or cascade arrangement, which in modern terminology is an all-pole or autoregressive (AR) model.

Specification of antiresonances may also be made, thus including zeroes as in a sampled data finite impulse response filter or moving average equation (MA). A 'poles plus zeroes' model is thus an ARMA model. Usually linear prediction uses an AR model, though ARMA formulations of LPC are now available (Green, 1976; Atashroo, 1976; Steiglitz, 1977; Song, 1983). Generally an ARMA model requires knowledge of both input and output of the system to be modelled. In speech the input information can at best be estimated.

Resonances may also be combined in parallel. Parallel combination gives rise to implicit (not directly specified) zeroes, which will be discussed in more detail in Chapter Four on the theory of parallel synthesis. Discussions of series vs parallel appear in Holmes (1982), Klatt (1980), Gold and Rabiner (1968) and Flanagan (1957).

A physical tube has an infinite number of resonances. Finite bandwidth models have a finite number of resonances and a correction term to replace the within-band effects of

the out-of-band resonances, the residuals. Handling of the 'higher order correction term' is a key feature of analogue series vs parallel configurations. The advent of digital simulations allows an alternative solution through the modelling of the higher terms via the inherent periodicity of the frequency response of a digital filter, thus simulating an infinite number of resonances (Gold and Rabiner, 1968). But Holmes (1982) makes the point that the actual location of higher order poles is unknown, and relying upon what amounts to upward aliasing is not a systematic way to estimate their position and effects. The crux of the matter is that without doing something about residuals the response of an analog series synthesizer with five resonances will be approaching -60dB/octave at the high frequency end of the working bandwidth, owing to a -12 dB/octave contribution from each second-order resonance term. Something must be done to correct this trend to a more nearly flat response; a uniform lossless tube has a flat trend. Holmes offers the parallel arrangement as a way of avoiding the problem.

Resonance models may also use explicit zeroes (rather than the implicit zeroes of the parallel arrangement) to simulate the effects of nasal/lateral branching and also secondary (fricative) sources within the vocal tract (Fant, 1960; Klatt, 1980).

Parallel resonance synthesizers allow considerable freedom of amplitude variation, owing to less resonance interaction than with series (cascade) resonance synthesizers. Series synthesis depends upon bandwidth variation to control amplitude. Both Stockholm and MIT currently use a hybrid formulation with a cascade section for vowels and a parallel section for most consonants (OVEII, Fant and Martony, 1962b; Klatt, 1980). It is interesting that the mechanical synthesizer of von Kempelin (reconstructed by Wheatstone) also had one path for vowels and vowel-like sounds, and four separate passages for consonants (including nasals) (Flanagan, 1972).

Automatic control of resonance synthesis began with Lawrence (1953) using optical control from hand-painted glass slides. His system had four parameters: three (parallel) frequencies plus excitation type/frequency. Bandwidth and amplitude were fixed, and all three resonances had equal bandwidth. Implementation was in the time domain using frequency shifting. Surprisingly, Lawrence describes the resultant intelligibility as "fairly good, probably quite adequate for commercial telephony."

It is interesting to note the similarities between the Lawrence method and the 'toy' synthesizer of Witten & Madams (1978) which also has fixed, equal bandwidths of two parallel resonances. In their case they needed to add amplitude control plus various fixed filters for fricatives in order to get "quite intelligible output" (Witten, 1982, p117).

Parallel synthesis has been used at MIT (Stevens, 1955) and Bell Labs (Flanagan,1955, 1965) and extensively developed by Holmes (1969, 1972; Rye and Holmes, 1982). A primary interest of Holmes has been the attempt to make samples of synthetic speech as close as possible to natural utterances; this attempt has been remarkably successful for particular utterances, though requiring great effort (Holmes, 1979).

Series synthesis has been studied at Stockholm (Fant and Martony, 1962a), MIT (Stevens, 1955; Klatt, 1972), and Bell Labs (Flanagan, 1957, 1962).

Linear prediction yields coefficients which define an all-pole cascade formulation (through use of prediction coefficients), though currently most implementations use reflection coefficients in a lattice filter. The two arrangements are formally equivalent (cf Chapter Three).

As a final point, the reflection coefficient parameters determined by LPC analysis can be transformed to an area function representation, connecting resonance and vocal tract synthesis models (Wakita, 1972; also Chapter Three, below). Limitations of the method are presented in Chapter Five.

2.2 Problem areas in speech synthesizers

The following discussion presents problems which may require examination in a comparison of synthesis models. The following areas will be considered:

Adequacy of the Source
Adequacy of the System
Source/System Interaction
Dynamics: Interpolation and Non-Stationarity

2.2.1 Adequacy of source

Speech generation can be considered from two aspects: source and system. This division is explicit in most synthesis models. Even in the case of self-oscillatory (Flanagan and Landgraf, 1968) and aerodynamic (Scully, 1979) models which incorporate excitation within the system, there is still a concept of source.

Synthesizers can be excited with signals of varying complexity: periodic or aperiodic excitations, or a combination of the two; spectral and temporal characteristics require consideration.

2.2.1.1 Periodic excitation

a) pulse shape and spectrum roll-off

One of the oldest controversies in synthesis is the nature of the excitation. A periodic source with a harmonically rich spectrum is an obvious requirement for voiced sounds, but here the agreement ends. A variety of all-pole and pole-zero shaping networks, as well as various time-domain characterisations have been used (Rosenburg, 1971). It is a commonplace in synthesis to improve naturalness by changing the excitation function (Holmes, 1973), and partly this must be a matter of supplying in the excitation what is lacking

elsewhere in the system. LPC systems, for instance, yield a perfect copy of an input (regardless of the accuracy of the analysis) if the entire error signal is used as excitation. Also vocoder naturalness is improved by transmission of a baseband (the bandlimited signal up to about 1kHz) rather than using just pulses or white noise as excitation (Flanagan, 1972).

There is no general agreement on specification of zeroes of the source (Fant, 1960). Another variable is the number of points of slope discontinuity in the excitation waveform (Rosenburg, 1971). Also one might have multiple excitation pulses within an excitation epoch, which has led to 'multipulse LPC' (Atal and Remde, 1982).

Finally, manipulation of source spectrum slope was one method of improving signal-to-noise ratios through hardware synthesizers. It was better to excite with a flat spectrum and shape at the end (Witten, 1982, pp 96-104). These considerations can be avoided in a computer implementation.

b) variation of spectrum with excitation rate

There is evidence that the spectrum roll-off is not independent of fundamental frequency (Sundberg & Gauffin, 1979). However voice quality is also related to excitation rate: people don't usually shout with a low pitch. This interaction makes it difficult when analysing real speech to determine just what relationship does hold, independent of voice quality. The problem is compounded by hardware characteristics which can cause excitation rate to effect spectrum shape and signal strength (Witten, 1982; p96-97). In general it would appear desirable if all such effects were eliminated and excitation rate, spectrum shape, and signal energy were kept separate.

c) type of phonation

There are many ways to describe overall voice quality (Laver, 1980). Many factors which do not properly characterise the source make a contribution, such as degree of pharyngeal constriction and amount of nasality. Those factors which do specify the operation of the source are described as phonatory types. Terminology is not uniform, but the underlying mechanism is variation in the operation of the vocal folds: they can snap together producing a rich spectrum (modal voice); they can be flaccid and consequently have little strength in the higher harmonics (breathy voice); they can move along their entire length or only a fraction of the folds may move (falsetto); the motion may be regular or irregular (modal voice vs harsh voice); the irregularity may be quite random or involve alternation of short and long periods (jitter); they can be closed for greater or lesser proportions of the total cycle time; varying amounts of air may flow through the larynx, producing varying amounts of turbulent airflow and hence superposition of a random noise (aspirant) component upon the regular excitation (whispery voice); and there can be considerable differences in subglottal pressure.

All of these changes may affect the spectrum of the resultant speech. One question deserving attention is the degree to which any synthesizer can accomodate research on these aspects. Holmes (1982) recommends parallel synthesizers over series arrangements because one spectral consequence - the difference in overall spectrum slope - of one phonatory effect, namely change in effort, can be simulated by adjusting formant amplitudes without altering the properties of the excitation. In general, however, phonatory variations of the various types mentioned above will require manipulation of the source regardless of system configuration.

2.2.1.2 Aperiodic excitation

a) noise spectrum and mixed excitation

It is common to use a noise generator with a flat spectrum in synthesis. For parallel resonance synthesizers in particular, any other choice would cause the amplitude controls for the fricative resonances to become frequency dependent. Yet in spectrograms for voiced fricatives it is easy to see a preponderance of periodic energy at the low frequency end, and random energy at the higher frequencies, above 3 hKz. Whispered speech also shows considerable elevation of second formant amplitude as compared with the first formant, enough indeed that it is usual in phonetics training to demonstrate the position of the second formant by whispering.

Some models (Scully, 1979) attack this issue from first principles, using aerodynamics to generate noise sources from turbulence, and thus the spectrum is whatever the aerodynamic model produces.

It may well be a simplification to use a rising characteristic on a noise source when modelling 'mixed' excitation. Many sounds have a mix of periodic and aperiodic excitation; the voiced fricatives notionally require such a mix. Phonetically the mix may vary greatly, and no voiced component at all need be present in syllable-final 'voiced' fricatives, as they may be cued by the length of the preceeding vowel (Ladefoged, 1972). Rye and Holmes (1982) use a separate mix parameter for each resonance, which might be approximated by simply adding the periodic and aperiodic sources (providing each had an appropriate spectral slope) and eliminating the mix parameters altogether. Or one might use a single mix parameter to satisfy the requirements for voiced fricatives as well as allowing for various degrees of mix to model narrow phonetic effects and overall phonatory effects.

Many simple synthesizers (the LPC chips, for instance) make

no allowance for mixed excitation; others have an additive mix (Holmes, 1972), and another possibility is a multiplicative mix to also model the interaction of the source and the system for voiced fricatives. The laryngeal source is opening and closing and hence releasing puffs of air, and the system has a constriction which produces noise (really a secondary source) from turbulent airflow whenever the laryngeal source supplies the air. This type of model has been used by Rabiner (1968a) and Klatt (1979).

b) frication and aspiration.

There is a distinction to be made between aspiration (turbulence at the glottis) and frication (turbulence at a point of constriction anywhere past the glottis). The main requirement is that the aspirated noise must pass through the ordinary vocal tract resonances, whereas the fricative noise has its own resonances in a terminal analogue system (the properties of the tract between the glottis and the point of constriction are relatively insignificant; Fant, 1960).

Thus to model aspiration vs frication in a terminal analogue model the random source must be able to feed both the vowel resonances as well as the fricative resonances. Such a distinction may be a quite minor effect, however. For instance, Rabiner (1968a) found it unnecessary to use a random component at all for the non-sibilant voiced fricatives /v/ and /D/. It becomes more important for voiceless plosives, however, where the formant motion during voiceless (aspirant) excitation is a primary cue to place of articulation. It is cumbersome and crude to approximate this effect through variation of the fricative network.

2.2.1.3 Temporal characteristics of source

The human excitation signal for speech is imperfect, with varying degrees of irregularity, including the phenomena of

alternating short and long periods (jitter) and of amplitude modulation (shimmer). The contribution of these aspects to naturalness is an open question in synthetic speech research, though some studies (Gill, 1961; Rosenburg, 1971) have concluded that temporal irregularity does not make a positive contribution. John Clark (Macquarie Univ, Australia, personal communication) has found that jitter improves naturalness if used only in certain places in an utterance.

Models of the temporal aspects of the excitation signal are of use in attempting to understand various types of speech pathology. Periodicity perception is also of interest in audiology as an important type of auditory processing.

2.2.2 Adequacy of system

The system or filter used to modulate the source signal will be considered from various points, beginning with the basic elements of a system and ending with how these elements are combined and implemented.

2.2.2.1 Lumped vs distributed elements.

The properties of the vocal tract constitute a distributed system. Early mechanical synthesizers were also distributed systems, but the advent of electronic models and computer simulations has always involved the approximation of distributed properties by lumped values. It is always worth considering, in any synthesizer, just what assumptions are involved in this approximation.

2.2.2.2 Modelling of resonances.

a) number and type of resonances

Resonance synthesizers vary in the number of resonances used for vocalic sounds, and in the number and type of resonance

used for nasals and fricatives. Explicit and implicit zeroes may be included in the nasal/fricative resonances. Typically three movable and one or two fixed resonances are used for vowels, usually working up to a 5 KHz bandwidth. Compensation may be introduced for the neglect of higher resonances.

Consonants are often approximated with a single complex pole-pair resonance, though detailed studies of nasals (Fujimura, 1962) and voiceless fricatives (Heinz and Stevens, 1961) typically require at least two complex pole-pairs and a complex zero-pair.

b) bandwidth variation

Parallel resonance synthesizers usually have fixed bandwidths. The implicit assumption is that bandwidth variation can be adequately represented by amplitude control. But detailed studies on a period-by-period basis exhibit bandwidth variations of as much as 100% (Pinson, 1963).

An even larger effect arises from bandwidth variation as a function of whether the glottis is closed (reflecting; narrow bandwidths) or open (absorbing; wide bandwidths). See section 2.3, below.

c) antiresonances

The principal source of antiresonances in speech synthesizers is the parallel resonance configuration. It is usually claimed that what matters is what a person can hear (determined by the resonances) rather than what cannot be heard (antiresonances). This view runs into difficulty when the zeroes approach the poles and their effects become distinctly audible. Thus it may be important to explicitly determine the locations of the implicit zeroes.

2.2.2.3 Complicated models

a) complicated synthesizers

Although it is convenient to speak of parallel vs series synthesizers, several major implementations are actually hybrids, using a series arrangement for vowel-like sounds and a parallel set of resonances for other sounds (OVE II, Fant, 1962; Rabiner, 1968a; Klatt, 1979).

b) complicated sounds

Three speech sound categories are departures from the uniform, unbranching tube model: nasal, lateral and retroflex sounds. These sounds may be difficult to implement in a single-tube vocal tract model (or any all-pole equivalent), although Kelly and Lochbaum (1962) demonstrated how simply nasals could be implemented through addition of a branching tube.

There is a general lack of attention to problems of lateral and retroflex configuration in the literature. Much of the work on vocal-tract modelling has been concerned specifically with vowels (Dunn, 1950; Stevens et al, 1953, 1955) or with modelling of an excitation driving a tract producing vowels (Flanagan and Landgraf, 1968; Ishizaki and Flanagan, 1972). Thus the eight papers on vocal tract simulation collected in Flanagan and Rabiner (1973) make no mention of either /r/ or /l/ sounds.

2.2.2.4 Implementation

a) analysis

In any synthesizer a question arises of obtaining data with which to operate the device. If a synthesis scheme is directly tied to a straightforward analysis procedure then

the problem is solved. The popularity of LPC for synthesis is an instance of how a rather rudimentary synthesizer has proliferated because the analysis problem was solved. A related issue is analysis-synthesis telephony, which has long been a primary motivation (and fundraiser) for synthetic speech research.

Several studies (Seeviour, Holmes & Judd, 1976; Moller, et al, 1977) use iterative methods to determine synthesizer parameters which produce a spectrum or waveform which is a good match to a natural utterance. The resultant speech is a sort of automated copy synthesis and as such usually has much higher naturalness than synthesis by rule. Given such data, one could then work backward and selectively remove or limit synthesizer properties such as mixed excitation or nasal branching or zeroes. Control parameter limitation could be introduced by fitting the derived parameter paths with suitable equations. Or the analysis could be used (in conjunction with segmentation) to determine an optimised phonetic inventory (Bridle and Chamberlain, 1983). This approach to synthesis might be called analysis-by-rule.

There is a common problem in dealing with alternative methods of synthesis by rule: comparisons must be made of obviously unnatural signals. (The same problem arises in before vs after therapy comparisons for cases of speech pathology.) The analysis-by-rule approach to synthetic speech avoids the unnaturalness problem by providing high quality speech which is still rule-governed.

b) range of variation

The design of hardware synthesizers involved developing control circuits that were linear (or at least regular in some fashion) over the range of desired operation. It was thus usual to simplify matters and operate only over a range of excitation and resonance frequencies appropriate to an adult male.

The digital implementation largely eliminates these constraints. Indeed the use of synthetic speech prosthetic devices requires the ability to generate voices appropriate to any size and shape vocal tract. This raises certain problems for articulatory synthesizers. Standard LPC analysis is tied to a fixed tract length for a given sampling rate. Conversion to a different tract length is not necessarily linear (Bladen, 1981; Fant, 1975). On the positive side, use of a shorter vocal tract (with the same number of sections) improves the ratio of wavelength to section length, increasing the usable bandwidth. This is fortunate as the resonant frequencies will be higher.

2.2.3 Source/system interaction

Speech synthesizers range from those in which the source and system are completely independent (such as LPC chips) to those in which the source is an intimate part of a large system that encompasses both laryngeal function and vocal tract resonances (Ishizaka and Flanagan, 1972).

Synthesizers with independent source and system are often referred to as source-filter models. These models are linear, whereas the interaction of supraglottal pressure with excitation waveform in the Ishizaka model is non-linear. A computationally more efficient method of modelling this nonlinear interaction is the combination of time and frequency domain representations used by Sondhi and Schroeter (1987). In practice certain synthesizers (Rye and Holmes, 1982) have made an approximation to the effect of temporal variation of source impedance by allowing resonance bandwidths to alter depending upon whether the source is deemed to be representing open phase (low impedance, high damping, wide resonance bandwidths) or closed phase (high impedance, low damping, narrow resonance bandwidths).

The impedance difference for open-phase vs closed phase is not a small effect. In ordinary speech the resonances are

really only active during the closed phase, damping abruptly when the glottis re-opens (Fant, 1978). Indeed the phonetician's trick of tapping on the neck to demonstrate excitation of the lowest resonance is only effective with the glottis closed.

2.2.4 Dynamics

Two problems will be considered: the fact that a moving vocal tract is a non-stationary system, and the requirement for interpolation of control data in a synthesis system.

2.2.4.1 Non-stationarity

A system whose characteristics change with time is difficult to describe. The usual specification of a system in terms of impulse response, frequency responce or system function is based on stationarity. A changing system does not have a single impulse response, spectrum and pole-zero plot. There is a formal analytical approach via the Wigner time-frequency distribution (Claasen and Mecklenbrauker, 1980), though it is not clear just how this description can be applied to speech synthesis.

The problem of stationarity is largely ignored, or quasi-stationarity is invoked. The practical consequence, however, is that a synthesizer simply may not produce what is expected from a specification based upon stationarity. Also the ordering of system elements in a series configuration will affect the output under transient conditions.

Some attempt has been made to quantify these effects. Jospa (1977) derives the increase in bandwidth as a function of the rate of frequency change of a resonance. With specific reference to speech synthesis, Maeda (1977) refers to a problem of temporal discontinuities caused by changing the parameters of a reflection coefficient synthesizer, and gives a solution based on two-dimensional reflections, which

preserve pressure and flow continuity. This approach is generalised in Strube (1982), in which continuity of longitudinal momentum is also considered, but he concludes that the effect of any discontinuities is rather slight, implying that the complexity of the solution may not be justified by the size of the problem. A further generalisation is made by Liljencrants (1985), in which eight distinct dynamic analogues are developed. Results for various tests tend to favour the Maeda formulation, as it produces smaller transient effects. Liljencrants does not test specifically speech-like motions, but uses stylised tasks in which marked inadequacies are demonstrated for the static analogue.

2.2.4.2 Interpolation

All synthesizers must update their parameters. Of interest in synthesis-by-rule is the derivation of a path from one target to the next. Also the effect of quantisation (step-size) along the path requires investigation. Finally, at least one widely used LSI synthesizer updates the transfer function one parameter at a time, introducing a time-skew to the interpolation (Brantingham, 1980).

Comparison of interpolation paths requires some common set of dimensions. There is a related problem of displaying a multidimensional effect. One method is to plot amplitude vs frequency vs time (as in a spectrogram) for transitions as produced by various synthesizers. A second method is to eliminate a dimension and plot just formant amplitude or frequency vs time. The formants tend to stay out of each other's way, allowing multiple parameters to occupy a single plot. Further, at least two interpolation methods can be compared on a single plot without undue clutter. This approach to the analysis of interpolation paths is used in Chapter Six.

A difficulty in formant interpolation is that the whole concept of an ordered set of formants with smooth

transitions does not necessarily accord with acoustic facts. With a uniform tube of varying cross section it is possible to hold one resonance steady while changing another resonance to actually 'cross' (in frequency) the first (Fujisaki, 1977).

Another difficulty is knowing what interpolation strategy to follow. Rabiner (1968b) postulated use of critical second-order damping for formant frequencies and amplitudes. A second degree equation was used to get exponential transitions; critical damping was used to control the transitions with a single parameter. It should be noted that there is nothing in the speech production apparatus that would necessitate these particular constraints on frequency and bandwidth motion. Observation based upon spectrographic analysis (Ohman, 1967) and articulatory measures (Sonoda, 1977) have been reduced to mathematical formulae. Witten (1982, p177) remarks: "The only thing that seems to be agreed is that the formant tracks should certainly not be piecewise linear. However, in the face of conflicting opinions as to whether exponentials should be decaying or increasing, piecewise linear motions seem to be a reasonable compromise! It is likely that the precise shape of formant tracks is unimportant so long as the gross features are imitated correctly. Nevertheless, this is a question which an articulatory model could help to answer."

The issue of interpolation and its implications for intelligibility and naturalness will be studied in Chapters Eight and Ten.

Chapter Three: Formal description and theoretical relationships

This chapter presents the formal specification of six types of speech synthesizer. The specification is in terms of a digital signal processing implementation. Formal equivalence is shown for four of the six types. The approximate relationship between these four models and the remaining two synthesizer types is presented in the next two chapters.

3.1 DIGITAL IMPLEMENTATIONS OF SYNTHESIZER MODELS

Two requirements must be met in order to implement a digital synthesizer:

1) specification of the synthesizer model in discrete form (section 3.1.1);

2) acquisition (from theory or measurement) of control data (section 3.1.2).

3.1.1 Specification of synthesizers

The types of synthesizer to be considered are:

1. Series resonance
2. Parallel resonance
3. Direct recursive form
4. Lattice form using reflection coefficients
5. Area function
6. Area function produced by articulatory parameters

The implementation of these synthesizers will be within the class of systems which are linear and time invariant. Such systems are completely described by their response to a unit

sample input (Rabiner & Schafer, 1978, p13-23).

Given an input x(n) and unit sample response h(n), the output y(n) is:

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) = x(n)*h(n) \qquad (3.1)$$

The z-transform of (3.1) is (Rabiner & Schafer, 1978, p13):

$$Y(z) = X(z)H(z) \qquad (3.2)$$

This can be simply rearranged to form the definition of the system function H(z):

$$H(z) = Y(z)/X(z) \qquad (3.3)$$

Speech synthesizers will be considered examples of digital filters. For all linear time-invariant systems of interest for implementation as filters, the system function will be rational: a ratio of polynomials.

The general formal description of a speech synthesiser is a rational system function, as in (3.4):

$$H(z) = \frac{(z^M + b_1 z^{M-1} + \ldots + b_M)}{(z^N + a_1 z^{N-1} + \ldots + a_N)} \qquad (3.4)$$

Here the numerator and denominator terms are written in the form of conventional polynomials. The polynomial coefficients a_j and b_i are real constants. A more compact representation is:

$$H(z) = \sum_{i=0}^{M} b_i z^{-i} \Big/ \sum_{j=0}^{N} a_j z^{-j} \qquad (3.5)$$

For a polynomial in standard form, the coefficient of the highest-order term is unity.

3.1.1.1 Series resonance

A polynomial with real, constant coefficients can be factored into linear or quadratic factors with real, constant coefficients (Jeffrey, 1979, p102). This is a corallary to a basic theorem that the roots of such polynomials are either real or occur in complex conjugate pairs.

Each quadratic denominator term is a <u>resonance</u>:

$$H(z) = G / (z^2 + a_1 z + a_2) \qquad (3.6)$$

A synthesizer which consists entirely of resonances can be expressed as a system function with only a constant in the numerator. The denominator is the only non-trivial polynomial, and it can be factored into linear and quadratic terms.

The factored system function for a denominator with no linear factors (no real roots) is then:

$$H(z) = G / \prod_{i=1}^{N} (z^2 - 2r_i \cos\theta_i z + r_i^2) \qquad (3.7)$$

$$\theta_i = 2\pi\tau C_i \qquad (3.8)$$

$$r_i = \mathrm{EXP}(-\pi\tau B_i) \qquad (3.9)$$

where: τ is the sampling interval;
C_i is the centre frequency of formant i (of N);
B_i is the bandwidth of formant i;
G is the gain;
EXP(x) is the exponential operation.

The simplest model for speech synthesis is an arrangement as in (3.7), because it is composed only of resonances, and has just a constant in the numerator. Roots of the denominator are called poles, as they cause the magnitute of the function H(z) to approach $+\infty$ as z approaches a root. When only the demominator has roots, H(z) is called an all-pole system.

The representation of H(z) as a product of resonances has an electrical equivalent in a serial (also called cascade) arrangement of circuits representing simple resonances.

The serial arrangement is the theoretically preferred arrangement for modelling speech, but only for those sounds consisting purely of resonances: vowels and vowel-like sounds /wrj/ (Fant, 1960). The remaining consonants introduce zeroes (roots of the numerator of the system function), either through constriction of the vocal tract to the point of turbulence or closure (fricatives and stops), or through multiple branching of the tract (nasals and the lateral approximant /l/). [The phonetic symbols and their categories are given at the end of the List of Symbols, page xii.]

Despite the fact that an all-pole model cannot properly represent the zeroes of consonants, the model is very widely used. The reason is the development of linear prediction (LPC; Markel and Gray, 1976), which provides an algorithm for determining the parameters of an all-pole model.

3.1.1.2 Parallel resonance

An alternative method of combining resonances is to use a parallel rather than a serial connection. This can be represented as the following system function:

$$H(z) = \sum_{i=1}^{N} [G_i/(z^2 - 2r_i\cos\theta_i z + r_i^2)] \qquad (3.10)$$

When the summation is carried out, the result is still a rational function, and so has the form of equation (3.5). However it will not in general have a constant for a numerator. It will instead have a polynomial of degree two less than that of the denominator, and so will have zeroes as well as poles:

$$H(z) = \frac{\sum_{i=1}^{N} G_i \prod_{j=1, j\neq i}^{N} (z^2 - 2r_j\cos\theta_j z + r_j^2)}{\prod_{i=1}^{N} (z^2 - 2r_i\cos\theta_i z + r_i^2)} \qquad (3.11)$$

There is one special case where a parallel system is the same as a serial one. This arises from consideration of the partial fraction expansion of a system function such as (3.5). In general any rational function can be represented as a sum of simple fractions, each with a constant in the numerator and a linear or quadratic denominator (raised to a power in the case of repeated roots) (Jeffrey, 1979, p338).

The partial fraction expansion can be said to show that a pole-zero parallel arrangement includes the all-pole possibilities as a proper subset. This argument is indisputably true as a statement of formal properties of polynomials. However as an argument for the parallel arrangement of a speech synthesizer, there is one difficulty: the parallel arrangement can only exactly model a serial arrangement if the numerator terms of the partial fraction expansion have specified values. As soon as these values are altered, the equivalence does not hold. In practical terms, these numerator terms represent the individual formant amplitudes. A principle reason for using a parallel arrangement is to have control of individual resonance gains. But the gain terms cannot be varied if the parallel form is to meet the partial fraction expansion conditions for equivalence with an all-pole form. Hence although the parallel arrangement formally includes the serial possibilities as a proper subset, in practical terms the parallel arrangement and serial arrangement will be different as soon as amplitude is varied.

Control of amplitude to minimise the disparity between the two forms of synthesis is one of the subjects treated in Chapter Four. That chapter will cover the general topic of approximate relationships between serial and parallel forms of resonance synthesizers.

3.1.1.3 Direct recursive form

The all-pole direct form is simply an unfactored system function:

$$H(z) = G / (z^{2N}+a_1z^{2N-1}+...+a_{2N-1}z+a_{2N}) \qquad (3.12)$$

This configuration is theoretically identical to the cascade arrangement, as it is only another way of implementing an all-pole system function. The implementation is as a single high-order filter rather than a cascade of second order stages. Although formally equivalent, in practice the signals generated by the direct form and the factored form may differ. In particular their dynamic behavior may be dissimilar, especially for a synthesis-by-rule system using control parameter interpolation to form the transition between one target value and the next. This difference in behavior extends to all the all-pole models to be considered, and is a question of research interest (cf Chapter 2, section 2.2.4, and Chapter 6).

Coefficients of the direct form filter are also known in the LPC literature as prediction coefficients, because the original linear prediction method yielded a direct form filter (Atal & Hanauer, 1971).

The direct form is also noteworthy in the digital signal processing literature for being a worst case arrangement for problems of instability, lack of noise immunity and sensitivity to quantisation effects (Kaiser, 1959; Witten, 1982, p136). A considerable part of the early development of the LPC approach concerned itself with avoiding or remedying the problems of the direct form, culminating in the use of reflection coefficients and lattice structures.

3.1.1.4 Lattice form with reflection coefficients

Another way to implement an all-pole system is with a lattice configuration using reflection coefficients. These filter

coefficients may be used to describe wave propogation in an acoustic tube, or the electrical analogue. The reflection coefficient can be defined in terms of wave properties, or obtained by conversion from area functions or from the direct form coefficients, as shown in section 3.2.1.

For plane wave propogation in a duct, the specific acoustic impedance is inversely proportional to the cross-sectional area S of the duct (Kinsler et al, 1982, pp124-131, 231-237). If adjoining sections of an N-section model of a vocal tract have specific acoustic impedances Z_i and Z_{i+1}, the pressure reflection coefficient k_{pi} at the boundary between the two sections (out of section i into section i+1) is (Makhoul, 1975, p566):

$$k_{pi} = \frac{Z_{i+1} - Z_i}{Z_{i+1} + Z_i} = \frac{S_i - S_{i+1}}{S_i + S_{i+1}} \qquad (3.13)$$

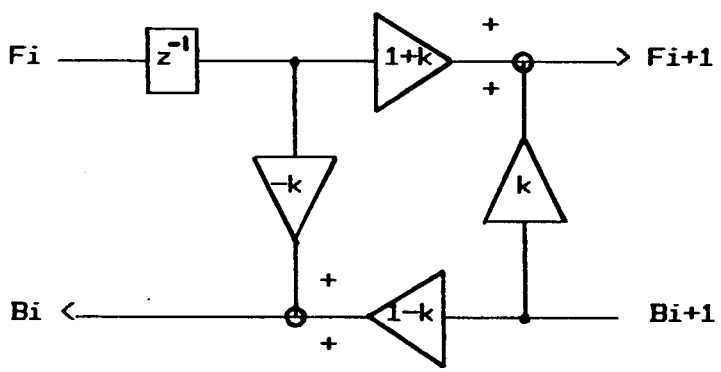
Alternatively, a volume velocity reflection coefficient may be defined:

$$k_{vi} = \frac{Z_i - Z_{i+1}}{Z_i + Z_{i+1}} = \frac{S_{i+1} - S_i}{S_{i+1} + S_i} \qquad (3.14)$$

The lattice of volume velocity reflection coefficients is in general use in linear prediction. Coefficients of magnitude less than one guarantee a stable filter; they can be reasonably compactly coded; they are used in a synthesis filter which has desirable noise and quantisation characteristics (Witten, 1978); finally, they can be directly determined from the autocorrelation function without proceeding via the predictor coefficients through Levinson's method (Makhoul, 1975).

The original use of reflection coefficients for speech synthesis was in a ladder arrangement using pressure reflection coefficients (Kelly and Lochbaum, 1962; Mermelstein, 1971). This configuration used four multiplies and two adds to implement the equations representing forward and backward wave propogation.

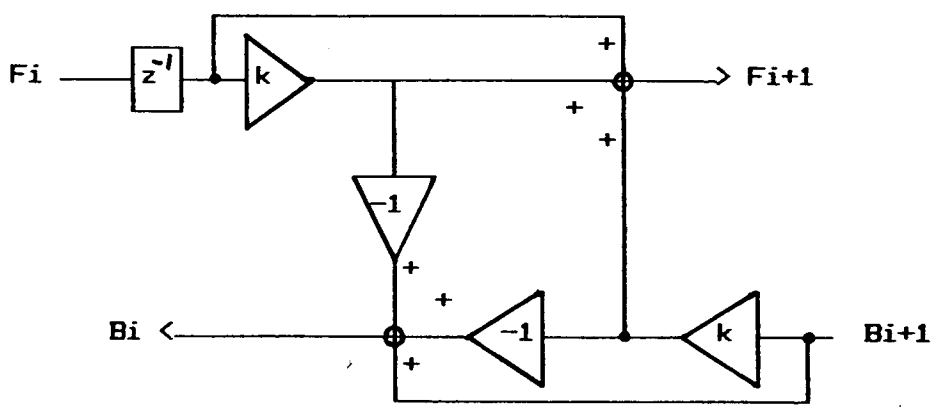
For an N-tube model of a vocal tract, let Fi be the forward volume velocity of section *i*, and Fi+1 be the forward volume velocity of the neighbouring section i+1. Let Bi and Bi+1 be similarly defined for volume velocity in the backward direction. Flow in the system may then be described according to the following diagram. In the following lattice diagrams the reflection coefficient k is a volume velocity reflection coefficient, to conform with current usage (Rabiner and Schafer, 1978; Witten, 1982). Also, only a single delay in the forward direction is used, as this was shown by Kelly and Lochbaum (1962) to be equivalent to two delays (each half as long) in both the forward and reverse directions.



$$Fi+1 = (1+k)\ Fi\ z^{-1} + k\ Bi+1 \qquad (3.15)$$

$$Bi = -k\ Fi\ z^{-1} + (1-k)\ Bi+1 \qquad (3.16)$$

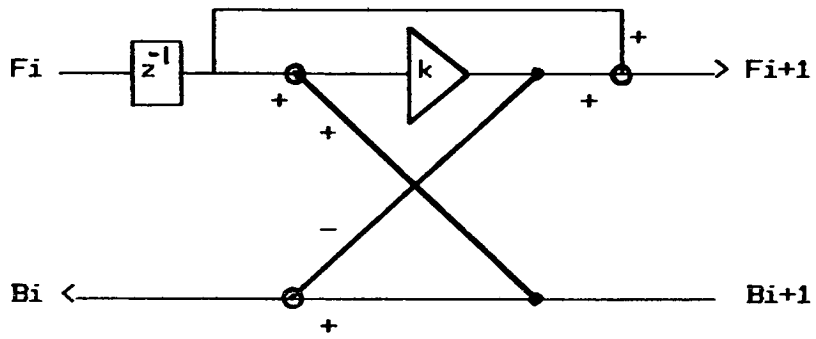
Alternatively, the equations can be arranged for two multiplies and four adds:



$$Fi+1 = Fi\ z^{-1} + k\ Fi\ z^{-1} + k\ Bi+1 \qquad (3.17)$$

$$Bi = -k\ Fi\ z^{-1} - k\ Bi+1 + Bi+1 \qquad (3.18)$$

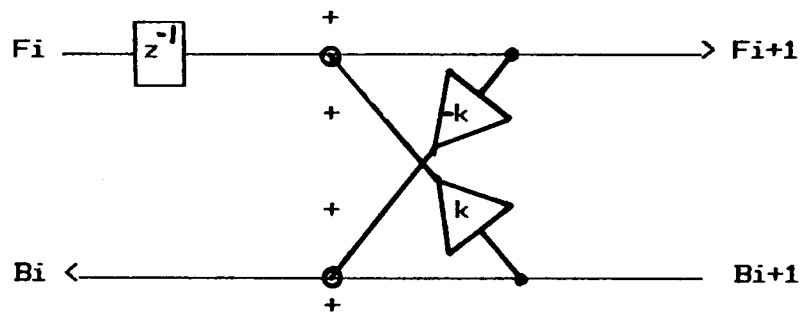
A further refinement (Itakura and Saito, 1971) used only one multiply and three additions by using a lattice (crossover) arrangement:



$$Fi+1 = Fi\ z^{-1} + k\ [Fi\ z^{-1} + Bi+1] \qquad (3.19)$$

$$Bi = Bi+1 - k\ [Fi\ z^{-1} + Bi+1] \qquad (3.20)$$

This configuration has been further refined (removing the stage gain of 1/(1+k) and compensating at the end for the product of all the stages by a single multiply) to yield the following 'standard' LPC lattice: (Rabiner and Schafer, pp 85-93; Witten, 1982, pp 138-144).



$$Fi+1 = Fi\ z^{-1} + k\ Bi+1 \qquad (3.21)$$

$$Bi = -k(Fi\ z^{-1} + k\ Bi+1) + Bi+1 = -k\ Fi+1 + Bi+1 \qquad (3.22)$$

All these configurations are formally equivalent. They use the same basic reflection coefficient, and are formally identical during interpolation. The arrangements with fewer multiplies have practical advantages: lower computational complexity, higher speed, and less effect from use of finite word length arithmetic.

The formulations given above all use a volume velocity reflection coefficient, for consistency. The original Kelly and Lochbaum (1962) simulation with a pressure reflection coefficient was identical to the four-multiplier ladder given above, but with sign changes on two of the four multiplies. Some texts (Witten, 1982, p137; Atal, 1985, p114) define reflection coefficients with the opposite sign from the definitions in (3.13, 3.14), causing a sign change on all four multiplies. We use the definition of Makhoul (1975, p566) and Rabiner and Schafer (1978, p84) because it agrees with the convention of general acoustics that as a plane wave approaches a rigid boundary, the reflected wave has equal pressure but opposite volume velocity.

The frequency response of a lattice filter can be determined by conversion to the direct form and evaluation on the unit circle in the z-plane (cf section 3.2.1 of this chapter).

3.1.1.5 Area function

A. Varieties of area function.

The direct link between terminal and line analogues of the vocal tract is by the interpretation of reflection coefficients (which could be considered as a lattice form of terminal analogue) as reflections at the boundaries of an N-section tract.

Areas, area ratios, and log areas have all been used as LPC data (Viswanathan and Makhoul, 1975). It is common to use (and interpolate) reflection coefficients, but the others have been considered (Rabiner and Schafer, 1978, pp 441-444). Their properties under interpolation have not all been studied in detail, however.

Implementation can simply be performed via the reflection coefficient structure considered in the previous section. Thus parameter specification can be in terms of any type of area parameters, which are then converted to reflection coefficients for actual synthesis. (See section 3.2.1 for conversion details.)

Given an area function for a lossless tube as shown above, losses may then be introduced as lumped effects in each section (Flanagan, 1972). Also, between any two sections a non-linear effect or a source may be introduced. Such a model becomes very like the original line analogues such as Steven et al (1953), as discussed in Chapter 1, section 1.1. But as soon as losses within the tube are introduced, the conversion between area function and direct form (cf section 3.2.1) is no longer possible. There is a distinction between the area function of a lossless tube, which is formally identical with terminal analogue models, and the area function of a lossy tube, which is a line analogue. The key difference between an area function derived from linear prediction (the pseudo-area function; Wakita, 1973) and an area function actually representing articulation is the question of modelling of losses within the vocal tract.

B. Articulatory models

An N-section area function description, like a discrete spectrum, is a representation of high dimensionality and considerable redundancy (correlation). Yet a speech

spectrum can be determined from a small number of resonances (and perhaps antiresonances). Similarly the area function can be determined from a small number of control functions. These may be derived from anatomical considerations (tongue shapes or musculature; Coker, 1968) or statistical analyses (Ladefoged & Harshman, 1979).

Articulatory parameters are important because they constrain area functions. They thus limit the range of allowed shapes, and smooth the transitions from one state to another (Haggard, 1979). It is of interest to examine transition paths subject to articulatory constraints, and to study the effects of such constraints upon naturalness and intelligibility.

Stevens and House (1955) present a three parameter model which has been extensively used (Atal et al, 1978). Their system determines 35 vocal tract areas from three control values: area at point of maximum constriction, distance from glottis to that point, and lip opening. A problem with this model is that it is 'one-way': area functions can be determined from articulatory parameters, but not the reverse. This fact makes it difficult to determine control data.

An articulatory model without this limitation is that of Ladefoged, et al (1978). Regression analysis was used to determine equations to produce articulatory parameters from formant data, though only for vowels. This system provides a 'two-way' articulatory model, and hence improved scope for adding articulatory constraints to synthesis from area function data. The extension of the Ladefoged et al method to a more general phoneme set will be the topic of Chapter Five. The intelligibility of the resultant synthetic speech was determined in Experiment IV, Chapter Nine. Finally, the naturalness of the speech is studied in Experiment V, Chapter Ten.

3.1.2 Acquisition of Control Data

A speech synthesizer is a model for the generation of a speech-like signal. To attempt to produce the sounds of speech, and their temporal patterns, the parameters of the model must be set to particular values. Acoustic data can be produced by analysis of the speech signal, and vocal tract data can be obtained from physiological measurements, including x-ray studies.

The earliest published set of data for the generation of synthetic speech from a phonetic description is that of Holmes et al (1964), which consists of resonant frequencies and amplitudes, as well as speech sound durations and transition control parameters. The data are tabulated for a phonemic inventory suitable for Southern British English. A small number of allophones are also included: clear vs dark /l/; aspirated and unaspirated voiceless stops. These data are based on the JSRU data for operation of a parallel resonance synthesizer. Ainsworth (1974) gives an alternative set of parallel resonance data. Parameter values for a combined series and parallel resonance synthesizer are given in Klatt (1980).

Supplementary data in the phonetics literature provides some indication of bandwidths, at least for vowel resonances (Peterson & Barney, 1953; Dunn, 1960). Considerable duration data also exists (Umeda, 1975; also numerous papers in Lehiste, 1967). Most papers on synthesis-by-rule do not publish the actual tabular data (with the exceptions of those mentioned above). The two other sources of data are direct measurement and trial-and-error. Both these approaches are very time consuming. The trial-and-error method also suffers from 'accomodation', an effect in which repeated listening destroys critical judgement.

A classic debate in speech synthesis is whether inadequacies in the final output are to be blamed on the synthesizer or the control parameters (Holmes, 1979). In the case of LPC analysis-synthesis, an extremely simple synthesizer

produces speech of high intelligibility and naturalness. This result is evidence for the paramount importance of the control data.

3.2 RELATIONS BETWEEN MODELS

Four of the six synthesizers studied are all-pole models for which an analytic conversion relationship exists, namely:

series resonance
direct form (prediction coefficients)
reflection coefficients
area function

The two remaining synthesizer types involve approximations. Combining resonances in parallel introduces a numerator polynomial into the system function, meaning there will be zeroes as well as poles, which increases the number of parameters. In the case of a simple articulatory model, there is a reduction in the number of parameters, from ten down to three or four. Thus neither the parallel nor the articulatory model can in general have an exactly equivalent series resonance formulation.

The analytical relations for conversion between the four all-pole models are given as Pascal subroutines in Appendix One, along with routines for the inexact conversion from series to parallel and between area function and articulatory parameters. The listings include specification of all conventions used, and the routines are mutually consistent.

3.2.1 Analytic relations for all-pole models

The equations for the formal equivalence of all the all-pole models begin with the serial resonance formulation. The equivalence is two-way: from a set of data for a series resonance synthesiser, the equivalent data for the other

three types can be produced, and vice-versa. These relations are used extensively in the experimental work of chapters six to ten. The forward direction allows all the four all-pole synthesisers to be driven by data originally published for serial resonance synthesis. The reverse direction allows the result to be converted back to formant centre frequencies and bandwidths, so that there is a common comparison space.

A. Serial resonance and direct recursive form

Given a series synthesizer in factored form, the direct recursive form is simply the product of the second order polynomials of the series (cascade) form (as in equations 3.7 to 3.9):

$$H(z) = G \Big/ \prod_{i=1}^{N} (z^2 - 2r_i\cos\theta_i z + r_i^2) \qquad (3.23)$$

$$= G \Big/ (z^{2N} + a_1 z^{2N-1} + \ldots + a_{2N-1} z + a_{2N}) \qquad (3.24)$$

One final detail is the convention for the coefficients of the denominator of the original system function. This thesis uses the convention of general mathematics, namely that the polynomial coefficients are all added, though they may have a negative value. A sort of standard in signal processing and LPC work is to have a minus sign on all but the first (unity) coefficient (Rabiner and Schafer, 1978, p19).

Conversion is also needed from the direct form to a series arrangement, to get back to the common reference representation of formant frequencies and bandwidths. This conversion requires factoring the denominator polynomial into individual resonances, individual second order sections. A simple Newton-Raphson root-finding algorithm has been completely successful for the polynomials encountered in our data, even without initial estimates of root location.

B. Direct recursive form and reflection coefficients

In the case of a tube with losses only at one end, and no losses along the tube (hardwalled or lossless tube model), the propogation from the lossless end of the tube to the lossy end is described by a product of N 2x2 matrices, which can be shown to be equivalent to the following simple recursion (Makhoul, 1975; Markel & Gray, 1976):

Let: a_i be the prediction coefficients (direct form polynomial coefficients);
k_i be the refection coefficients;
subscripts denote coefficients;
superscripts denote iterations;
i from N down to 1; $1 \leq j \leq i-1$;

Then: $k_i = a_i^{(i)}$ (3.25)

$$a_j^{(i-1)} = (a_j^{(i)} - k_i a_{i-j}^{(i)}) / (1 - k_i^2) \quad (3.26)$$

Reflection coefficients can be converted to direct form coefficients through the reverse iteration, with i from 1 to N:

$$a_i^{(i)} = k_i \quad (3.27)$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \quad (3.28)$$

C. Reflection coefficients and area functions

The reflection coefficient is an impedance ratio, and for simple hardwalled tubes this can be taken to be determined by the ratio of areas at the boundary between two cylindrical sections (Makhoul, 1975):

Let: k_i be the reflection coefficients;
S_i be the i^{th} cross-sectional area;

Then: $k_i = (S_{i+1} - S_i) / (S_{i+1} + S_i)$ (3.29)

The relation can be turned around to compute the area ratios from the reflection coefficients:

Let: R_i be the area ratio;

Then: $R_i = S_i/S_{i+1} = (1-k_i)/(1+k_i)$ (3.30)

Finally, given a starting point, the areas themselves are determined:

$$S_i = R_i \cdot S_{i+1} \quad (3.31)$$

or: $S_{i+1} = S_i \cdot (1+k_i)/(1-k_i)$ (3.32)

A standard starting point is $S_i = 1$.

3.2.2 Approximate relationships for other models

The problem of relating the series and parallel resonance models is the topic of Chapter Four. The problem of converting data for any other synthesizer into appropriate values of articulatory parameters is discussed in Chapter Five.

Chapter Four: Parallel resonance parameters

4.1 FORMAL EQUIVALENCE OF SERIAL AND PARALLEL ARRANGEMENT

This chapter presents a theoretical and practical comparison of the two types of resonance synthesizer, serial and parallel. It is shown that the serial arrangement cannot be exactly matched by a parallel configuration for practical synthesis. Gain control and compensation is discussed for five different forms of a resonance. Finally the method used in this research to convert from series to parallel parameters is presented.

Chapter Three presented the discrete system function representation of digital filters, of which speech synthesizers may be considered a special case. As discussed in that chapter, any rational function may be represented as a sum of simple system functions by the method of partial fractions expansion. The terms in the expansion will be rational functions with linear or quadratic denominators, and numerators of degree one less than the denominators. For repeated roots of the original denominator polynomial, the linear or quadratic roots will be raised to the appropriate power.

For speech synthesis, a term with a quadratic denominator (and a constant or linear numerator) is a resonance. The parallel arrangement is, in general terms, the addition of resonances.

Typically in modelling of speech for synthesis one begins not with the complete system function, but with a few resonances. A simple resonance (as shown in section 3.1.1.1, equation 3.7) is:

$$H(z) = G / (z^2 - 2r \cos\theta\, z + r^2) \qquad (4.1)$$

$\theta = 2\pi\tau C$; C = centre frequency of the resonance;
$r = \mathrm{EXP}(-\pi\tau B)$; B = bandwidth of the resonance;
τ = sampling interval; G = gain (a real constant).

These resonances can then in engineering terms be connected in series or in parallel. The system function for the serial arrangement is the product of the individual terms, which still just has a constant in the numerator and so has no zeroes. Thus it is an all-pole system.

The parallel connection is the sum of the individual system functions, which will in general have a numerator polynomial of degree two less than the denominator polynomial, and so will have zeroes as well as poles. However a special case of the parallel connection arises when the numerator terms (gains) for the individual resonances are exactly the requisite coefficients of a partial fraction expansion of a serial arrangement system function. This is then a very special system: a parallel connection without zeroes.

4.1.1 Parallel Case Without Zeroes

In the discussions of serial vs parallel resonance synthesizers, one could argue that the parallel is the obvious choice: since any serial system function can be turned into an equivalent parallel arrangement, the parallel synthesizer has the most generality and the serial synthesizer it represents is only a proper subset.

The exact relation provided by the partial fraction expansion is of little practical value, because it only holds for fixed values of the individual resonance gains. Fixed gains are only appropriate for individual phoneme targets, to match a steady-state spectrum. Thus Weibel (1955) shows good matches to measured natural speech vowel spectra using a parallel connection of resonances. For generation of speech, rather than just matching steady-state spectra, the system must change from one target position to the next. This change will involve alteration of the gains, and the appearance of zeroes.

There is thus a very strong limitation on the practical use of the all-pole parallel arrangement to exactly model the

behavior of a serial arrangement via the partial fractions expansion: it only holds for steady-state values, not for continuous speech. In practice the system functions for the series connection are not a subset of the possibilities of the parallel case; rather they are disjoint: series synthesizers are all-pole systems, parallel ones have additional zeroes.

4.1.2 Parallel Case With Zeroes

The general case for a parallel arrangement will have zeroes, the roots of the numerator polynomial.

The system function for a parallel arrangement of a set of simple resonances such as (1) has a denominator polynomial of degree 2N, where N is the number of resonances. The numerator polynomial is of degree 2N-2, and has the following general formula:

$$N(z) = \sum_{i=1}^{N} G_i \prod_{j=1, j\neq i}^{N} (z^2 - 2r_j \cos\theta_j z + r_j^2) \qquad (4.2)$$

For the simple addition of resonance with equal sign, as in equation (3.10), the zeroes will fall between the poles (Rabiner, 1968, p827; Flanagan 1957, p309). This effect can be most easily seen in the case of the addition of two resonances with equal amplitudes:

$$\frac{1}{(z^2+B1z+C1)} + \frac{1}{(z^2+B2z+C2)} = \frac{2z^2+(B1+B2)z+(C1+C2)}{(z^2+B1z+C1)(z^2+B2z+C2)} \qquad (4.3)$$

Roots of the final numerator polynomial are roots of the following normalised equation:

$$z^2 + [(B1+B2)/2]z + (C1+C2)/2 = 0 \qquad (4.4)$$

Equation (4.4) has polynomial coefficients between those of the two denominator polynomials in (4.3), and simple application of the quadratic formula gives roots r at:

$$r = \{-(B1+B2)/2 \pm [\ (B1+B2)^2/4 - 2(C1+C2)\ ]^{1/2}\}/2 \qquad (4.5)$$

The case of main interest is complex roots, as otherwise the expressions do not meet the conventional definition of a resonance. If both addend denominators in (4.3) have complex roots, then $B1^2<4C1$, $B2^2<4C2$, and the expression within square brackets in (4.5) is also negative with a value between $(B1^2-4C1)$ and $(B2^2-4C2)$. Thus the zeroes are between the poles. For the more general case of varying gain of the original resonances, the zero location is simply pushed toward the poles associated with the larger gain.

Another problem associated with summing resonances is the question of phase. The simple addition of two resonances with equal sign is also an in-phase connection: there is no phase difference between the inputs of any of the resonances. In a serial connection the opposite is true: a second-order system produces a 180° phase difference between the response on either side of the resonant frequency, and so a high frequency component becomes phase-reversed as it passes through a lower frequency resonance. The nearest parallel equivalent is to reverse the sign of alternate terms in the summation, forming the 'alternate phase' connection.

The alternation of phases in a parallel resonance synthesizer not only improves the model of phase behaviour, it also releases the zeroes from their constraint of being interleaved:

$$\frac{D}{(z^2+B1z+C1)} - \frac{E}{(z^2+B2z+C2)} = \frac{(D-E)z^2+(DB2-EB1)z+(DC2-EC1)}{(z^2+B1z+C1)(z^2+B2z+C2)} \qquad (4.6)$$

The quadratic coefficients are no longer constrained to values between poles, and indeed the minus signs in the numerator coefficients of the right hand side of (4.6)

allow unconstrained positioning of the zeroes as the relative gains D and E are altered.

The importance of using the alternate phase connection is stressed by both Klatt (1980, p984) and Holmes (1982). An early mention is the paper by Weibel (1955), in which the partial fraction expansion interpretation of the parallel configuration is described, but with the constraint of alternating signs on the coefficients. The Weibel paper (also discussed in Lynggard, 1985, p73) does not stress the limitation of the method to the steady state.

Another omission in the literature concerns the statement in Flanagan (1957, p309), Rabiner (1968, p827) and Klatt (1980, p982) that the zeroes for the parallel case fall between the poles. None of the authors mentions that this is only true for the case of simple in-phase addition of resonances.

A main effect of interleaved zeroes is to make deep notches in the spectrum between formant peaks. Several examples are given in Klatt (1980,p984) for vowel sounds for a serial and both sorts of parallel configuration. However the antiphase connection does not eliminate all such problems.
Experimental data from parallel synthesis (using alternating signs) of eight nonsense words are given in Chapter 6, including examples of the effects of zeroes on the region below the first formant. (Cf Chapter 6, Section 6.3.2 and Figure 6.7.)

4.2 AMPLITUDE CONTROL

In synthesis using resonances connected in series there is no explicit setting of resonance amplitudes. The amplitude at a given resonant frequency is controlled both by the bandwidth of the resonance in question, and by the frequency response at that same frequency for all the other resonances in the system. Thus as formant frequencies are moved about, the amplitudes go up and down as any particular resonance 'rides the skirts' of the remaining terms in the series.

One of the stated advantages of a parallel connection is the ability to set individual formants to particular amplitudes. This is especially useful if the basic data for the synthesis comes from spectrographic measurements, in which amplitudes are much more apparent than are bandwidths.

There is a difficulty in actually setting gain. It is not sufficient to simply use the gain factor G of the simple resonance of equation (4.1). The gain at resonance would then be:

$$\left|H(z)\right|_{z=EXP(j\theta)} = G/\left|z^2-2r\cos\theta\, z + r^2\right|_{z=EXP(j\theta)} \qquad (4.7)$$

which is a function of centre frequency and bandwidth as well as the factor G. If the amplitude of the final acoustic output at a resonant frequency is to be determined solely by G, the dependence upon frequency and bandwidth needs to be eliminated.

One solution is a normalised resonance. However there are at least two possibilities:

(1) A frequency domain criterion: use a resonance whose gain at the resonant frequency is unity.

(2) A time domain criterion, as used in the JSRU synthesis (Holmes, 1982): use a resonance whose impulse response is normed, so that waveforms out of the resonance will have uniform size regardless of resonant frequency and bandwidth.

The approach of the Klatt synthesizer (1980) does not normalise the gain, but follows the following strategy:

(3) Set unity gain at DC for the basic resonance; modify by a fixed set of amplitude corrections based on achieving equal amplitude peaks for a neutral vocal tract; modify further according to the distance between resonant frequencies.

These three approaches will be presented in the next three subsections, and compared in the following subsection.

4.2.1 Frequency Domain Normalisation

For cascade synthesis, it is sensible to have a gain term G in the system function such that gain at DC (z=1) is unity. This allows any number of resonances to be cascaded without introducing any shift in response at DC.

Unity gain at DC requires the gain term G in equation (4.1) to take the value:

$$G = \left| z^2 - 2r\cos\theta\, z + r^2 \right| \Big|_{z=1} = 1 - 2r\cos\theta + r^2 \qquad (4.8)$$

For parallel synthesis, one method of normalising the gain of a resonance before applying an amplitude control factor is to require unity gain at resonance. This in turn means that the system function needs to be evaluated at z=EXP(jθ), and have a G such that the magnitude is unity:

$$\left|H(z)\right| \Big|_{z=EXP(j\theta)} = 1 \qquad (4.9)$$

Then:

$$G = \left|EXP(2j\theta) - 2r\cos\theta\, EXP(j\theta) + r^2\right| \qquad (4.10)$$

4.2.2 Time Domain Normalisation

Rather than starting with a frequency domain system function representation for synthesis, Rye and Holmes (1982) begin with a (continuous) time-domain description in terms of addition of impulse responses:

$$f(t) = \sum_{i=1}^{N} A_i \, EXP(\sigma_i t) \, \sin(w_i t) \qquad (4.11)$$

where: $\sigma_i = -\pi B_i$
$w_i = 2\pi C_i$
B_i = bandwidth of resonance i
C_i = centre frequency of resonance i
A_i = amplitude of resonance i
N = number of resonances

This equation is simply a statement that the synthetic waveform f(t) will be a set of damped exponentials added with amplitudes A_i. Then to ensure that the term A_i is the only factor controlling gain, the implementation of the resonances must be such that they produce exponentials of unit amplitude:

$$h(t) = EXP(\sigma t) \, \sin(wt) \qquad (4.12)$$

Equation (4.12) is a formal statement of the requirement in Rye and Holmes (p11) that formants "keep the initial height of the envelope of their impulse responses, EO, independent of formant frequency and bandwidth".

The time domain requirement in equation (4.12) can be converted to an equivalent in the frequency domain. First expressing (4.12) for a sampled signal:

$$h(n\tau) = EXP(\sigma n\tau) \, \sin(nw\tau) \qquad (4.13)$$

where: σ,w are as in (4.11)
τ = sampling interval
n = sample number
$h(n\tau)$ = output sequence

This can be converted by a Z-transform to:

$$H(z) = z \sin\theta \, / \, (z^2 - 2r \cos\theta \, z + r^2) \qquad (4.14)$$

where r and θ are as in equation (4.1).

The z in the numerator can be neglected, as removal of any numerator delay terms simply amounts to a redefinition of the point of time origin in the output stream. Removing the numerator delay leaves just a constant gain term:

$$H(z) = \sin\theta \,/\, (z^2 - 2r\cos\theta\, z + r^2) \qquad (4.15)$$

A comparison of equations (4.1), (4.8), (4.10), and (4.15), and of the actual gain normalisation method used in the JSRU synthesizer, will be made in subsection 4.2.4.

4.2.3 Series connection equivalence (Klatt method)

The Klatt (1980) method of speech synthesis has a serial as well as a parallel set of resonances. A feature of his hybrid synthesizer is that the serial path may be replaced by appropriate settings for the parallel path. Thus the primary consideration for Klatt is equivalence of the parallel and serial connections.

The method is to establish parallel gain adjustment factors such that equal gain parameters (60dB) applied to the five parallel resonances (arranged to produce the neutral vowel schwa) will give the same formant amplitudes as in the serial configuration. This method does not reduce to equations, and in fact is implemented simply from a table of correction factors. These factors modify the gain of resonances which are normed for unity gain at DC, so that identical recursion parameters can be used in both the serial and parallel sections of the synthesizer.

Klatt goes on to modify amplitudes for the case when two resonant frequencies are within 550 Hz of each other. The amplitude is increased by 1dB for every 50 Hz decrement in spacing between formant frequencies, to a maximum of 10 dB.

One might suspect that a single correction factor per formant, established for a single configuration (the neutral vowel), would not correctly compensate for gain changes across the range of centre frequency and bandwidth variation. Indeed this is exactly the case, but Klatt simply adjusts the gain parameters, phoneme by phoneme, to achieve the desired response. Thus the tabulated gain parameters (as in Klatt, 1980, Table III, p987) achieve the desired acoustic effects, but only in a Klatt synthesizer! It is very difficult in the Klatt scheme to separate the factor of gain variation produced by centre frequency and bandwidth changes (inherent gain) from the factor of parametric control of resonance amplitude (external gain).

4.2.4 Comparison of gain adjustment criteria

A simple resonance has the form of equation (4.1). The numerator term G (for gain) can be adjusted to compensate for the fact that response at resonance varies according to centre frequency and bandwidth. Four different gain terms have so far been mentioned:

a) Uncompensated: $G = 1$ equation (4.1)

b) Unit gain at DC: $G = 1 - 2r\cos\theta + r^2$ equation (4.8)

c) Unit gain at resonance: equation (4.10)
$$G = |EXP(2j\theta) - 2r\cos\theta\ EXP(j\theta) + r^2|$$

d) Unit amplitude impulse response: equation (4.15)
$$G = \sin\theta$$

These four gain terms have in common the fact that they are all real constants. Implication for gain setting will be considered in the following section. The actual JSRU method involves a complex expression in the numerator, which will be separately considered in section 4.2.4.2.

4.2.4.1 Simple resonance

It is clear by inspection that an uncompensated resonance (equation 4.1) will differ significantly from the unit amplitude impulse response (equation 4.15). The uncompensated resonance will have the same gain as (4.15) for $\theta=\pi/2$, but will be more than 3dB down at the low and high ends of the frequency range, $0 < \theta < \pi/4$ and $3\pi/4 < \theta < \pi$.

The unit gain at DC (equation 4.8), is approximately $2(1-\cos\theta)$, because r would be about 0.95 for speech resonances. This gain term differs from the unit amplitude impluse response by the fixed factor 2, but $(1-\cos\theta)$ is similar to $\sin\theta$ in the first quadrant, though very different in the second quadrant. For a 20 kHz sampling frequency and speech resonances only below 5 kHz (as used in the JSRU research), these two criteria differ most at $\pi/4$=2500Hz, by a factor of about 7dB.

The difficult term to approach analytically is the expression for unity gain at resonance (equation 4.10). Expanding the operation for magnitude using $|X| = [X X^*]^{1/2}$:

$$G=[1-4r\cos^2\theta+4r^2\cos^2\theta+2r^2\cos 2\theta-4r^3\cos^2\theta+r^4]^{1/2} \qquad (4.16)$$

Using the half-angle formula $\cos 2\theta=\cos^2\theta-\sin^2\theta$:

$$G=[1-4r\cos^2\theta+6r^2\cos^2\theta-4r^3\cos^2\theta+r^4-2r^2\sin^2\theta]^{1/2} \qquad (4.17)$$

Use $(1-r)^4 = 1 - 4r + 6r^2 - 4r^3 + r^4$

$$G = [(1-r)^4\cos^2\theta + \sin^2\theta(1-r)^2(1+r)^2]^{1/2} \qquad (4.18)$$

Which reduces to:

$$G = (1-r)\,[(1-r)^2\cos^2\theta + (1+r)^2\sin^2\theta]^{1/2} \qquad (4.19)$$

For r approaching 1.0, (4.19) approximates to:

$$G \approx (1-r)[(1+r)^2\sin^2\theta]^{1/2} = (1-r^2)\sin\theta \qquad (4.20)$$

Thus the frequency domain criterion is approximately the same as the time domain criterion, for r near 1.0 and for θ not near 0 or π, except for the constant factor $1-r^2$; r is related only to bandwidth, which for conventional parallel synthesizers is a constant. Numerical evaluation of formulae (4.15) and (4.10) have shown that for r as low as 0.8, the two formulae are within 3 dB for $0.15 < \theta < 3.0$; for a 10 kHz implementation this would be the range of about 225-4775 Hz. For 20 kHz sampling the two criteria would be equivalent down to nearly 100 Hz. The conclusion is that for the normal range of speech formants, the frequency domain and time domain criteria are equivalent for practical purposes.

4.2.4.2 Resonance plus a zero at DC

Both the Klatt (1980) and JSRU (Holmes, 1973) synthesizers add a zero at DC to the basic specification of a resonance. This zero is added to the upper formants only, mainly to prevent more than one formant contributing to response at DC. This zero is referred to in the JSRU scheme as a spectrum weighting filter.

The system function is no longer as in (4.1), but becomes:

$$H(z) = (z-Q) / (z^2 - 2r \cos\theta\, z + r^2) \qquad (4.21)$$

To complicate the situation in the JSRU synthesis scheme, the first formant and the nasal formant do not follow the time domain criterion. Rather they have fixed gain at DC, at a value chosen to give a unit amplitude impulse responses at (and only at) a resonant frequency of 500 Hz. The justification is given in Holmes (1982), in which he discusses the importance of not having changes in the level of the signal in the frequency region below the first formant. If the DC response moves, the signal can have a variable 'booming' quality which is unnatural and disruptive. Thus Holmes adheres to fixed gain at DC as a primary requirement, and uses the unit impulse response only

for the higher formants (which in the JSRU synthesizer have no output at DC).

Rye and Holmes give the effect of the (z-Q) term in (4.21) on the amplitude term A (as in equation 4.11) of the impulse response as:

$$A = SQRT(r^2+Q^2-2rQ\cos\theta) \;/\; (r\sin\theta) \qquad (4.22)$$

They go on to state that for their usual value of Q, A is very nearly unity, so they dispense with a normalisation factor.

Equation 4.22 can be derived from convolution of the impulse presponse of the resonance (4.14) and the impulse response of the zero:

Given: $H_1(z) = z \;/\; (z^2 - 2r\cos\theta\, z + r^2)$, a resonance;

and $H_2(z) = (1-Qz^{-1})$, a zero at Q;

Then: $h_1(nT) = EXP(\sigma nT)\sin(wnT) \;/\; \sin\theta \qquad (4.23)$

$$h_2(nT) = \delta(n)T - Q\delta(n-1)T \qquad (4.24)$$

where $\delta(n)$ is the delta function. The fact that the zero only contributes delta functions to the impulse response makes direct evaluation of a convolution possible. Equation (4.23) follows from (4.13) and (4.14), and (4.24) is just the impulse response of $H_2(z)$. The total response is the convolution:

$$h(nT) = h_1(nT) * h_2(nT) \qquad (4.25)$$

$$= \frac{1}{\sin\theta} \sum_{k=-\infty}^{\infty} [\{(EXP(\sigma(n-k)T)\sin(w(n-k)T)\}\{\delta(k)T - Q\delta(k-1)T\}] \qquad (4.26)$$

$$= (1/\sin\theta)\,[\,EXP(\sigma nT)\sin(nwT) - Q\,EXP(\sigma nT)\,EXP(-\sigma T)\sin(nwT-wT)\,] \qquad (4.27)$$

Using the trigonometric identity $\sin(a+b) = \sin(a)\cos(b) - \cos(a)\sin(b)$ on $\sin(n\omega_T - \omega_T)$, and using the Euler formula on $\cos(-\omega_T)$ and $\sin(-\omega_T)$ yields:

$$h(n_T) = (1/\sin\theta)\ [\ \ EXP(\sigma n_T)\ \sin(n\omega_T) - Q\ EXP(\sigma n_T)EXP(-\sigma_T)EXP(-j\omega_T)\sin(n\omega_T)] \qquad (4.28)$$

$$= (EXP(\sigma n_T)/\sin\theta)\ [1 - Q\ EXP(-\sigma_T)EXP(-j\omega_T)]\ \sin(n\omega_T) \qquad (4.29)$$

Using $r = EXP(\sigma_T)$ and $\theta = \omega_T$, as in (4.1) and (4.11):

$$h(n_T) = [(1-Qr^{-1}\ EXP(-j\theta))/\sin\theta]\ EXP(\sigma n_T)\ \sin(n\omega_T) \qquad (4.30)$$

The expression in brackets in (4.30) is the gain term, and only the magnitude is relevant:

$$mag[(1-Qr^{-1}\ EXP(-j\theta))/\sin\theta] = SQRT\ [1 + Q^2r^{-2} - 2Qr^{-1}\cos(\theta)]\ /\sin\theta \qquad (4.31)$$

$$= SQRT\ [r^2 + Q^2 - 2Qr\cos(\theta)]\ /\ r\ \sin\theta \qquad (4.32)$$

This is result (4.22), the amplitude of a resonance with a zero. Rye and Holmes present this result, and go on to state that this amplitude term " is very nearly unity". Numerical evaluation of (4.22) has given values within 3dB of unity, for $0.9<r<1.0$ and $0<\theta<\pi/2$. But beyond $\pi/2$ they diverge, reaching 9dB difference at $3\pi/2$ and heading for infinity at π. This point is significant, because while the original JSRU synthesis used a 20 kHz sampling rate and had no speech resonances in the second quadrant, the hardware implementation commercially available (Quarmby and Holmes, 1984) runs at 10 kHz.

To summarise, Rye and Holmes specify a time domain gain normalisation criterion (4.12), and present (4.22) as the appropriate gain normalisation factor for a resonance plus a zero at DC. They do not apply the normalisation to the first formant and the nasal formant because they consider it more important to maintain constant DC response for these two resonances. Neither do they apply it to the higher

resonances, because their unnormalised response (for a resonance which includes a zero at DC) is close to a unit impulse response, for their usual parameter values, and providing the sampling rate for synthesis is a least four times higher than the highest resonant frequency.

4.2.5 Conversion of Klatt serial resonance data

Parallel vs serial relationships are more than an academic point. The experiments reported in Chapters Six to Ten require the use of a common phonemic inventory to drive a multiplicity of synthesizer types, including both serial and parallel resonance configurations. This chapter concludes with a brief description of the method for generating parallel resonance data from series data given in Klatt (1980).

An exact conversion from series form to parallel resonance form is possible through partial fraction expansion, but only for one particular set of amplitudes and bandwidths as discussed above. The usual form of parallel resonance synthesizer in practical use has fixed bandwidths, and only amplitudes are variable. In this case a parallel synthesizer can be made approximately equivalent to a serial form by simply using the resonance amplitudes of the series form.

Amplitudes at resonance for the Klatt data were computed by evaluating the magnitude of the serial resonance system function (cf section 3.1.1.1, equation (3.7) with gain G=1) at the formant centre frequencies. The bandwidths used (in order from F1 to F5) were: 50, 150, 250, 350 and 450 Hz. The resultant amplitudes are given in Table 6.1 of Chapter 6.

Synthesis proceeded by applying these amplitudes as gain terms in a parallel resonance configuration with alternating gains. As discussed throughout section 2 of this chapter, the question is: what is a resonance, and what is its gain? In our research the simplest form of a resonance was used,

equation (4.1), without a zero. The unity gain at resonance compensation was applied, equation (4.10).

The results of synthesis with this 'parallel driven by serial parameters' approach are presented in Chapters Six, Seven and Ten.

Chapter Five: Determination of articulatory parameter values

5.1 INTRODUCTION

This chapter presents a method for deriving control data for an articulatory synthesizer from formant data, and gives the result in Table 5.3. The object of this thesis research is an examination of synthesizer parameters. A representative subset of the possible all-pole parameters sets has already been discussed: series resonance, direct form, reflection coefficients and area function. There is no problem using a common set of phoneme data for synthesis with these all-pole descriptions, because they are all formally related in an exact way, as presented in Chapter 3.

Another important synthesis model is the parallel resonance formulation. This can be exactly equated to an all-pole model, but in the more practical case of fixed bandwidths the conversion from series to parallel is approximate. Details of the parallel parameters were discussed in Chapter 4.

The last of the six parameter types to be investigated is a set of articulatory parameters. Articulatory modelling is a large subject, with a considerable history as presented in Chapter 2. It is challenging to attempt an investigation of the articulatory approach because historically it has been difficult to produce natural and intelligible synthesis using articulatory parameters (Witten, 1982, p23). One reason is that while there is general agreement on acoustic parameters for phonemes, there is little data (and less agreement) for articulatory parameters.

Nevertheless an articulatory approach should be attempted, because such a model differs in several important respects from the other five models studied:

1) parameter values for all the other models could be derived directly from acoustic data; the articulatory representation is a separate level, a separate link in the speech chain.

2) articulatory models typically have fewer dimensions than do resonance or LPC models; the other five models in this study have ten parameters, whereas an articulatory model can have as few as three parameters.

3) articulatory models involve the motion of physical bodies through space and time, and so are subject to physical constraints during transitions; resonances and LPC parameters are not similarly constrained.

The key problem for an articulatory model is to maintain, to the extent possible, uniformity with the synthesis from the other five models. Therefore the same basic phoneme data should be used for all six synthesizers. As formant data exists, and all the other five models can be manipulated with formant data or an exact equivalent (or near equivalent in the case of parallel resonances with fixed bandwidths), a major problem for this study of an articulatory model is converting from formant data to articulatory parameters.

Four methods were examined in detail, and will be presented in sections three to six of this chapter. The first used equations published by Ladefoged (1978) which convert formant frequencies to tongue and lip parameters. The second used the vocal tract weighting factors of Harshman (1977), which produce an area function from tongue and lip parameters. In our work this process was reversed and the weighting factors were used as basis vectors, to allow conversion from an area function into tongue and lip values. The third method was to compute formant data from articulatory parameters, collect the data graphically, and then manually search for articulatory configurations having the requisite formants. The final approach was to use a gradient search algorithm to

replace the manual search.

The result of this effort is a set of articulatory parameters that provides a good match to the original resonance data for most sounds, so far as formant frequencies are concerned. The data are presented in section six of this chapter. These values allowed synthesis of speech data from an articulatory representation. The resultant intelligibility and naturalness were measured in Experiments IV and V, Chapters nine and ten.

The study uncovered a basic difficulty with formant bandwidths and the lossless tube vocal tract model (Rabiner & Schafer, 1978, p82). This matter is discussed in section five of this chapter.

5.2 ACOUSTIC-ARTICULATORY INVERSION

There are two problems to solve before proceeding with synthesis-by-rule from an articulatory model: which model to select, and how to determine phoneme target values for parameters of the model. In our case the choice of model is governed by the need for parameter data. If any of the available articulatory models already had phoneme data compatible with that used for the other five synthesizers in this study, then that would be the natural choice.

The alternative is to have a method for determining articulatory parameters from acoustic data (such as the Klatt table of formant values for phonemes). The difficulty is that going from an acoustic to an articulatory specification is an unsolved problem in speech science, though much work has been done (Ladefoged, 1978; Atal et al, 1978; Charpentier, 1984). The problem is known as the acoustic to articulatory inversion; the search for articulatory parameters is in essence the search for an articulatory model for which an inversion method exists.

5.2.1 Articulatory Models

As discussed in Chapter 2, electronic models of the vocal system have been under development since Dunn (1950). However there is a basic distinction to be made between vocal tract analogues and articulatory models.

Vocal tract analogues (Stevens et al, 1953; Kelly and Lochbaum, 1962; Ishizaka and Flanagan, 1972) are transmission line models of the entire vocal tract. As such they typically contain 20 or more sections, each with one to three parameters. Thus these are models with rather more parameters than is the case for resonance synthesis.

By contrast, articulatory models specify a vocal tract shape in terms of a small set of parameters (Dunn [3 parameters], 1950; Stevens and House [3], 1955; Coker [5], 1968; Lindblom and Sundberg [2 tongue factors], 1971; Mermelstein [6], 1973; Harshman et al [3], 1977; Flanagan et al [6], 1980). Typically the tongue position is described by two to four parameters. Additional parameters may describe lip opening and protrusion, jaw position, and length.

In the present research, a type of vocal tract model is already available through the area function parameters. This is a special sort of vocal tract, one in which there are no losses except at the lip end: the lossless tube (Rabiner & Schafer, 1978, p82).

Further consideration must be given to the low-dimensionality articulatory models, in particular whether any have a phoneme table for synthesis, or alternatively have an acoustic-articulatory inversion method. Unfortunately the available literature on articulatory models did not include a phoneme table. Most papers limit published results to data for vowels (Ladefoged, 1978; Charpentier, 1984), or to analysis-synthesis of short phrases (Sondhi and Resnick, 1983). We were thus led to consideration of inversion methods. There are several possibilities, which will be described in the next sub-section.

5.2.2 Inversion Techniques

Several approaches to determining articulatory shapes from acoustic data have been studied. The simplest is a set of equations published by Ladefoged et al (1978). A related study by Harshman et al (1977) determined a 'best' set of articulatory parameters from vocal tract data. A reversal of this method potentially allows conversion from formants to lossless tube area function to articulatory parameters. A large study by Atal et al (1978) used a vocal tract model to compute formants, and then 'reverse sorted' the data. Finally Charpentier (1984) also used a vocal tract model to compute acoustic consequences, but used an adaptive search procedure to attempt to match formants.

Each of these approaches was tried (with modifications) as a possible solution to the problem of determining articulatory parameters whose acoustic consequences would match the Klatt resonance data for phoneme synthesis. The methods are presented in the next four sections of this chapter.

5.3 LADEFOGED/HARSHMAN METHOD

The method of Ladefoged (1978) consists of a simple set of three equations for determining two tongue parameters plus lip opening from the first three formant frequencies. The method originated with the factor analysis of vocal tract shapes of Harshman et al (1977). The two tongue parameters are the two dimensions which best described ten vowels from each of five subjects. These tongue parameters will be considered again in the next section, as they provide a link between vocal tract area functions and articulatory parameters.

In the Ladefoged (1978) work, formants were measured for the same set of 50 vowels. A set of multiple regression analyses were then performed between formant frequencies and tongue or lip parameters. The analysis considered 25 acoustic variables (the first three formants and various cross and

triple products and reciprocals). The published equations are the resultant best three-term equations for predicting an articulatory parameter from formant data.

The articulatory parameters are tongue front raising (FR), tongue back raising (BR), and lip opening (LIP). These are computed from the first three formant frequencies (F1-F3) by the following relations:

$$FR = 2.309*(F2/F3) + 2.105*(F1/F3) + 0.117*(F3/F1) - 2.446 \quad (3.1)$$

$$BR = -1.913*(F1/F2) - 0.245*(F2/F1) + 0.118*(F3/F1) + 0.584 \quad (3.2)$$

$$LIP = 3.0E-3*F2 - 3.43E-7*F2*F3 + 4.143*(F1/F2) - 2.865 \quad (3.3)$$

The tongue parameters can then be used to determine 16 vocal tract diameters, as in the original Harshman et al (1977) study. That study had an 18-section vocal tract model. Section 18 is determined by the parameter LIP, and section 17 is the arithmetic mean of LIP and section 16.

In our case we require a ten section area function to conform with all the other ten-parameter synthesizers. Accordingly the method was simplified to produce eight vocal tract areas instead of sixteen, simply by using every other term in the weighting functions. A ninth area was determined by lip opening. A fourth articulatory parameter represented the 'area' beyond the lips, which is the term representing radiation loss (the only loss in the lossless tube model), and is the tenth and final area parameter.

Thus the original Klatt data went into the Ladefoged equations to produce articulatory parameters, and these resultant parameters were used to produce a ten-term vocal tract area function. This area function could then be converted by the analytic relations (given in Chapter 3) to the direct form, and the root finding routine would then determine the formants. Comparison of obtained formants with the original Klatt data provides a check on the accuracy of the inversion.

Table 5.1: Acoustic consequences (for a lossless tube) of articulatory parameters estimated by the Ladefoged equations.
F1-F5 = formant frequency, Hz; B1-B5 = formant bandwidth, Hz;
FR = tongue front raising; BR = tongue back raising;
LIP = lip opening; LOSS = reflection coefficient at lips.

	F1	B1	F2	B2	F3	B3	F4	B4	F5	B5	FR	BR	LIP	LOSS
i	134	22	2277	3	3233	1063	3683	19	5000	42	0.57	0.49	1.82	0.31
a	619	151	1128	647	2777	204	3725	107	4482	451	-0.36	-0.24	2.08	0.38
u	0	332	745	248	2705	16	3695	31	4443	48	-0.39	0.51	0.63	0.39
w	0	186	565	98	2931	5	3741	22	4392	19	-0.64	0.55	0.48	0.42
j	13	0	2342	0	3676	1507	3715	0	5000	0	0.68	0.58	1.72	0.29
r	349	515	1536	472	2290	238	3581	61	4571	204	0.32	0.02	1.02	0.39
l	0	162	456	422	2814	1	3953	6	4640	41	-0.29	0.94	0.47	0.36
v	0	116	500	587	2506	2	3774	2	4837	68	0.10	0.75	0.48	0.24
b	0	2	176	15	2775	0	3986	0	4954	3	-0.03	1.08	0.09	0.07
g	27	1	2197	0	3633	0	4007	189	5000	0	1.31	0.16	2.33	0.21
m	0	199	735	197	2686	12	3708	27	4458	46	-0.35	0.56	0.55	0.32

The general result of this method was that it worked quite well for vowel formants, as shown in Figure 5.1. This was encouraging, considering there were a number of differences between this application and the original Harshman et al study, as follows:

1) formant data from a new subject
2) consonant as well as vowel sounds
3) eight vocal tract parameters plus lips instead of the original 17 plus lips
4) resonances computed from the lossless tube model

However bandwidths were not well estimated for vowels, and were worse for consonants, as can be seen in Table 5.1. This table gives the four articulatory parameters and their resultant resonance centre frequencies and bandwidths. Only

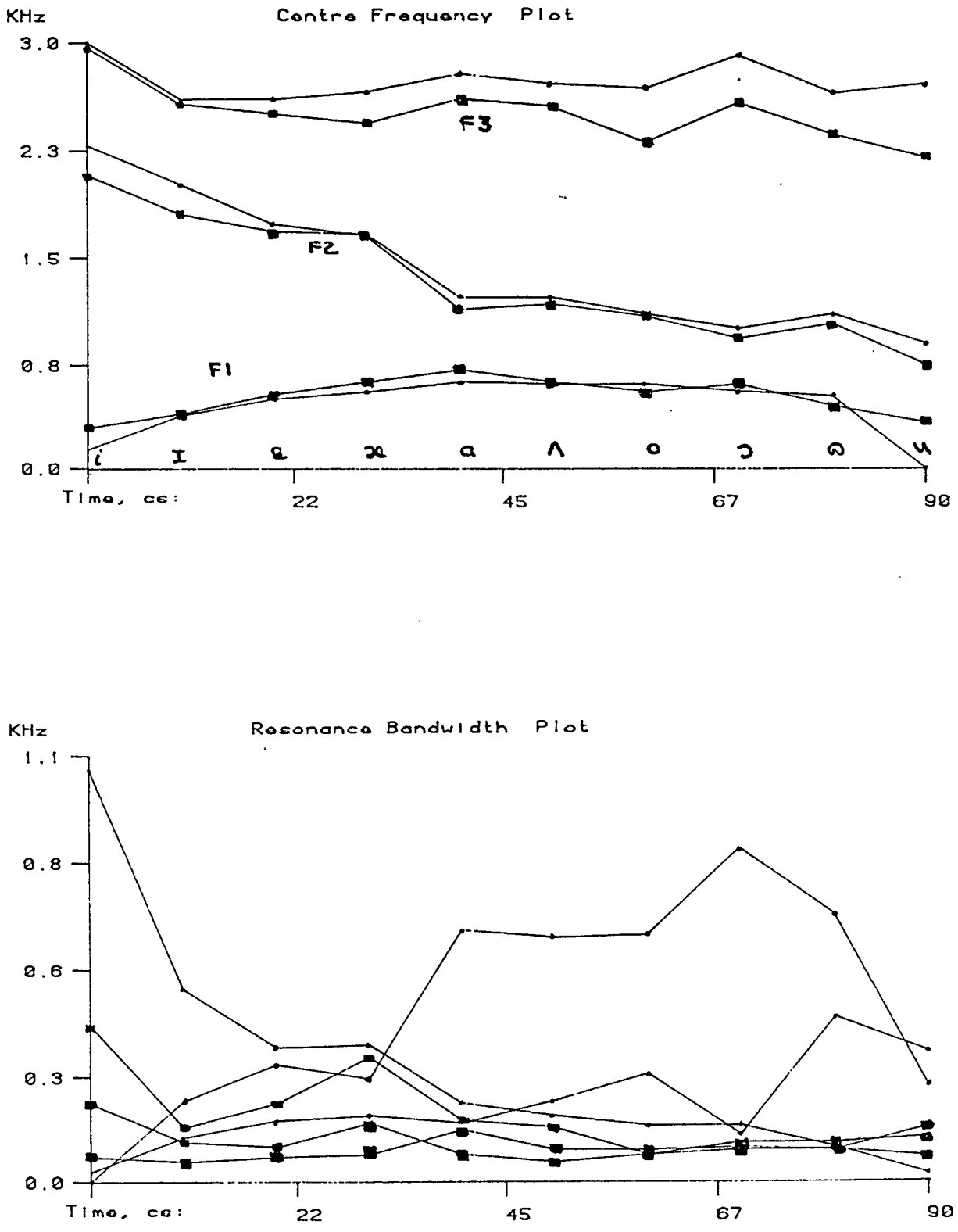


Figure 5.1: First three formant frequencies for nine vowel sounds, as determined from the Klatt data using the Ladefoged equations method. Articulatory parameters (•) vs original Klatt series resonance data (▪). Plotted in frequency vs time as a stylised spectrogram.

eleven phonemes are listed, constituting extreme articulatory positions. These are the eleven sounds studied in detail in Experiment I, Chapter Six.

Figure 5.2 gives the positions of these sounds (and six additional vowels) in the FR, BR plane, for comparison with Ladefoged's original results (Figure 5.3) and with the results of the methods described below (Figures 5.5 and 5.10).

The resonant frequencies were not as well estimated for consonants as for vowels. The result is not a close match to the Klatt table, and so another method was tried.

5.4 BASIS-FUNCTION METHOD

One way to try to determine articulatory configuration from acoustic data is through linear prediction. As discussed in Chapter 3, the results of a linear prediction analysis can be converted to the reflection coefficients of a lossless tube, and thence to the areas. This is not necessarily a true vocal tract area, for at least two reasons:

1) a real vocal tract has significant losses at the walls (Rabiner & Schafer, 1978);

2) even for a lossy tract there are many shapes which have identical first, second and third formant frequencies (Atal et al, 1978).

Because of the reservations about the lossless tube model, the area function computed from linear prediction is often referred to as a pseudo-area function. However it can be a reasonable tract shape, especially for vowels, as shown by Wakita (1973). Sondhi (1977) provides an extensive review and criticism of the method.

It is straightforward to compute a 'pseudo-inversion' and get a pseudo-area function. This approach gets close to determining articulatory parameters, even if only

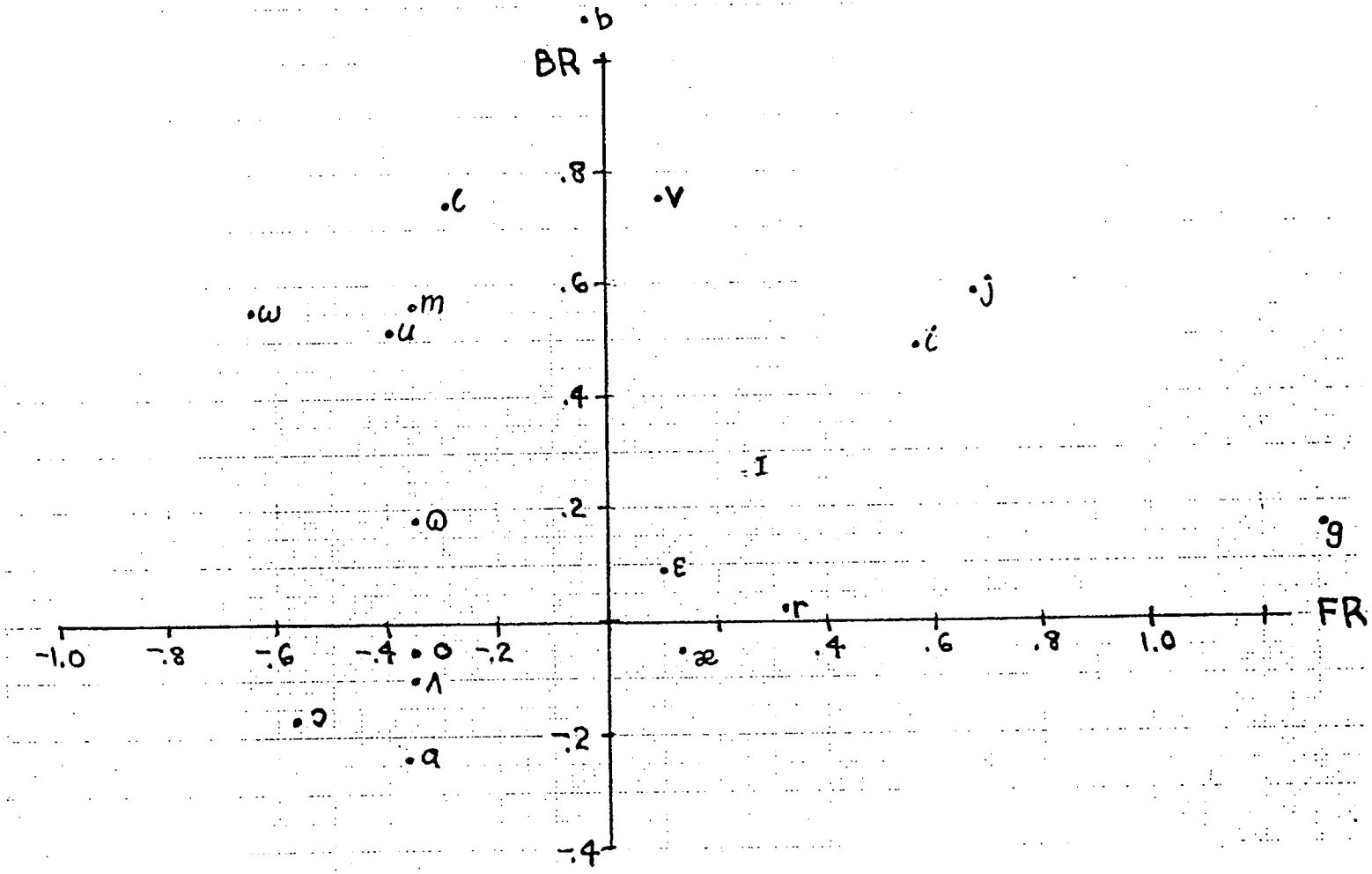


Figure 5.2: Loci in FR vs BR plane of 11 phoneme target values, as determined from the Klatt data using the Ladefoged equations method.

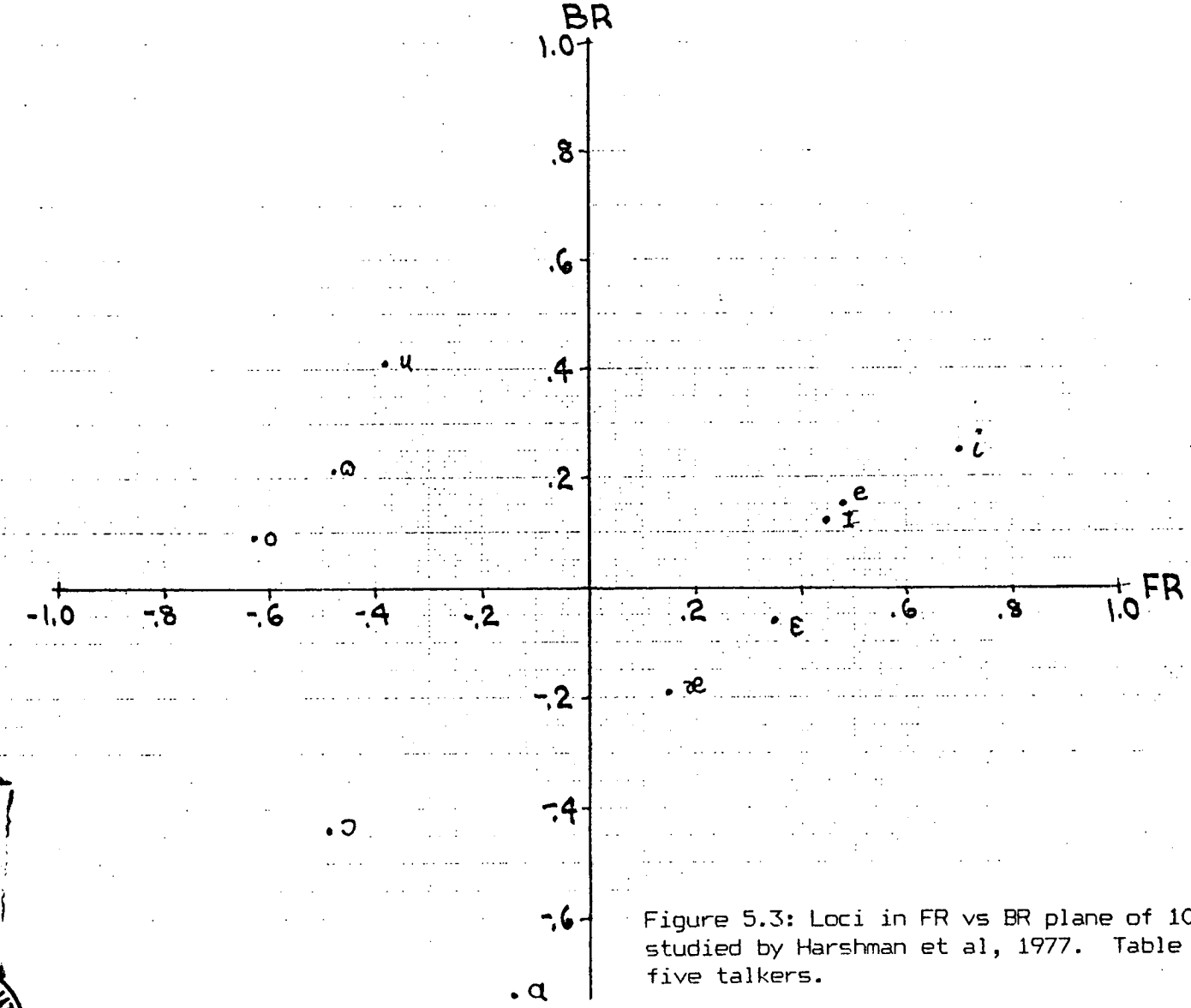


Figure 5.3: Loci in FR vs BR plane of 10 vowels originally studied by Harshman et al, 1977. Table V, p705; averages of five talkers.

pseudo-parameters. Can articulatory parameters be computed from an area function?

All the articulatory models provide a method to go the other direction, and compute a vocal tract from the model parameters. However there is no built-in method for reversing this process in the usual models.

5.4.1 Tongue-Hump Models

The problem is that most such models are based on describing a constriction along a tract. The models of Dunn, Stevens and House, Coker, and Lindblom & Sundberg all have two parameters which manipulate a single constriction. We shall consider one of these models as an example.

The Stevens and House (1955) articulatory model is used to compute area functions given three articulatory parameters: place and degree of constriction; lip opening. The reverse process is not explicitly considered. One could estimate articulatory parameters by simply taking that point in an area function where the area is minimum, but that does not necessarily represent the point of constriction. The minimum area may be at one end, independent of a constriction elsewhere. There may be more than one constriction. Finally, it may well be that a completely unrelated set of articulatory parameter values would actually give a better match (in terms of spectrum) than would those resulting from trying to fit a constriction model to a vocal tract (Atal, 1978).

Attempting to proceed from tracts to 'tongue humps' is not guaranteed to yield a sensible solution, and certainly not guaranteed to find a best solution. The same problem is characteristic of the other constriction models. One can compute a tract from a tube with a single constriction, but not necessarily vice-versa.

5.4.2 Basis Vector Approach

A fundamentally different approach is to describe a vocal tract in a more formal, less geometrical way. This approach is represented by the studies of Mermelstein (1967) and Schroeder (1967), which determined the eigenvectors of vocal tract shapes. For a uniform, lossless tube these have the form of cosine functions. These functions can be interpreted as the basis vectors for a space of vocal tract shapes. Further, analysis of an arbitrary vocal tract shape is a matter of taking projections onto these vectors. The process is an example of decomposition using orthogonal functions; Fourier analysis is a related example.

But cosine functions are not necessarily relevant to actual vocal tracts. Here the study of Harshman et al (1977) is relevant, because it is a factor analysis of vocal tract shapes to determine something like basis vectors. They are used in the Harshman study as vectors of loadings, where each loading (vector component) represents the amount a particular vocal tract section varies as the articulatory parameter varies.

Our second approach to the acoustic-articulatory inversion was to interpret the Harshman factors (vectors of loadings) as actual basis vectors, and then use them for the analysis of pseudo-area functions.

The Harshman et al (1977) approach to the description of vocal tract shapes resulted in two main tongue factors, front raising (F) and back raising (B). These factors (each is a vector of loadings onto each vocal tract section) were established from multidimensional scaling (PARAFAC; Harshman, 1970) of x-ray data for fifty vowels (five speakers, ten vowels each). The two dimensions accounted for 94% of the variance or the original data, and a third dimension only accounted for an additional 1%.

A weight on each tongue factor produces a vocal tract shape estimate from the following formula:

$$\underline{E} = \underline{T}\underline{W} + \underline{N} \qquad (5.4)$$

where: $\underline{E}$ is an n-component vector of estimated vocal tract diameters, for an n-tube model;

$\underline{T}$ is an nx2 matrix; the first column is $\underline{F}$, the second is $\underline{B}$; each is a vector of loadings;

$\underline{W}$ is a 2x1 matrix, consisting of the weights w1 and w2, one for each loading vector $\underline{F}$ and $\underline{B}$; w1, w2 are estimates of the articulatory parameters FR and BR.

$\underline{N}$ is a vector giving a neutral tongue postion.

The problem for acoustic-articulatory inversion is to reverse the above procedure with minimum squared error of estimation. The starting point is the Klatt resonance data. The exact relations (from Chapter 3) are then used to compute a pseudo-area function. Appropriate choice of area at the glottis (to agree with the Ladefoged and Harshman scaling) allows conversion to the vector of vocal tract diameters, $\underline{X}$.

Given $\underline{X}$, $\underline{T}$ and $\underline{N}$, the problem is to compute $\underline{W}$, the weights on the $\underline{F}$, $\underline{B}$ loading vectors which in turn produce $\underline{E}$, the best estimate of $\underline{X}$ (in the sense of least squared error).

$\underline{X}$ is the vector to be estimated by $\underline{E}$, as in (5.4).

Squared error: $J = \langle \underline{X} - \underline{E} \rangle = (\underline{X} - \underline{E})' \ (\underline{X} - \underline{E})$ (5.5)

Notation: $\underline{V}'$ is the transpose of any vector $\underline{V}$;
$\langle \underline{X}, \underline{Y} \rangle$ is the inner product of vectors $\underline{X}$ and $\underline{Y}$.

J is a quadratic surface. Using (5.4) in (5.5) and rearranging in terms of $(\underline{X}-\underline{N})$, the displacement from a neutral tongue position, yields:

$$J = \langle \underline{X} - (\underline{T}\underline{W}+\underline{N}) \rangle = \langle (\underline{X}-\underline{N}) - \underline{T}\underline{W} \rangle$$

$$= (\underline{X}-\underline{N})'(\underline{X}-\underline{N}) - (\underline{X}-\underline{N})'\underline{T}\underline{W} - \underline{W}'\underline{T}'(\underline{X}-\underline{N}) + \underline{W}'\underline{T}'\underline{T}\underline{W} \qquad (5.6)$$

A minimum for J is sought where the derivative of J (with respect to each weight w) is zero:

$$\delta J/\delta \underline{W} = \begin{bmatrix} \delta J/\delta w1 \\ \delta J/\delta w2 \end{bmatrix} = -2\underline{T}'(\underline{X}-\underline{N}) + 2\underline{T}'\underline{T}\underline{W} \qquad (5.7)$$

Setting (5.7) to zero and solving for $\underline{W}$:

$$\underline{T}'(\underline{X}-\underline{N}) = (\underline{T}'\underline{T})\underline{W} \qquad (5.8)$$

$$(\underline{T}'\underline{T})^{-1}\underline{T}'(\underline{X}-\underline{N}) = \underline{W} \qquad (5.9)$$

Equation (5.9) is considerably simplified for certain types of basis vectors $\underline{T}$. If the vectors are orthogonal, $\underline{T}'\underline{T}$ is a diagonal matrix. As such its inverse is another diagonal matrix, and each element on the diagonal is simply the reciprocal of the corresponding element in $\underline{T}'\underline{T}$. Because the inverse is diagonal, (5.9) leads to independent rather than simultaneous equations:

$$w1 = \underline{F}'(\underline{X}-\underline{N}) \;/\; \underline{F}'\underline{F}; \quad w2 = \underline{B}'(\underline{X}-\underline{N}) \;/\; \underline{B}'\underline{B} \qquad (5.10)$$

For an orthonormal basis set, $\underline{T}'\underline{T} = \underline{I}$ and the solution is simply:

$$\underline{W} = \underline{T}'(\underline{X}-\underline{N}) \qquad (5.11)$$

In practice, a simplification was used. Referring back to equation (5.9), if we do not have an orthonormal set of vectors then the term $\underline{T}'\underline{T}$ does not drop out. However this is in our case only a 2x2 matrix which can be explicitly inverted, so it is a simple matter to directly solve equation (5.8) using Cramer's rule:

Let $\underline{P} = \begin{bmatrix} p1 \\ p2 \end{bmatrix} = \underline{T}'(\underline{X}-\underline{N})$ ($\underline{P}$ for projection).

Then: $$\underline{W} = \frac{\text{Adjoint } (\underline{T}'\underline{T}\ \underline{P})}{\text{Determinant } (\underline{T}'\underline{T})} \qquad (5.12)$$

Because $\underline{T}'\underline{T}$ is only 2x2, (5.12) can be written out explicitly as follows:

$$\text{Det } (\underline{T}'\underline{T}) = t't11 \cdot t't22 - t't12 \cdot t't21 \qquad (5.13)$$

$$w1 = (p1 \cdot t't22 - p2 \cdot t't21\ ) \ / \ \text{Det } (\underline{T}'\underline{T}) \qquad (5.14)$$

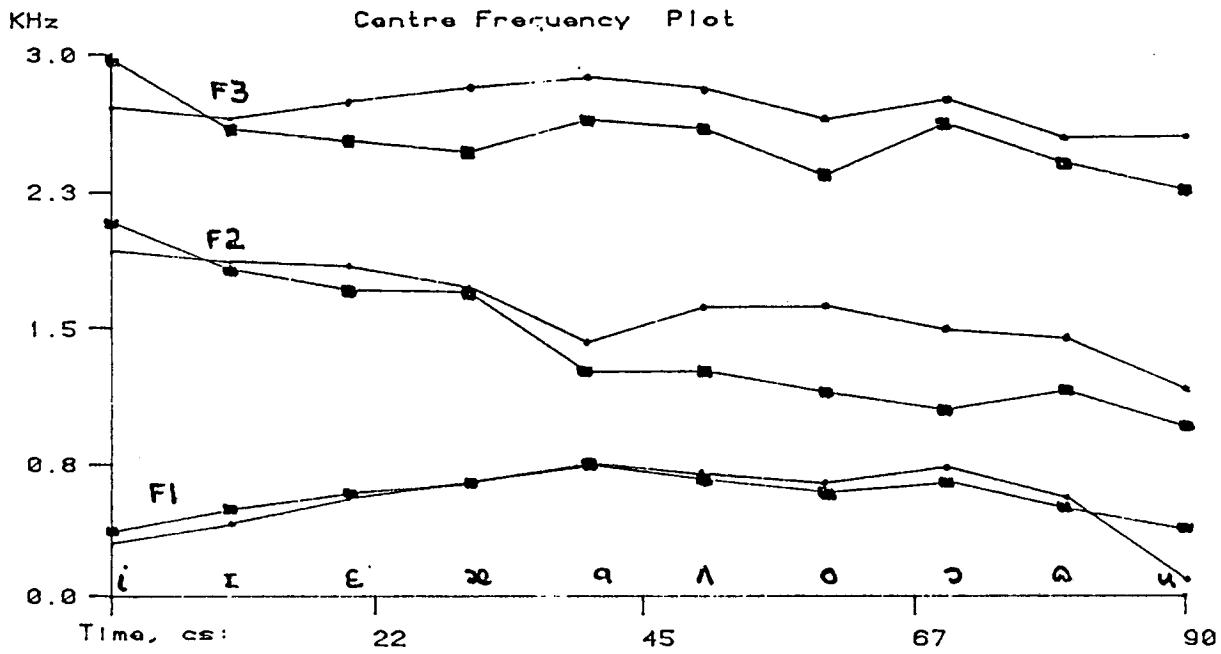
$$w2 = (p2 \cdot t't11 - p1 \cdot t't12\ ) \ / \ \text{Det } (\underline{T}'\underline{T}) \qquad (5.15)$$

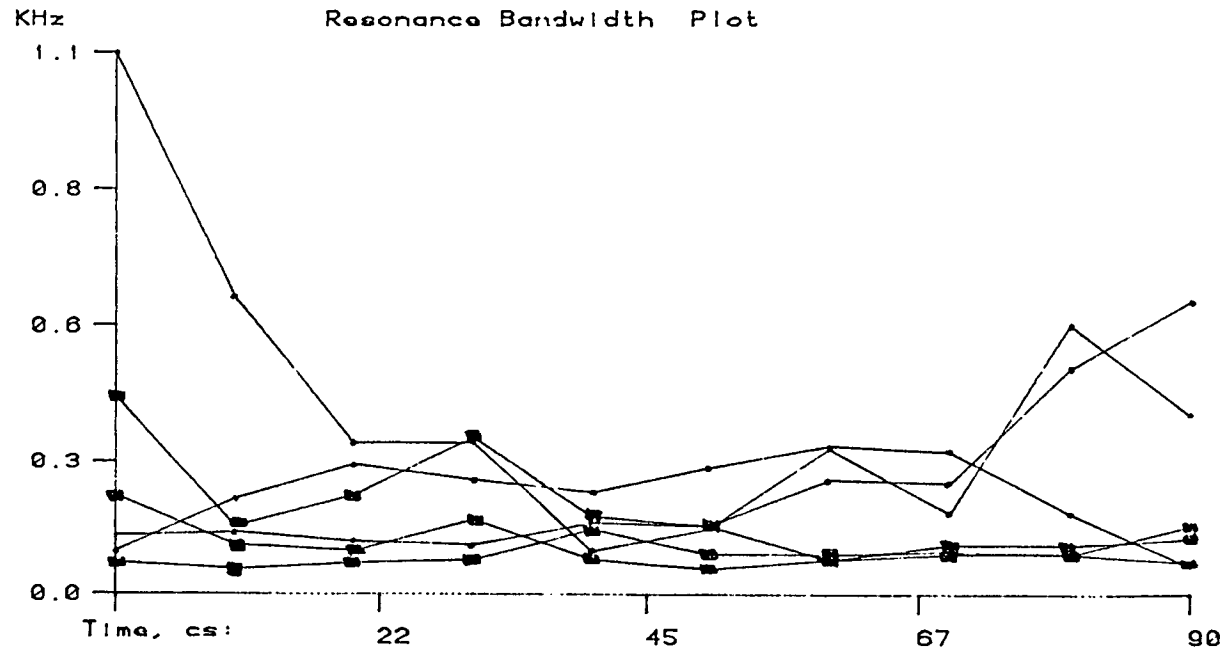
The algorithms of section 3.2.1 produce a vocal tract area function directly from formant frequencies and bandwidths. The method just described produces tongue parameters from a vocal tract as described by a set of diameters (rather than areas). Also there are differences in the assumed tube size at the glottis. Once appropriate compensation has been introduced for these differences of detail, the Klatt formant data can be used to produce tongue parameters via projection of pseudo-areas onto the front-raising and back-raising tongue parameters.

The third articulatory parameter, lip opening, is determined immediately from the pseudo-area function as the area just before the end-most area. The end-most area represents the outer world, and it is the ratio of these two areas which is determined by the final reflection coefficient, representing loss at the lips. The fourth and last articulatory parameter is this loss term.

As before, the method was tested by using the resultant parameters to determine a new vocal tract shape, and then interpreting this shape as a lossless tube and thence by the formal methods of Chapter 3 to a solution for formant frequencies and bandwidths. The results for nine vowels are shown in Figure 5.4. The results for 11 selected sounds including eight consonants are given in Table 5.2, which lists the articulatory parameters as well as the resultant

Figure 5.4: First three formant frequencies for nine vowel sounds, as determined using the basis vectors method. Articulatory parameters (•) vs original Klatt series resonance data (▪). Plotted in frequency vs time as a stylised spectrogram.





formant frequencies and bandwidths. These articulatory data are plotted in the FR vs BR space in Figure 5.5.

Table 5.2: Acoustic consequences (for a lossless tube) of articulatory parameters estimated by the basis vector method. F1–F5 = formant frequency, Hz; B1–B5 = formant bandwidth, Hz; FR = tongue front raising; BR = tongue back raising; LIP = lip opening; LOSS = reflection coefficient at lips.

	F1	B1	F2	B2	F3	B3	F4	B4	F5	B5	FR	BR	LIP	LOSS
i	315	118	1861	1101	2652	88	3793	107	4635	446	−0.21	0.66	1.59	0.31
a	665	146	1330	87	2788	205	3780	94	5000	344	0.41	−0.69	3.15	0.38
u	0	598	1046	362	2429	56	3610	38	4497	90	−0.07	0.24	0.76	0.39
w	258	709	1250	428	2337	96	3592	41	4527	116	0.11	0.13	0.84	0.42
j	331	182	1424	1329	2748	50	3805	129	4505	299	−0.35	0.66	1.35	0.29
r	268	763	992	528	2534	49	3644	47	4487	103	−0.20	0.36	0.85	0.39
l	0	804	1028	426	2454	66	3617	47	4493	111	−0.11	0.25	0.83	0.36
v	0	163	853	166	2537	24	3644	29	4481	63	−0.21	0.41	0.56	0.24
b	0	124	835	130	2539	30	3640	45	4474	100	−0.22	0.40	0.59	0.07
g	232	155	1580	1856	2686	25	3844	55	4693	369	−0.19	0.81	1.21	0.21
m	0	229	961	196	2442	35	3610	27	4487	63	−0.11	0.26	0.59	0.32

The results were similar to those of the first method (Ladefoged equations; Table 5.1, Figures 5.1 and 5.2). Vowel formant frequencies were well approximated, resonant frequencies for consonants were less well approximated, and bandwidths were poorly matched. Thus there was no improvement over the Ladefoged equations, and another approach was sought.

5.5 CONTOUR PLOTS

Method three began with a consideration of the approach of Atal et al (1978). They were able to provide a very detailed analysis of the relation between acoustic and articualtory parameters, using the method of reverse sorting.

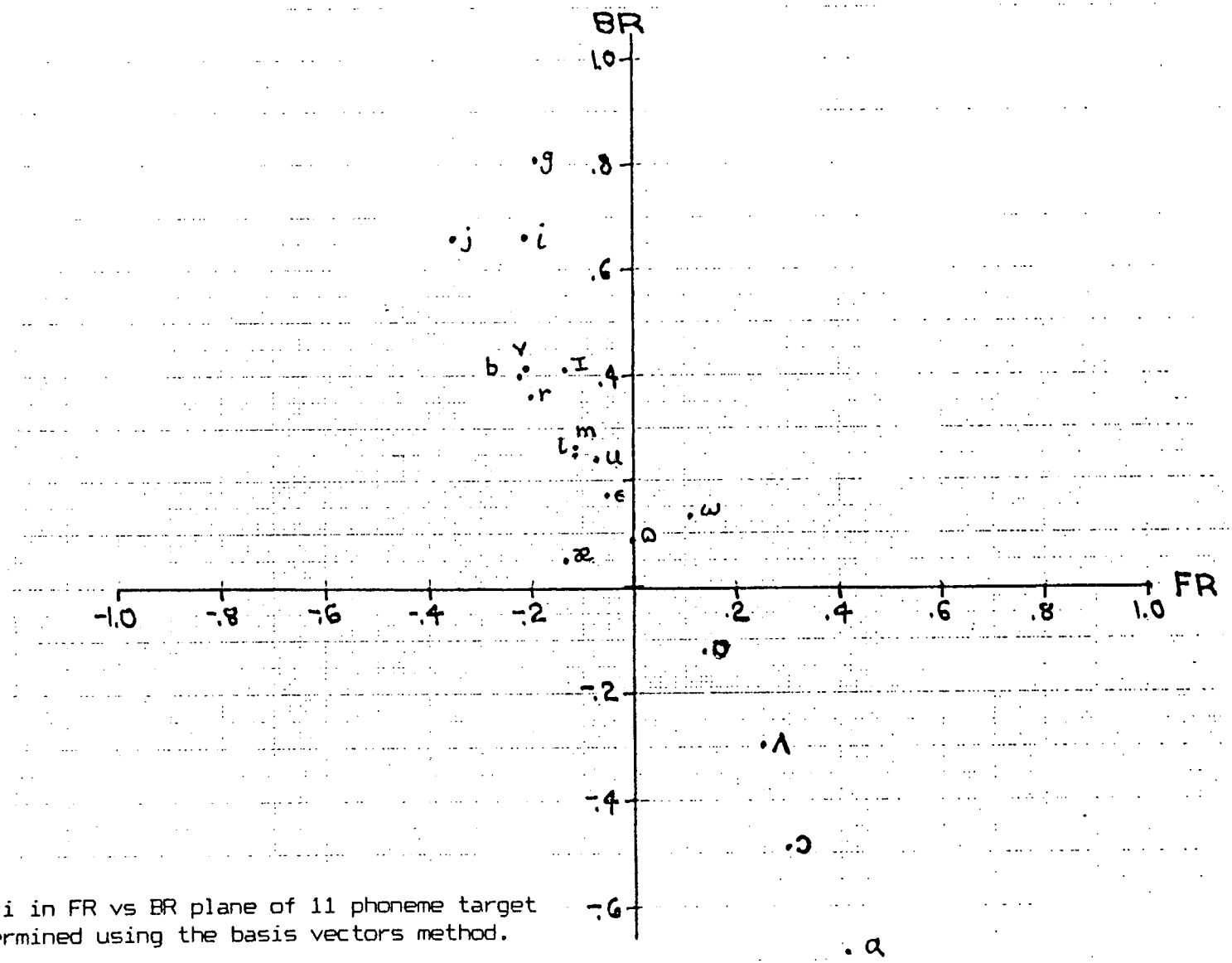


Figure 5.5: Loci in FR vs BR plane of 11 phoneme target values, as determined using the basis vectors method.

It is perfectly possible to compute the acoustic consequences of a particular vocal tract shape, at least if the vocal tract is approximated as N cylindrical sections. If the tube is lossless (except at one end) then the system function can be computed in a simple fashion. This function will be rational and all-pole, and the roots of the denominator correspond to vocal tract formants.

If the tube is lossy then the problem is more difficult. Each section has a representation as a 2x2 transmission matrix involving hyperbolic functions (Flanagan, 1972, p27). The whole system function is the product of N matrices, plus a term to account for the glottal impedance (which is assumed to be constant over time). The important point is that while the lossy-tube model is much more involved than the lossless tube, and cannot be written out in a closed form, it is still possible to numerically evaluate the product of transmission matrices to search for zeroes in the reciprocal. These values are roots of the denominator of the system function, even though the system function is not expressed as a polynomial; thus formant frequencies and bandwidths can still be estimated.

5.5.1 Reverse Sorting

Atal et al computed resonances for more than 30000 vocal tract shapes, and then used a sorting procedure to prepare a massive table which constitutes the acoustic-to-articulatory inversion method. Thus they solved the problem by reducing it to a matter of table look-up.

The difficulty is that there is no convenient way to publish the result. As they say: "Our computer programs are the principal tools for querying this information" (Atal et al, 1978, p1555). They do publish selected results in the paper (vowels only), but not enough to allow the interested reader to start from a phoneme table of formants and end up with articulatory data. One reason for not publishing more data is the difficulty of presenting a relationship between three

or more acoustic parameters and four or more articulatory ones: the representation does not lend itself to the printed form.

5.5.2 Graphical Sorting

A simplified version of the Atal et al procedure was used to explore the Ladefoged articulatory model (two tongue parameters plus lip opening). The parameters were varied over their sensible range and acoustic resonances (for the lossless tube) were computed. The results were then examined in terms of the relation between two articulatory parameters and a single acoustic parameters.

The effect of restricting consideration to one acoustic parameter at a time is that the result can be put on the printed page. The format used was that of contour plots: points of equal formant value in the two-dimensional tongue factors space were joined. The result is shown in Figure 5.6.

There are two points to be emphasised:

1) this procedure allows the full acoustic-articulatory relationship to be put in a visually accessible format;

2) the problem of finding an articulatory position corresponding to specific formant values does not then require a computer reverse sort of the data, but a human inspection of the contour plots.

5.5.2.1 Data presentation

There are many plots of acoustic vs articulatory data: major examples are Stevens and House (1955), Fant (1960), and Charpentier (1984), reprinted in Figure 5.7. They tend to be difficult to interpret because they plot one acoustic dimension vs one articualtory dimension, with possibly one

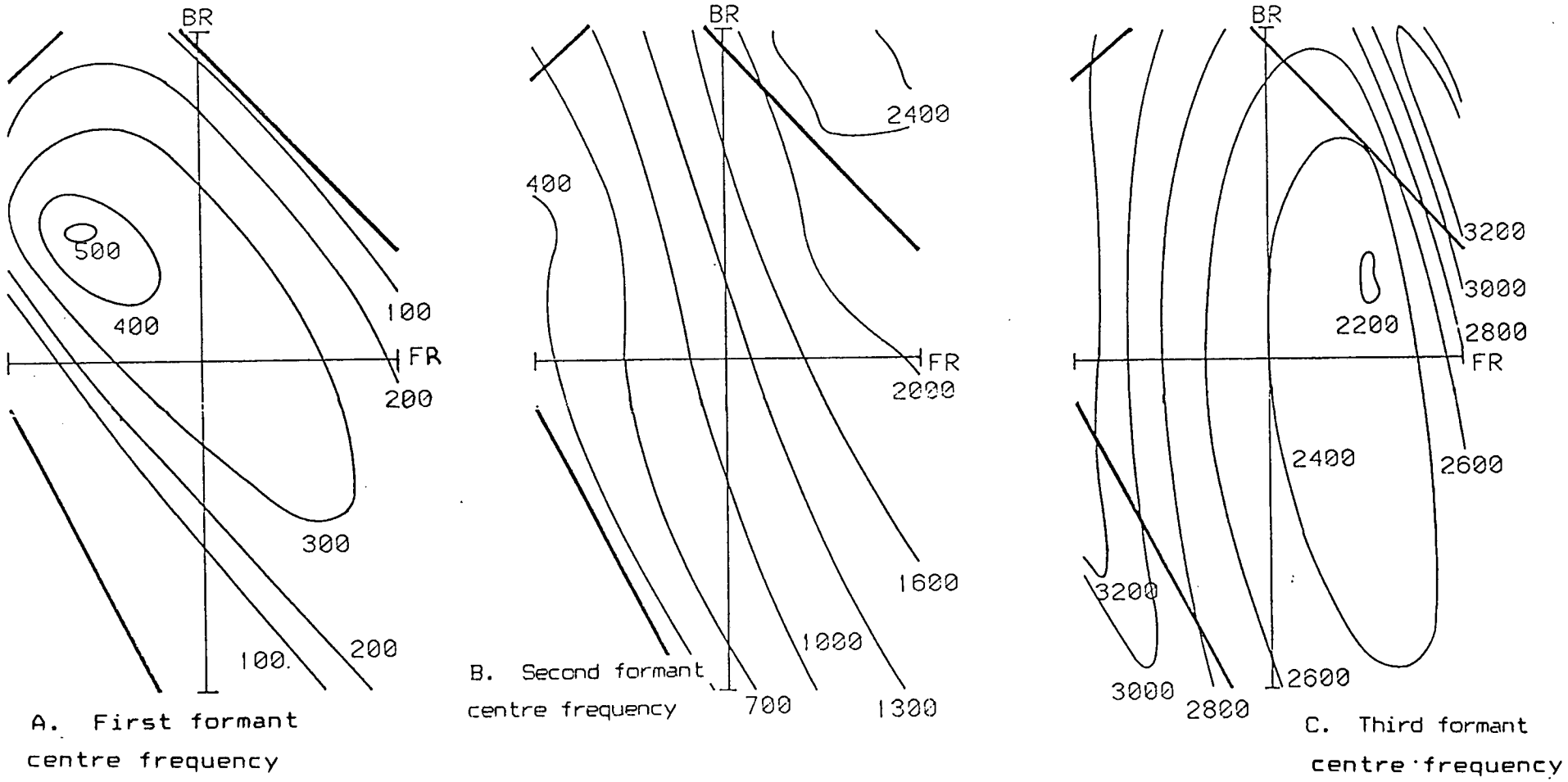
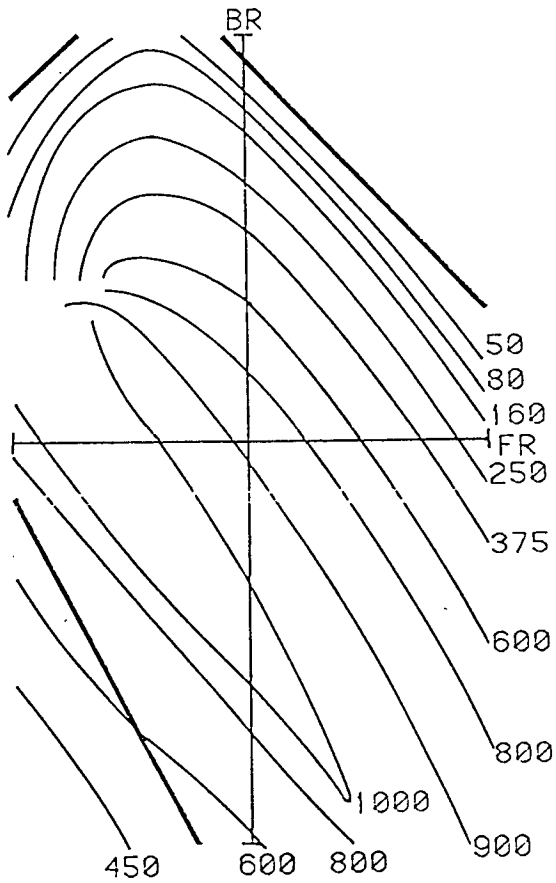
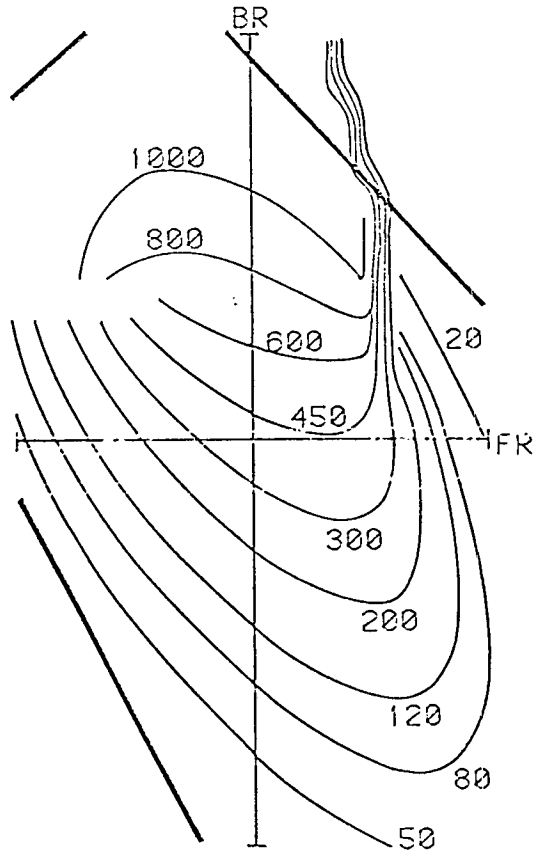


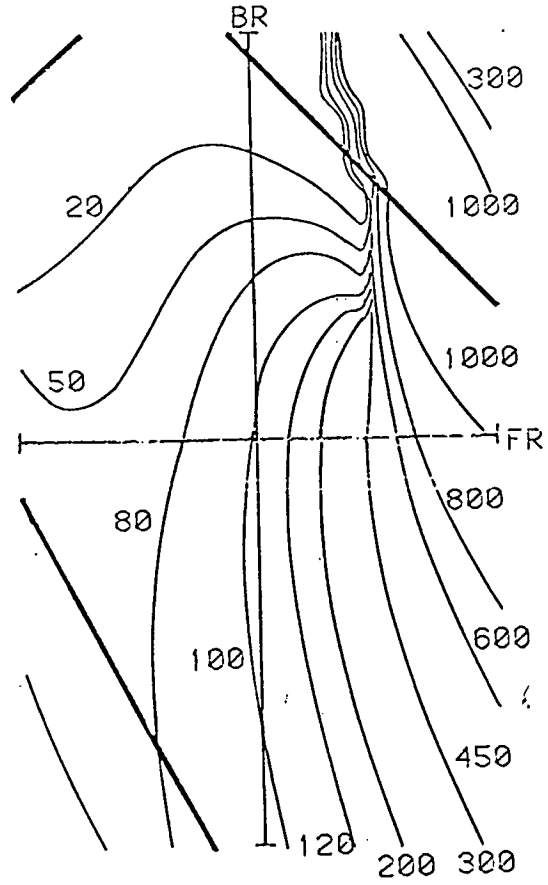
Figure 5.6A-F : Acoustic consequences of articulatory parameters. Resonance frequencies and bandwidths plotted as contours over the two-dimensional space of articulatory control parameters FR (front raising) and BR (back raising). Thick lines show limit of region of positive vocal tract areas. All values in Hz.



D. First formant bandwidth



E. Second formant bandwidth

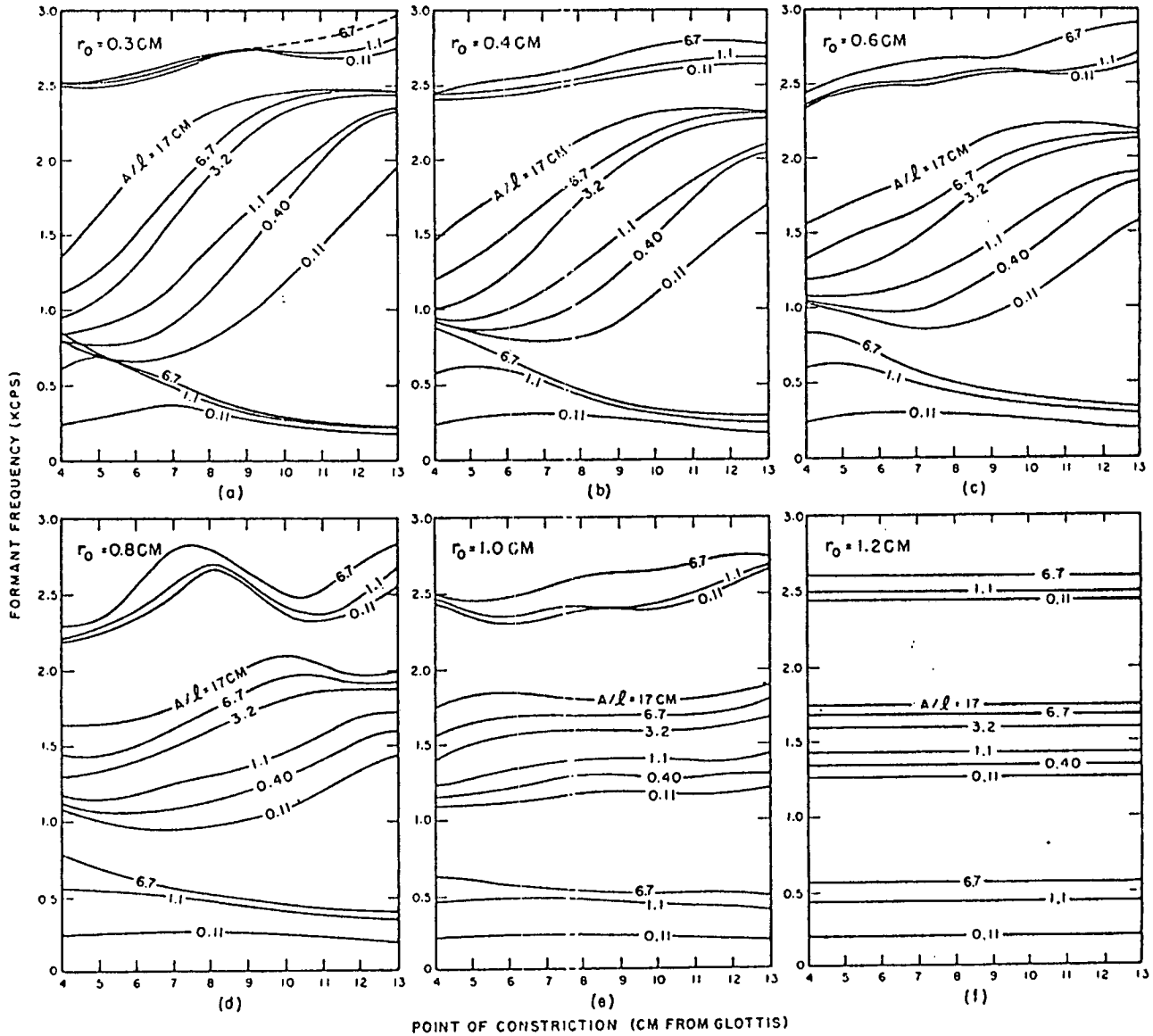


F. Third formant bandwidth

Figures 5.7A-C: Standard presentation of acoustic consequences of articulatory parameters.

Figure 5.7A: Stevens & House, 1955, Figure 3 p487;

Original caption: These curves show the frequencies of the first three formants as a function of r_o, d_o, and A/l. In each section data are presented for a given degree of constriction (r_o) as indicated, with mouth opening (A/l) as the parameter. Three families of curves corresponding to F1, F2, and F3 are plotted in each section. The abscissa is d_o, the distance from the glottis to the point of constriction.



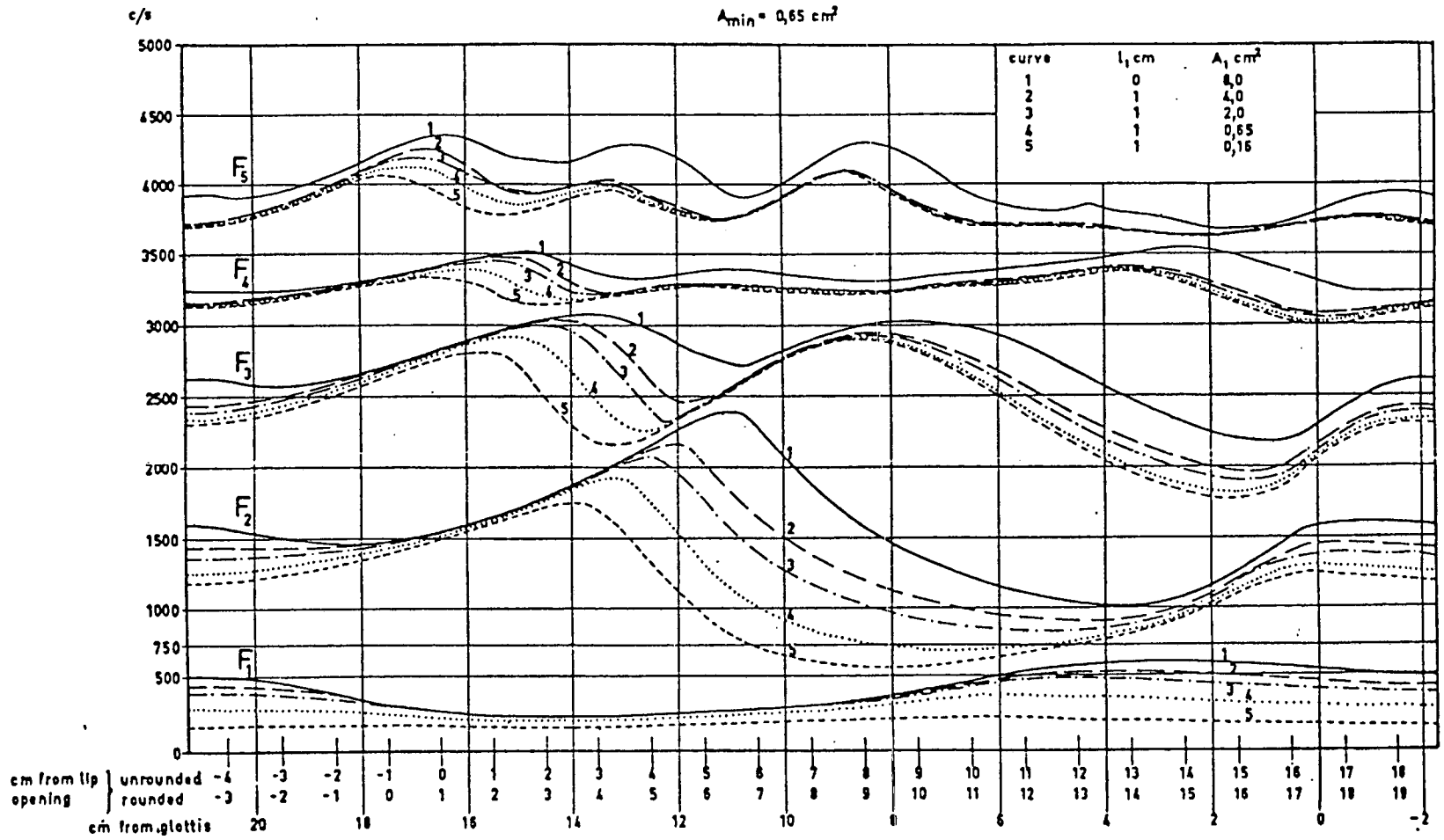


Figure 5.7B: Fant, 1970, Figure 1.4-11a p 82;
Original caption: Axial coordinate of the tongue constriction centre.

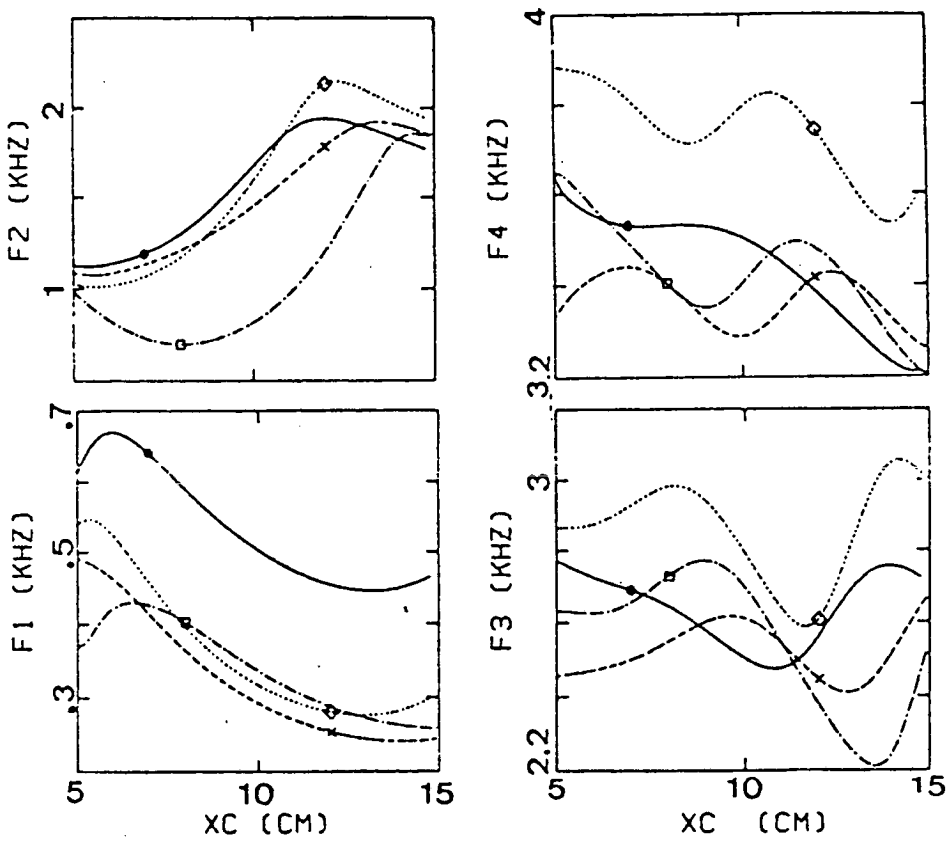


Figure 5.7C: Charpentier, 1984, Figure 6 p300.

Original caption: Variations of the first four formant frequencies as parameter Xc varies over the 5-15 cm range. These curves, and the variations of the other acoustic parameters (F5, P1, P2, P3), specify a multidimensional trajectory in the acoustic space. On each separate plot, notice the four curves, corresponding to different values of the other five articulatory parameters. These values are chosen so that the articulatory trajectories pass by synthetic vowels /a/,/i/,/y/,/o/; these vowels are indicated by a circle, a rhombus, a cross and a square, respectively.

more articulatory dimension as a parameter. Usually several plots are needed to give the full information. The Fant nomagraph is a single plot, but with values for five resonances on a frequency scale, and five values of lip opening as a parameter, making 25 lines to represent the graph of acoustic consequences of articulatory parameters.

The first use of contour plots of an acoustic variable as a surface over a two-dimensional articulatory space is also Stevens and House (1955), reprinted in Figure 5.8. First and second formant contours are given, but superimposed! Also they chose to make degree of constriction a parameter, rather than using it as one of the two dimensions of the figure. This makes it difficult to see the effect of motion of the constriction in its two dimensions (place of constriction and degree of constriction). There is no single plot which shows the acoustic effect of tongue (constriction) motion.

The new plots presented in this chapter are related to the Stevens & House approach, but are meant to add clarity. The two tongue control dimensions are used as horizontal and vertical coordinates, and lip opening is made a parameter. Then for simplicity only a single acoustic variable is plotted as a contour over the tongue parameter surface.

The significance of the difference in approach is meant to be apparent by comparison of the Stevens and House graphs of Figure 5.8 with Figure 5.6. Ladefoged et al (1978) did not use contours in presenting results of their model, as seen in their figures reprinted as Figure 5.9. They followed the 'acoustic vs articulatory' method with multiple graphs. Their paper does give one plot using tongue parameter dimensions (their Figure 5, p1029, which is equivalent to Figure 5.3 of this chapter) but not using contours.

The acoustic-articulatory mapping given in this chapter is essentially a clarification of the Steven & House contour plots method, but applied to the Ladefoged articulatory dimensions.

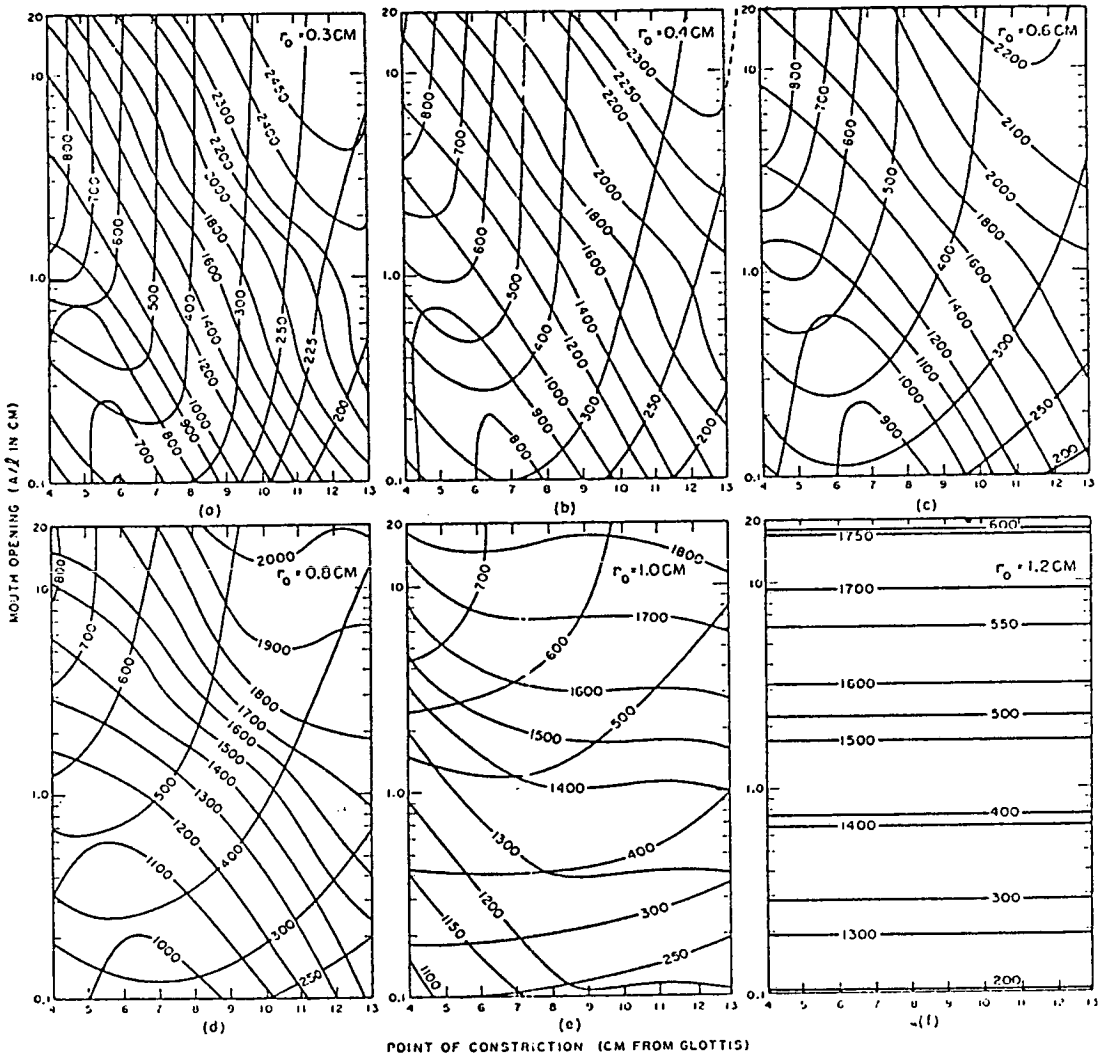


Figure 5.8: First use of contour plots to represent a relation between acoustic and articulatory parameters, Stevens and House, 1955, Figure 5 p489.

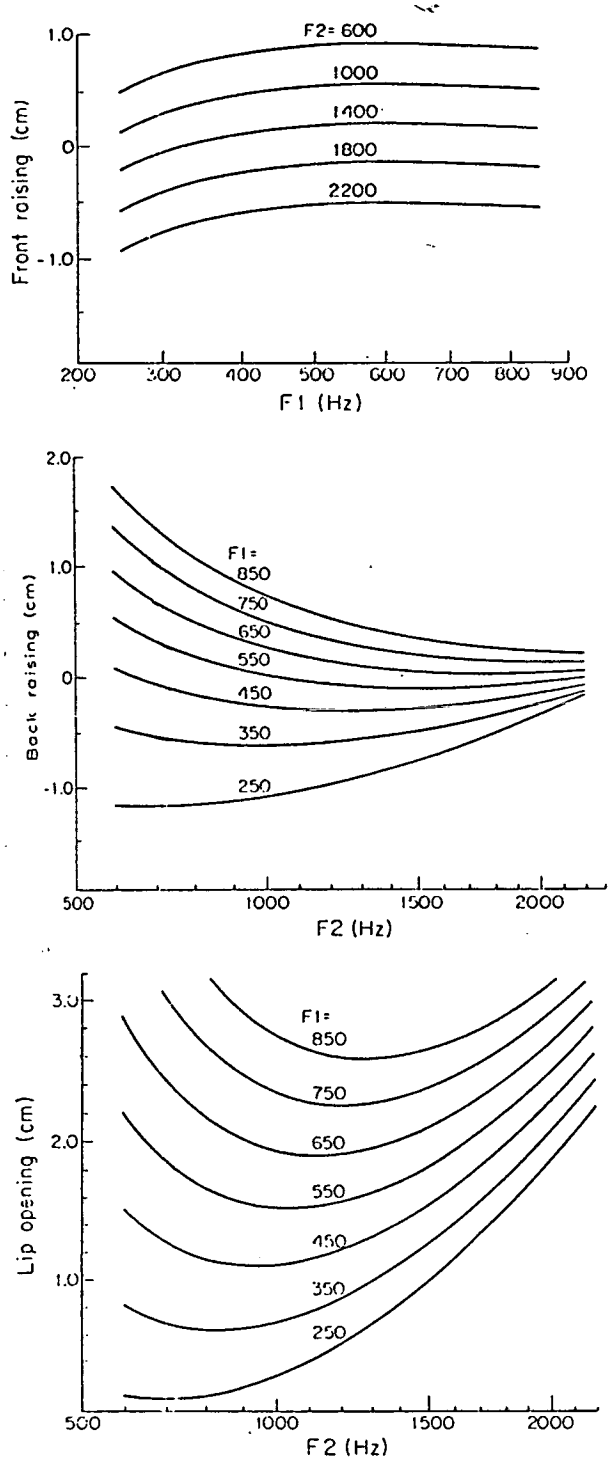


Figure 5.9: Original presentation of acoustic consequences of tongue front-raising and back-raising parameters, Ladefoged et al, 1978, Figures 6,7,8[pp1030-1031].

An examination of the formant frequency contours in Figure 5.6 shows a reasonably smooth surface for each. This is in sharp contrast to the contours for bandwidths. Here there are areas of very rapid bandwidth change. This is the explanation for the problems with bandwidths in sections three and four of this chapter: the bandwidths of the lossless tube model can change dramatically for a small change in articulatory position.

5.5.2.2 Visual search of articulatory space

The plots in Figure 5.6 are adequate for getting rough values for articulatory positions corresponding to vowels. Thus for the high-front vowel /i/ one looks in the upper-right quadrant and finds high F2 and low F1. This is also just the region where second and third formant bandwidths have their most precipitous change.

For a vowel like /u/ with low F2, it is apparent that the F2 contours intersect the F1 contours. Thus there are two positions with F2=1000 Hz and F1=300 Hz. In this case F3 information is required to determine which to select.

While articulatory positions for vowels can be determined by inspection, there are problems with respect to consonants. Figure 5.6 is just for one particular lip opening value. The full articulatory space is not just the two tongue parameters, but also lip opening. This third parameter is relevant to rounded vowels and labial consonants such as /bwrmv/.

To explore the possibility of using contour plots for consonants as well as vowels, a further set of plots were made for nine more degrees of lip opening. These are given in Appendix 3, fifty plots for five formant frequencies times ten values of lip opening. Bandwidth plots were not made, as the experience of the first two methods and the data in Figure 5.6 showed that bandwidths of the lossless tube were too unreasonable to merit further investigation.

The third approach to obtaining articulatory data corresponding to the Klatt resonance data was then 'simply' a matter of scanning the fifty plots, manually looking for the best point in a three-dimensional space for optimising from three to five formants (only three are significant for vowels, but F4 and F5 play important roles in the fricatives, as discussed in Chapter Seven).

At this point visual inspection of contour plots became a search of fifty plots rather than five, and a manual approach was abandoned in favour of a computer search.

5.6 ADAPTIVE SEARCH

The method chosen was a steepest-descent adaptive search of an error function (or cost function). In the present case the error function was a measure of how close the resonances of a vocal tract were to the target values of the Klatt phoneme data.

An adaptive search is not necessarily a sensible approach to finding an answer, unless something is known about the surface being explored. There were two sources of relevant information:

(1) the set of contour plots described in the previous section;

(2) the successful result of a related search approach by Charpentier (1984).

5.6.1 Charpentier Procedure

A six-parameter articulatory model (after Ishizaka; Flanagan et al, 1975) was studied, using a computer optimisation. The method used a squared-error between target and actual formant frequencies and amplitudes, iterating articulatory values to minimise the error.

The approach was sensitive to starting values, and so a set of articulatory values was chosen, acoustic consequences were computed, and the results held in a table. A 'good' starting value for the iterative procedure was then chosen from the table. This is reminiscent of the Atal et al reverse sorting, except Charpentier managed with only 36 selected articulatory positions, three orders of magnitude fewer than in the Atal et al study.

The method was tested on known articulatory and acoustic data. A test of 630 sets of formant data (computed from the articulatory model itself, so that the possibility of an exact match was ensured) resulted in less than 1% error at the first iteration, and perfect matches in all cases after at most six iterations. A further test of x-ray and acoustic data from Fant (1960) matched well on /i/ but poorly on /u/. This was attributed to a limitation of the articulatory model, rather than a problem with the iterative search method.

5.6.2 Procedure Used to Fit Klatt Data

The method used in the present study is similar to the Charpentier approach, with these differences:

(1) Starting estimate - rather than use a table or an arbitrary starting postition, the Ladefoged equations method of Section 5.3 was used to provide an initial starting position.

(2) The error function was based only on formant frequencies, not amplitudes or bandwidths.

The method used was a gradient search of three dimensions, using a small step in each direction to estimate the gradient. The error function is a sum of the discrepancies of all the formant frequencies of interest. As a computer was to do the searching, the factor of radiation loss at the

lips was also made a variable, a fourth dimension to be searched.

The error function J in formant frequency estimation was:

$$J = \sum_{i} \log (F'i/Fi) \qquad (5.16)$$

where: i is the number of formants of interest;

Fi represents target frequencies from the Klatt data;

F'i represents resonance frequencies computed from the currect articulatory configuration.

The steepest descent method as implemented had the following form:

$$\underline{W}(n+1) = \underline{W}(n) - \mu \delta J/\delta \underline{W} \qquad (5.17)$$

$\underline{W}(n)$ is the vector representing articulatory position at iteration n;

the partial derivative $\delta J/\delta \underline{W}$ was defined in equation (5.7) on page 5.15.

μ is a real scalar value controlling rate of adaptation.

A simple estimate of the derivative was used. For each element w_i of $\underline{W}$, the estimate was:

$$\delta J/\delta w_i = (J(w_i+\epsilon) - J(w_i)) / (w_i+\epsilon - w_i) \qquad (5.18)$$

meaning that a step ϵ was taken for each articulatory dimension, the acoustic consequences were computed, and the difference was the estimate of the gradient.

The next step in the iteration was then:

$$w_i(n+1) = w_i(n) - \mu/\epsilon (J(w_i+\epsilon) - J(w_i)) \qquad (5.19)$$

and the process iterated until a stopping condition was met.

The stopping conditions were:

(1) $|J|$ < minimum error. A value of $|J|<1.0$ gives an 8% error in three formants. Actual stopping values varied from 0.1 to 1.0, depending upon the phoneme.

(2) MAX (ABS(w(t+1) - w(t))) < minimum increment. The largest absolute change in any articulatory parameter is not causing significant motion in the search space. The standard minimum increment was 0.01.

(3) n > maximum allowed number of iterations, usually 100.

The biggest limitation to the adaptive search was that for a whole region of 'small' lip openings, the first formant frequency is zero. It was then impossible with a locally-determined gradient estimator to find a direction which improved F1 and reduced the error score.

This problem affected /u/ and the /u/-like sounds: /bvwm/; namely all the labials with low F1 and smallish lip opening. Manual intervention was used to find a lip-opening and radiation loss which give an adequate F1, at which point the problem reduces to two dimensions (just the two tongue factors once lip opening and loss are determined) and so it was possible to find articulatory values for /ubvwm/ by hand.

An unsolved problem was /r/, which in the Klatt table has a very low third formant frequency which the tongue factor constraints simply do not allow.

Although in the end an automatic search rather than a manual search of the contour plots was adopted, it was useful to have the plots in order to understand the space which the adaptive search algorithm was exploring. Further, the plots allowed manual intervention to untangle the problem with labials, and also showed that the problem with /r/ was unsolvable.

5.6.3 Results

Figure 5.10 shows how closely vowel formant centre frequencies were matched using the optimisation technique. Formant bandwidths are not included in the figure (as they were in figures 5.1 and 5.4), because bandwidths were not computed. Visual inspection of the contour plot data of figure 5.6 had shown that bandwidths were very sensitive to articulatory position when using the lossless tube model. Consequently bandwidths were not considered in the implementation of the gradient search approach.

The loci in the FR vs BR tongue space for 17 vowel and consonant sounds are given in Figure 5.11.

The main result of the automated search (with manual intervention) is a table of tongue and lip parameters which give a reasonable approximation to the Klatt data for most sounds. These data are presented in Table 5.3. These articulatory parameters were used to produce the synthetic speech for Experiments IV and V, Chapters 9 and 10, on intelligibility and naturalness.

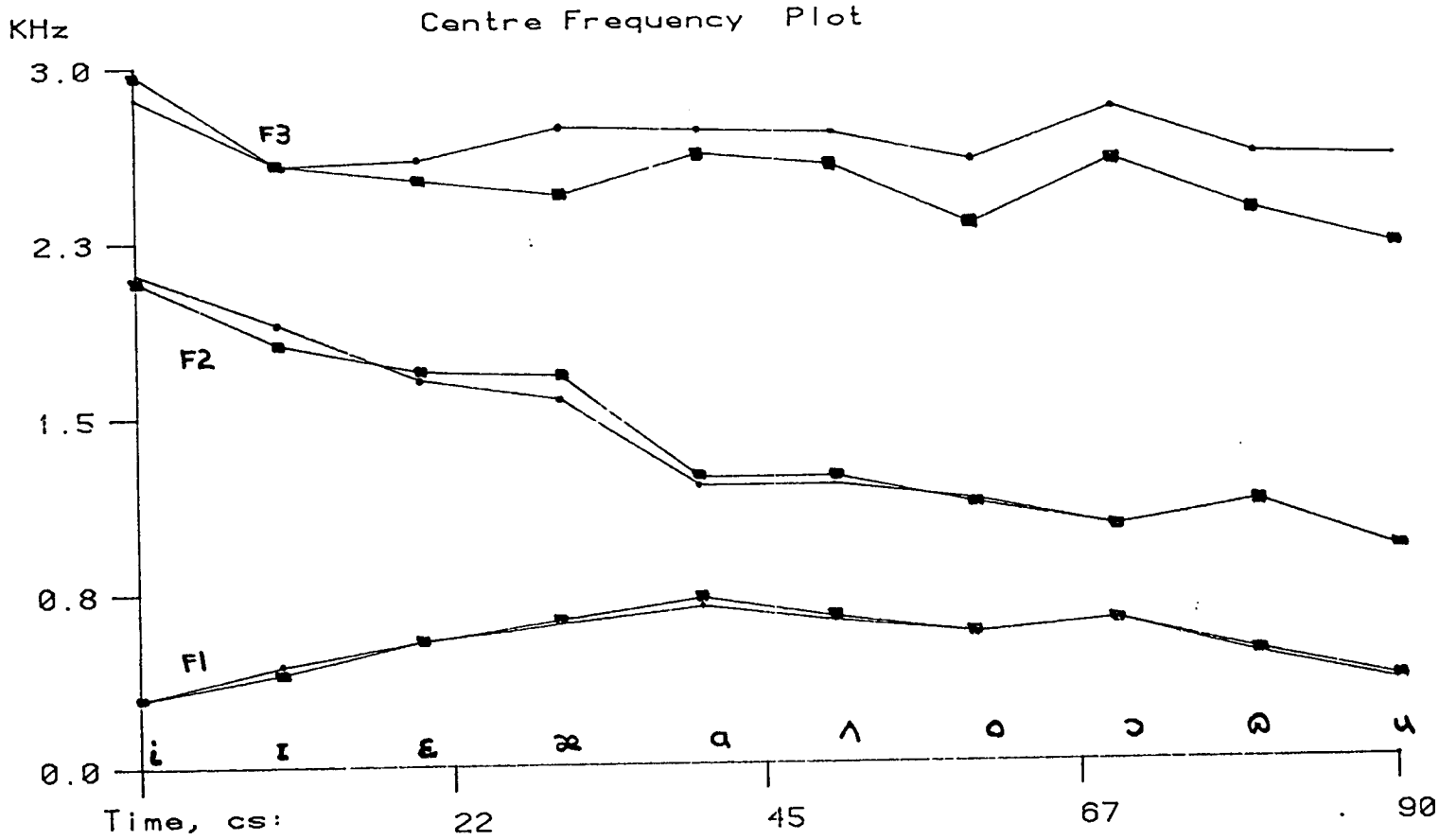


Figure 5.10: First three formant frequencies for nine vowel sounds, as determined using the gradient search method. Articulatory parameters (•) vs original Klatt series resonance data (▪). Plotted in frequency vs time as a stylised spectrogram.

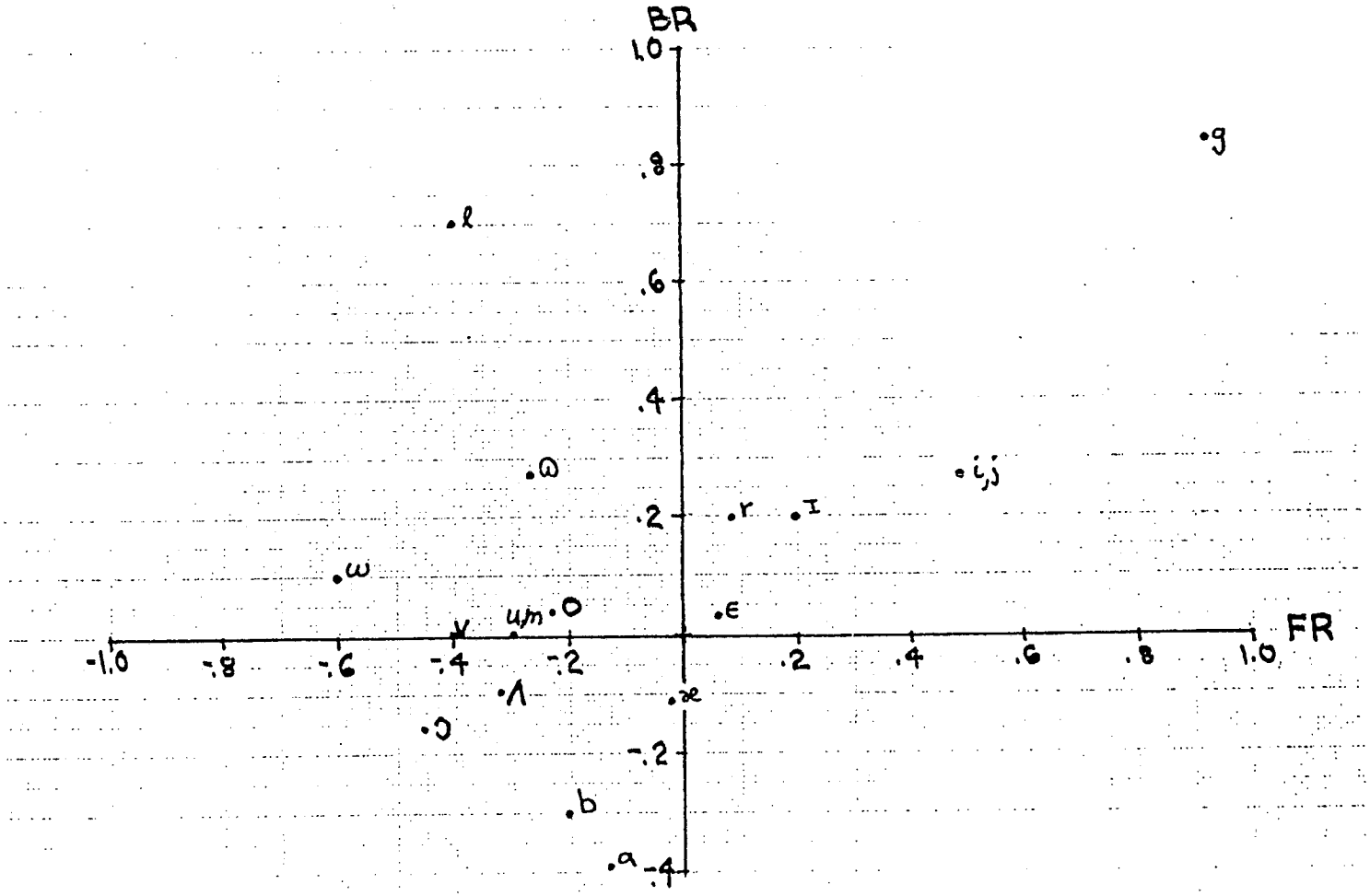


Figure 5.11: Loci in FR vs BR plane of 11 phoneme target values, as determined using the gradient search method.

Table 5.3: Articulatory parameters for approximating the (modified) Klatt phoneme target data. F1-F5 are formant frequencies. T = target data, E = estimate produced by the listed articulatory values. The last four columns give the four articulatory parameters: FR = tongue front raising; BR = tongue back raising; LIP = lip opening in cm^2; LOSS = reflection coefficient at lips.

		F1	F2	F3	F4	F5	FR	BR	LIP	LOSS
vowels										
i	T	290	2070	2960	3300	3850	0.485	0.270	1.827	0.310
	E	291	2102	2855	3650	5000				
I	T	400	1800	2570	3300	3850	0.198	0.199	1.863	0.392
	E	425	1893	2569	3661	4720				
e	T	530	1680	2500	3300	3850	0.058	0.029	2.036	0.3776
	E	532	1646	2595	3675	4653				
æ	T	620	1660	2430	3300	3850	0.020	-0.110	2.728	0.3351
	E	603	1571	2713	3732	5000				
a	T	700	1220	2600	3300	3850	0.130	-0.387	2.211	0.375
	E	667	1187	2700	3728	4548				
^	T	620	1220	2550	3300	3850	0.321	-0.099	1.845	0.397
	E	606	1185	2697	3688	4479				
O	T	540	1100	2300	3300	3850	-0.234	0.043	1.32	0.40
	E	540	1114	2567	3638	4467				
o	T	600	990	2570	3300	3850	0.450	-0.156	1.732	0.397
	E	598	990	2785	3702	4435				
Ω	T	450	1100	2350	3300	3850	0.266	0.269	1.134	0.399
	E	456	1099	2583	3648	4472				
u	T	320	900	2200	3300	3850	-0.30	0.00	1.00	0.40
	E	337	889	2578	3626	4432				
approximants										
w	T	290	610	2150	3300	3850	-0.60	0.10	1.00	0.40
	E	297	655	2847	3659	4406				
j	T	260	2070	3020	3300	3850	0.485	0.270	1.827	0,310
	E	291	2102	2855	3650	5000				
r	T	310	1060	1380	3300	3850	0.090	0.195	0.926	0.392
	E	316	1304	2355	3604	4542				
l	T	310	1050	2880	3300	3850	-0.40	0.70	1.00	0.40
	E	304	1097	2789	3811	4468				

fricatives									
f	T	340	1100	2080	3300	3850	-0.20	0.00	1.00	0.40
	E	355	975	2510	3616	4449				
v	T	220	1100	2080	3300	3850	0.00	-0.40	1.00	0.40
	E	232	1001	2439	3636	4442				
θ	T	320	1290	2540	3300	4900	0.123	0.529	1.029	0.269
	E	320	1204	2532	3695	4563				
ð	T	270	1290	2540	3300	4900	0.123	0.529	1.029	0.269
	E	320	1204	2532	3695	4563				
s	T	320	1390	2540	4000	5000	0.057	0.562	1.236	0.142
	E	316	1486	2494	3712	4656				
z	T	240	1390	2540	4000	5000	0.057	0.562	1.236	0.142
	E	316	1486	2494	3712	4656				
S	T	300	1840	2750	2500	4900	0.481	0.205	1.38	0.267
	E	307	2059	2496	3616	4802				
Z	T	240	1840	2750	2500	4900	0.659	0.184	1.15	0.30
	E	238	2081	2501	3600	4771				
stops										
p	T	400	900	2150	3300	3850	-0.30	0.20	1.00	0.40
	E	405	930	2588	3634	4452				
b	T	200	900	2150	3300	3850	-0.20	-0.30	1.00	0.40
	E	212	884	2537	3650	4415				
t	T	400	1600	2600	3300	4900	0.059	0.347	1.543	0.350
	E	394	1856	2457	3667	4646				
d	T	200	1600	2600	3300	4900	0.791	1.296	1.496	0.436
	E	146	1580	2886	4207	5000				
k	T	300	2400	2850	3300	4900	0.636	0.129	2.20	0.183
	E	300	2033	3199	3683	5000				
g	T	200	2400	2850	3300	4900	0.916	0.845	2.293	0.213
	E	149	2453	3236	4021	5000				
nasals										
m	T	300	900	2150	3300	3850	-0.30	0.00	1.00	0.40
	E	337	889	2578	3626	4432				
n	T	300	1600	2600	3300	3850	0.097	0.456	1.194	0.303
	E	317	1747	2393	3669	4661				
N	T	300	2400	2850	3300	3850	0.605	0.156	2.11	0.303
	E	299	2042	3061	3659	5000				

Chapter Six: Parameter interpolation in speech synthesis. Experiment I

6.1 INTRODUCTION

This chapter describes a comparison of the interpolation properties of six types of speech synthesizer parameters.

1 parallel resonance
2 serial resonance
3 prediction coefficients
4 reflection coefficients
5 area functions
6 articulatory parameters

The six synthesizers can be made to produce identical steady-state sounds (targets), but interpolation paths between targets will differ. Each synthesizer was tested on nonsense words spanning a wide range of parameter variation. For each parameter type, linear interpolation was used to determine a path between target values. The resultant data were then converted to formant values and plotted as a spectrographic (frequency vs time) representation. Small differences in formant frequency (vs linear transitions of formant frequency and bandwidth) were common, and there were some quite large differences in formant bandwidths. Each type of synthesizer parameter tended to exhibit characteristic path differences. Quantitative analyses of these formant path differences were also performed, showing that average path differences tended to exceed the Just Noticeable Difference for steady formants. Finally, the problems of a formant description versus an articulatory description are discussed.

An expanded version of the material in this chapter (including some sections of Chapters Three and Seven) has been provisionally accepted for publication (Wright and Elliott, in preparation).

6.1.1 The Transition Problem

An electronic speech synthesizer is essentially a parametric representation of speech: the synthesizer consists of a set of parameters, usually representing some articulatory or acoustical model of speech production. A particular set of parameter values determines a particular sound quality. A sequence of sets of these target values can then be used to produce continuous speech, providing there is some method for defining parameter values between targets. The most usual method is simply to interpolate parameter values between targets. (Holmes et al, 1964; Rabiner, 1968; Klatt, 1980a).[1]

In an effort to improve the intelligibility and naturalness of synthetic speech, considerable attention has been given to good target values (Holmes, 1979; Klatt, 1982). Attention has also been paid to the modification of target values as a function of phonetic context (Holmes et al, 1964; Klatt, 1976). The consideration of contextual effects raises the issue of coarticulation, which is an area of general phonetics interest (Lindblom, 1963; Ohman, 1967; Schouten & Pols, 1977, 1978, 1980; Stevens et al, 1966; Pickering, 1986). Finally, the attempt to produce a variety of synthetic voices involves the speaker-dependent aspects of voices and of phonemic targets, which in turn is of interest

---

1 - An alternative method of synthesis is the use of diphones (or dyads), in which the basic elements represent transitions rather than targets (Peterson et al, 1958; Estes et al, 1964; Stella, 1985). In this case intermediate values are already known, as they are part of each synthesis element; the price is a larger inventory of elements, of the order of N^2 rather than N, where N is the number of target values in the language. An interpolation problem for diphone synthesis is 'endpoint mismatch', where the ending value of one diphone does not match the starting value of the next, causing a discontinuity. The same problem arises in concatenation of parametrically coded words (Fallside and Young, 1978).

for speaker identification, and speaker normalisation for automatic speech recognition. Thus the problem of 'good target values' for speech synthesis has considerable overlap with problems in other areas of speech processing.

There has been rather less attention, however, to the problem of what happens between targets. The usual approach is simply to postulate a plausible method of interpolation, something that appears to give reasonable formant motion in spectrograms of the resultant synthesis. However, both decaying exponentials and increasing exponentials have been advocated as being the most plausible (Witten, 1982).

This chapter concentrates on the question of transitions between targets. It is not an investigation of formant motion in natural speech. Rather, formant motion in various types of synthetic speech has been investigated in detail, to show what happens to formants during synthesis.

6.1.2 Parameter Types

Between 1955 and 1970, speech synthesizers were either vocal tract models or formant synthesizers (Flanagan and Rabiner, 1973). With the advent of various linear prediction (LPC) parameters, the situation became rather more complicated: Makhoul (1975) mentions eight varieties of LPC parameters; Rabiner & Schaefer (1978) list nine.

Alternatives to formant synthesis are older than LPC, however. Analogue articulatory models were constructed in the early 1950's (Dunn, 1950; Stevens et al, 1953), though these were not used for the production of continuous speech because they were manually controlled. The early computer-controlled synthesizer of Kelly & Lochbaum (1962) was a vocal tract model using reflection coefficients, a parameter which re-emerged with LPC and eventually became the predominant parameter set for synthesis when it was incorporated in the first 'chip' for speech synthesis (Wiggins & Brantingham, 1978).

The increasing use of digital implementations for speech synthesis has increased the possibilities for parameter sets. Formant synthesizers are no longer necessarily either series or parallel, but can be both (Klatt, 1980b). Parallel synthesizers could in principle have variable bandwidth as well as variable frequency and amplitude. A variety of articulatory synthesizers are in use (Coker, 1968; Mermelstein, 1973; GALF Symposium, 1977; Liljencrants, 1985; Sondhi and Schroeter, 1987) and synthesis parameters have been extended to include aerodynamics (Scully, 1979; Shadle, 1986), complex models of excitation (Titze,1973, 1974), and of source/system interaction (Ishizaki & Flanagan, 1972; Holmes, 1973).

The development of new and more detailed parametric descriptions of speech has not, in general, deepened our understanding of the nature (or even the very existence) of targets and the problem of transitions between targets. Rather, it has increased the complexity of the problem by adding to the number of parameters to be considered.

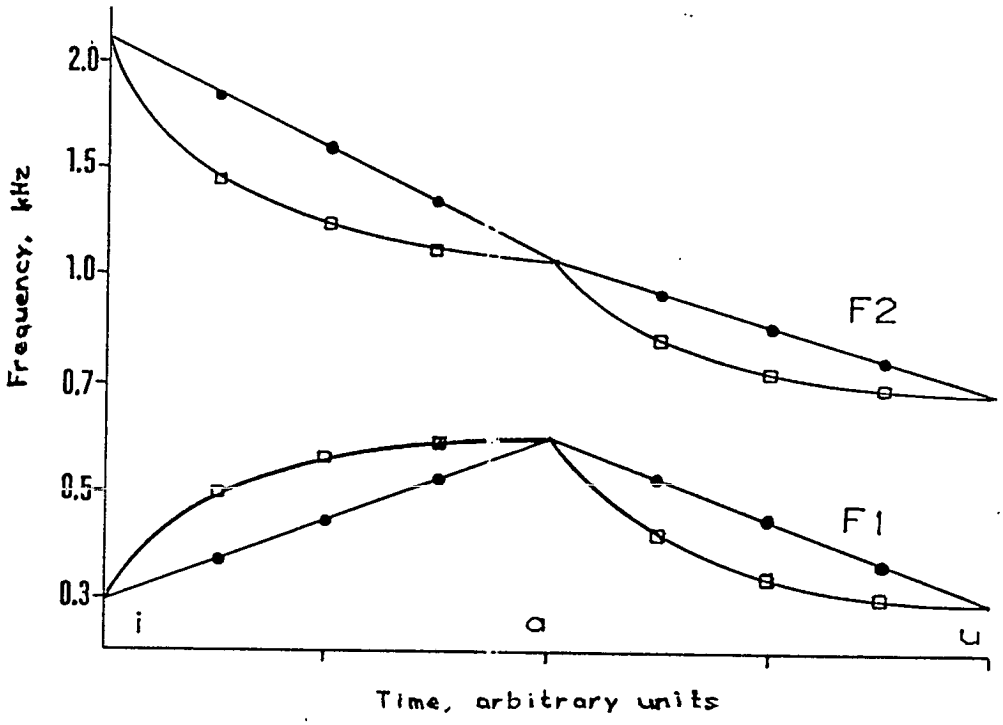
6.1.3 Interpolation Methods

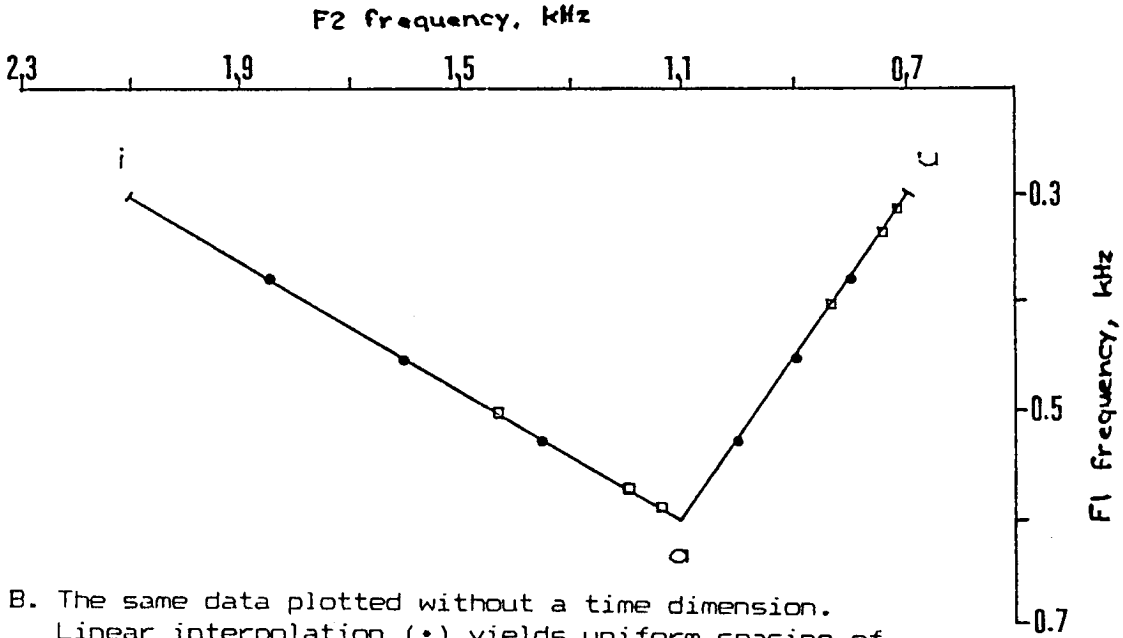
It was originally intended to study two aspects of synthesis parameters: parameter type and interpolation method. Various interpolation methods have been considered for speech synthesis: piecewise linear (Klatt,1980a; Holmes, et al, 1964); decaying exponential (Rabiner, 1969); increasing exponential (Lawrence, 1974). Another alternative is linear interpolation on a log frequency scale, which is either an increasing or decreasing exponential, depending upon the direction of the transition.

These interpolation methods appear to be quite different. However when viewed in terms of variation of one parameter against another (motion in parameter space or state space), all such methods produce the SAME path. This is shown in Figure 6.1 for linear and exponential interpolation of the

Figure 6.1. Interpolation in parameter space.

A. Parameter variation vs time, for linear (•) and exponential (▪) interpolation.





B. The same data plotted without a time dimension. Linear interpolation (•) yields uniform spacing of samples, whereas exponential interpolation (▪) yields points on the same path, but not uniformly spaced.

first formant (F1) and second formant (F2). It can be seen that the paths in the F1 vs F2 parameter space are identical; it is simply the time sampling (or rate of motion) along the path that differs. This result generalises to any interpolation method, so long as the same interpolation function is applied to all the parameters. Thus different interpolation methods share a common path in parameter space. It is only the interpolation of different types of synthesizer parameters that leads to distinct paths in a reference parameter space (such as formant frequencies and bandwidths).

Because interpolation method differences sample an identical path in parameter space, whereas different synthesizer parameters give rise to distinct paths, we concluded that interpolation method was a secondary consideration. Further, linear interpolation samples the paths in their own parameter space at points which are an equal distance apart (as is also apparent in Figure 6.1). Thus in this study only linear interpolation was used.

Six different types of synthesizer parameters were studied, using serial synthesis parameters as a common reference. The interpolation of the other types of parameters gives rise to separate paths through the ten-dimensional space of serial parameters (five resonance frequencies and five bandwidths). Although it would be interesting to investigate these path differences directly, traces in a ten-dimensional space do not lend themselves to graphical interpretation. So for examining the results of interpolation of the various parameters, we plot formant motion vs time, as in spectrograms. This gives a coventional representation, and still allows several parameters to be plotted simultaneously.

6.2 METHOD

The starting point was series resonance data taken from Klatt (1980b), Tables II and III. The Klatt data provide centre frequency and bandwidth targets for American English phonemes, and a subset of the data was used to specify requisite targets for each of eight nonsense words. Then for each of the five remaining synthesizer types (the five which do not use series resonance parameters), a conversion was performed to determine targets in that synthesizer's own parameter type. Linear interpolation was then used to determine intermediate values between target points. Finally, results were converted back to formant frequencies and bandwidths. These results were then plotted as stylised spectrograms, and quantitatively analysed.

Table 6.1. Phoneme targets for a series resonance synthesizer. Resonance frequency (Fn) and bandwidth (Bn) values in Hz; amplitudes (An) in dB. For all the phonemes, F4=3300, F5=3850, B4=500 and B5=700.

sound	F1	F2	F3	B1	B2	B3	A1	A2	A3	A4	A5
i	290	2070	2960	60	200	400	14	7	12	11	1
a	620	1220	2550	80	50	140	20	25	16	8	0
u	320	900	2200	65	110	140	16	6	-7	-19	-24
w	290	610	2150	50	80	60	18	9	-9	-28	-34
j	260	2070	3020	40	250	500	17	2	9	10	0
r	310	1060	1380	70	100	120	15	11	5	-28	-33
l	310	1050	2880	50	100	280	17	5	-2	-4	-13
v	220	1100	2080	60	90	120	12	1	-8	-23	-29
b	200	1100	2150	60	110	130	11	-3	-10	-24	-29
g	200	1990	2850	60	150	280	11	2	5	2	-8
m	480	1270	2130	40	200	200	13	17	-18	-34	-60

(/m/ has a resonance at 270 Hz, 100 Hz bandwidth, and no F5)

6.2.1 Speech Sound Categories

An effort was made to select a detailed and comprehensive sample of relevant speech patterns. The parameters studied represent only the 'system', not the source. Thus the prosodic (suprasegmental; stress and intonation) aspects of speech are not being investigated, but only the segmental aspect, a matter of speech sounds and their combinations. As the primary concern is parameter motion, those speech sounds representing maximum and minumum parameter values were selected. Combinations involving these extreme sounds will have the largest parameter motion.

To cover the possible combinations in a comprehensive but still efficient way the following strategy was used:

(1) Six categories were considered: vowel, approximant, fricative, stop, nasal, and affricate.

(2) Only extreme values (in terms of formant data) in each category were used. Thus the 18 vowels and diphthongs were represented by /i/, /a/, and /u/.

(3) For the stop, fricative and nasal groups, extreme formant values were mainly associated with front place of articulation (and with voicing in the case of stops and fricatives); thus only /v/, /b/, and /m/ were selected, plus /g/ to reach the extreme of the second formant range for stops. Affricates were discarded as they have the same transitions as the stop/fricative components from which they are synthesized.

(4) Each chosen consonant need not pair with each of the three vowels: two pairs span the extremes in most cases.

(5) Finally, three transitions can be represented on one plot, so sequences of up to four sounds can be analysed at once, reducing the required number of nonsense words and associated plots.

This strategy resulted in the following nonsense words:

/iaui/	Vowels; /iau/ maximum F1 motion, /ui/ maximum F2
/wiju/	Approximants; maximum F2 motion, F1 low
/waja/	Approximants; maximum F1 motion
/rilu/	Approximants; maximum F3 motion
/viva/	Fricatives; /vi/ has maximum F2, F3 motion; /va/ has maximum F1 motion
/iba/	Stop; /ib/ has maximum F2, F3 motion; /ba/ has maximum F1 motion
/agu	Stop; /ag/ has maximum F1 motion; /gu/ has maximum F2 motion
/ima/	Nasal; /im/ has maximum F2, F3 motion; /ma/ has maximum F1 motion

6.3 RESULTS

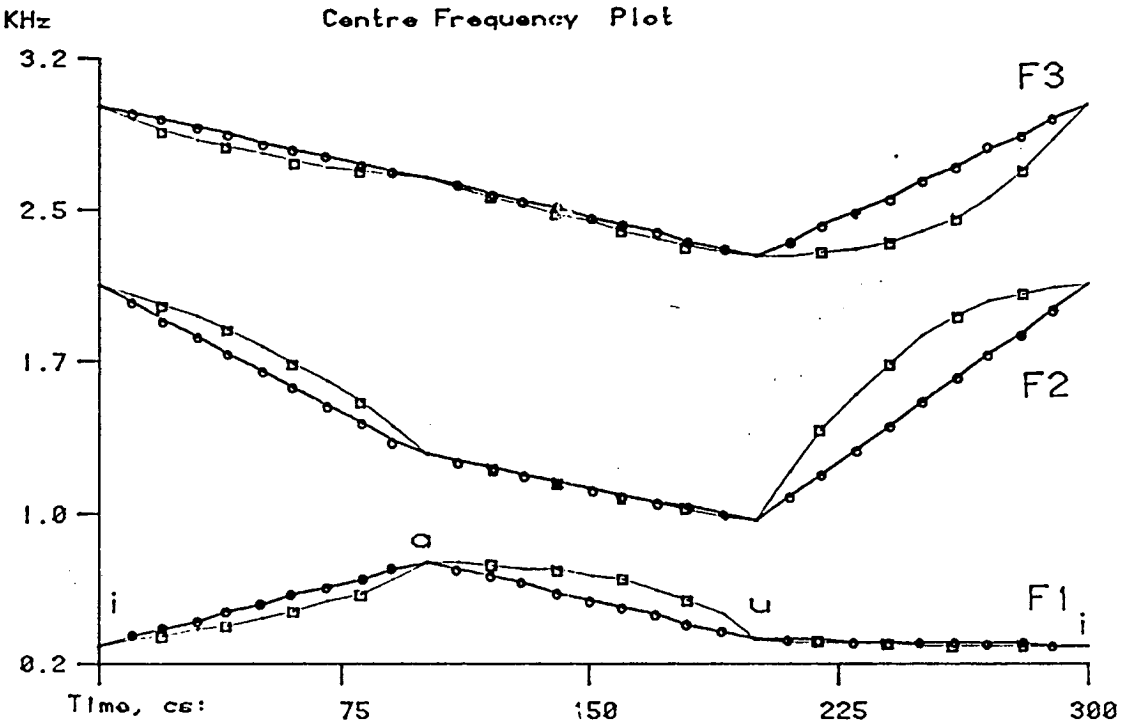
The results of the parameter motion study were analysed in two ways:

1 - graphically, using a spectrographic representation of resultant formant motion vs time.

2 - quantitatively, using measures of average formant differences.

6.3.1 Graphical Analysis

A typical result is shown in Figure 6.2, which compares a series resonance synthesizer with synthesis using an area function representation. Note that both synthesizers reach identical formant values at the target points, as an exact conversion is possible between centre frequency and bandwidth data (the series resonance parameters) and the area function data. The paths between targets are far from identical, however. General results for all the tokens and all the synthesizers are summarised in Table 6.2. Graphical results for all the nonsense words are given in Appendix 2.1.



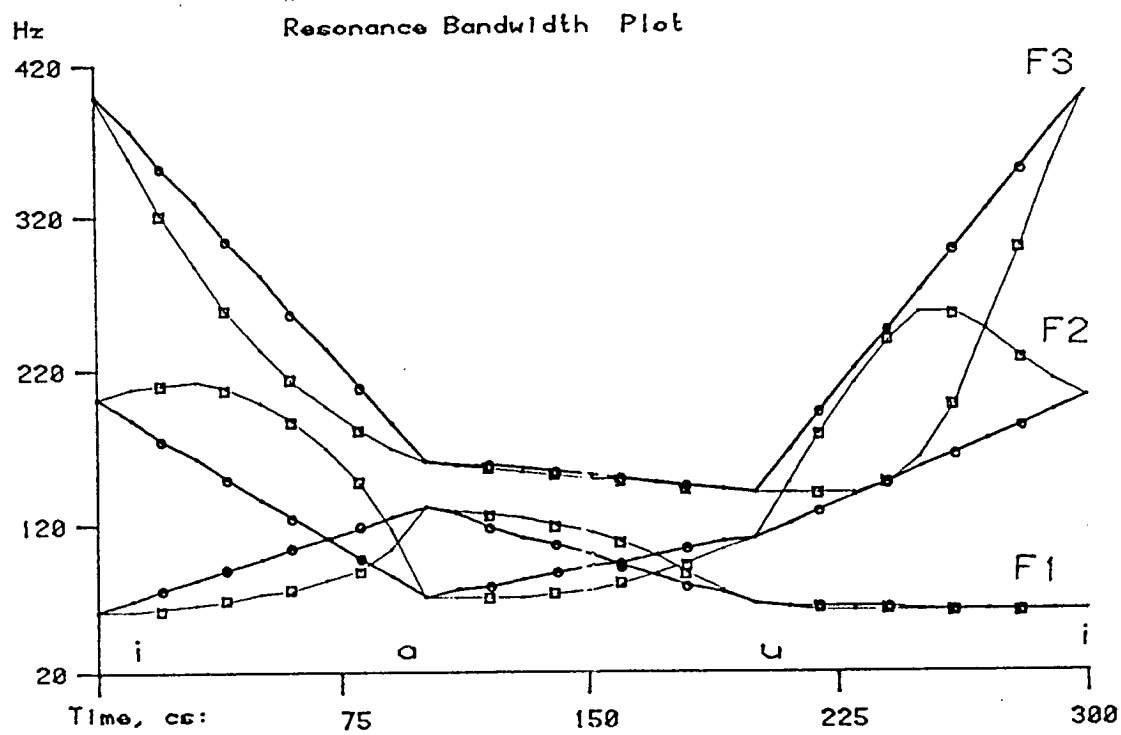


Figure 6.2. Formant trajectories resulting from linear interpolation of series resonance and direct form synthesizer parameters, for the word /iaui/.
Legend: series (•) ; direct form (▪) .

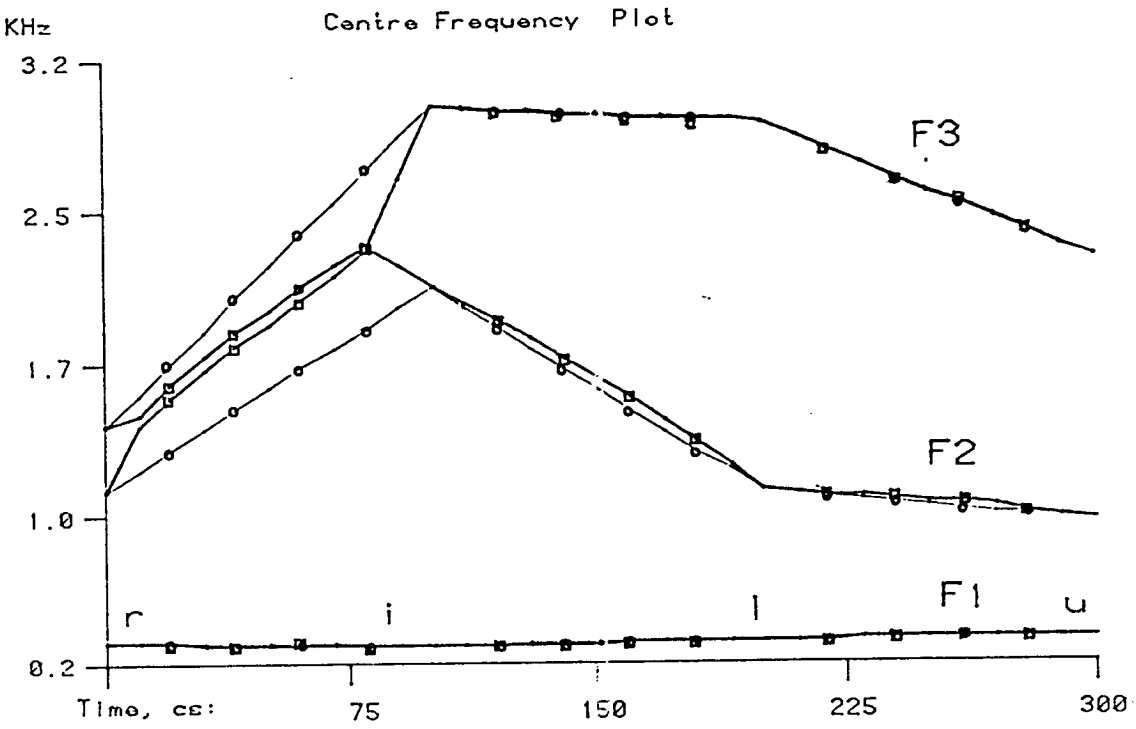
Table 6.2: Summary of main differences between parameter paths for a series form synthesizer as compared to five other types.

Series	(Reference path)
Parallel	Amplitude differences
Direct form	Instability (under certain conditions)
Reflection coeffs	Bandwidth differences
Area function	Frequency and bandwidth differences
Articulatory	Large bandwidth differences

The most marked effect of interpolation was observed for the direct form: during two of the eight nonsense words the resonance damping became negative, corresponding to instability in the steady state. This shows that potential instability can arise simply by trying to get from one stable position to another; this problem does not affect reflection coefficients and area functions. Figure 6.3 shows the direct form going unstable during /rilu/.

A stability problem can also occur with articulatory parameters, as the method used computes displacements about a neutral value for vocal tract diameter. There is nothing in the formulae which limits diameters (or areas) to positive values, and indeed reflection coefficients and hence direct form and cascade form resonance parameters can all be computed.

A less disastrous but more general effect is the observation that bandwidths are proportionally more affected than frequencies, as shown in figure 6.4. In the original series resonance data the upper two formants (F4 and F5) did not vary. These fixed resonances remain fixed for interpolation in the direct form. They began to move slightly when reflection coefficients were used, and were very much affected when the representation was in terms of area functions.



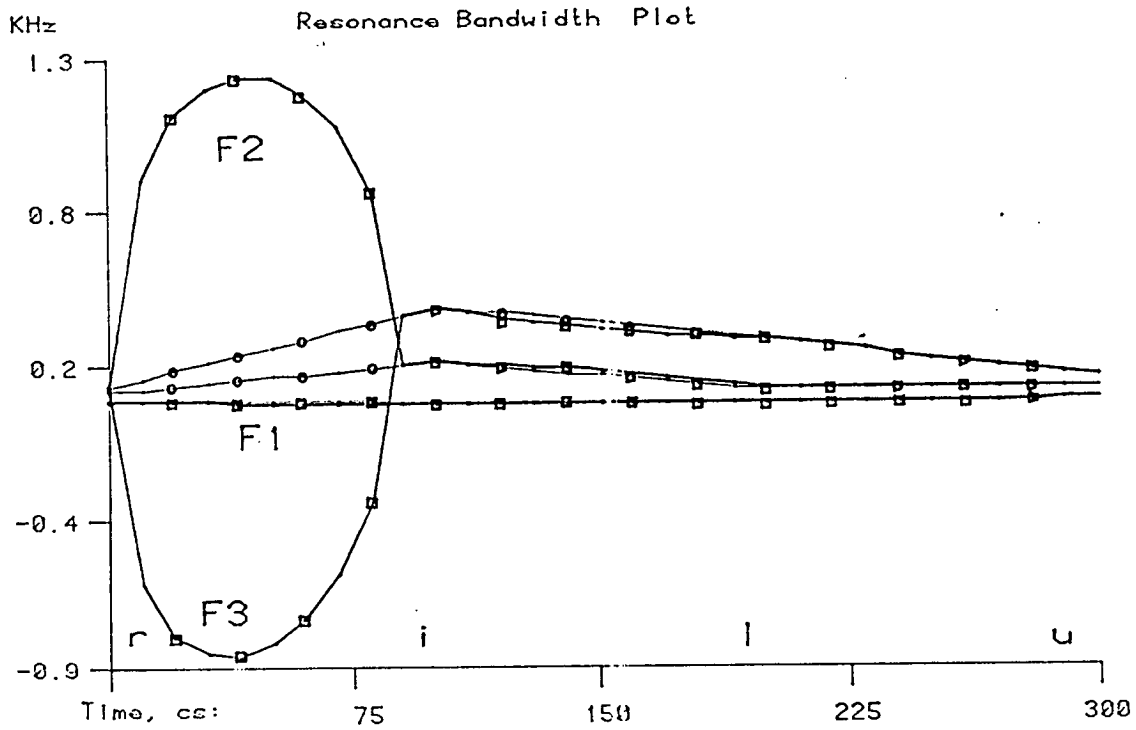
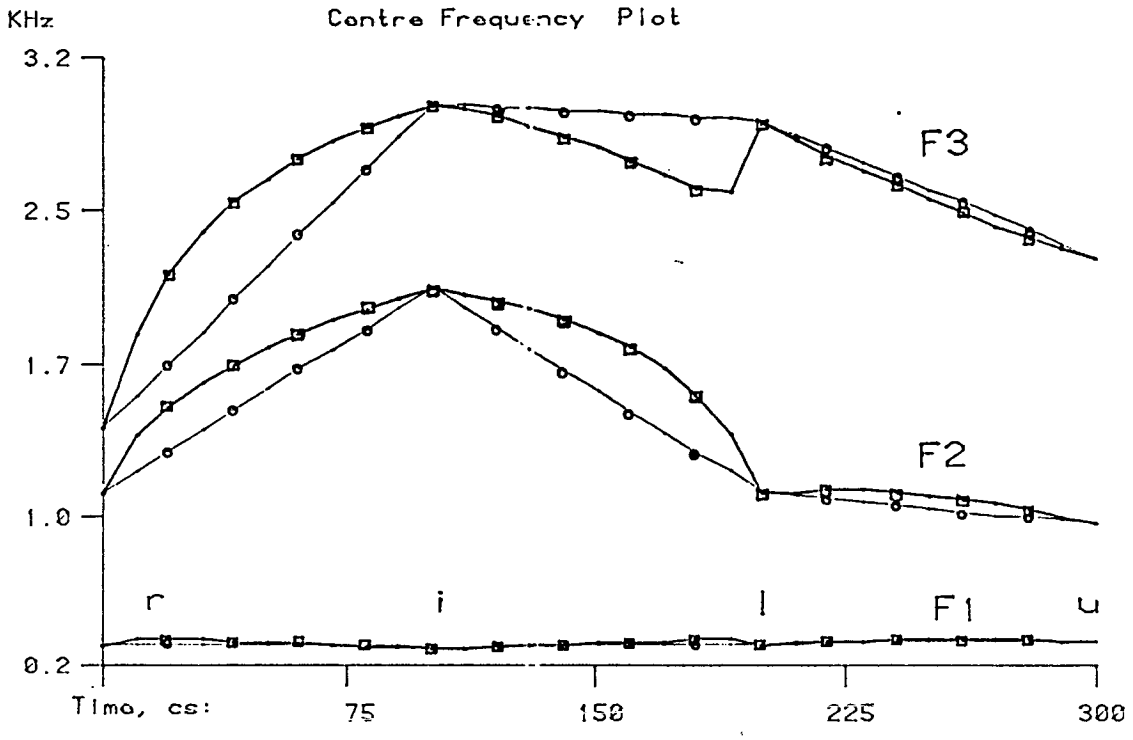


Figure 6.3. An instance of instability during a transition for direct form coefficients. The word is /rilu/.
Legend: series (•) ; direct form (▪) .



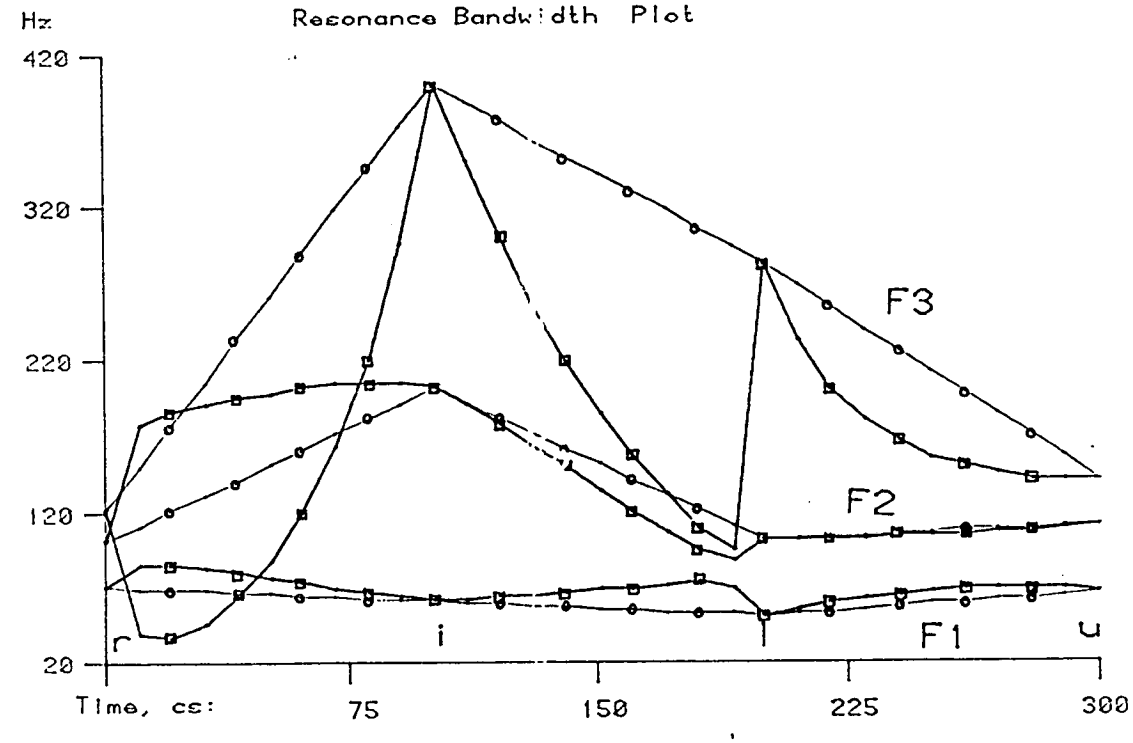


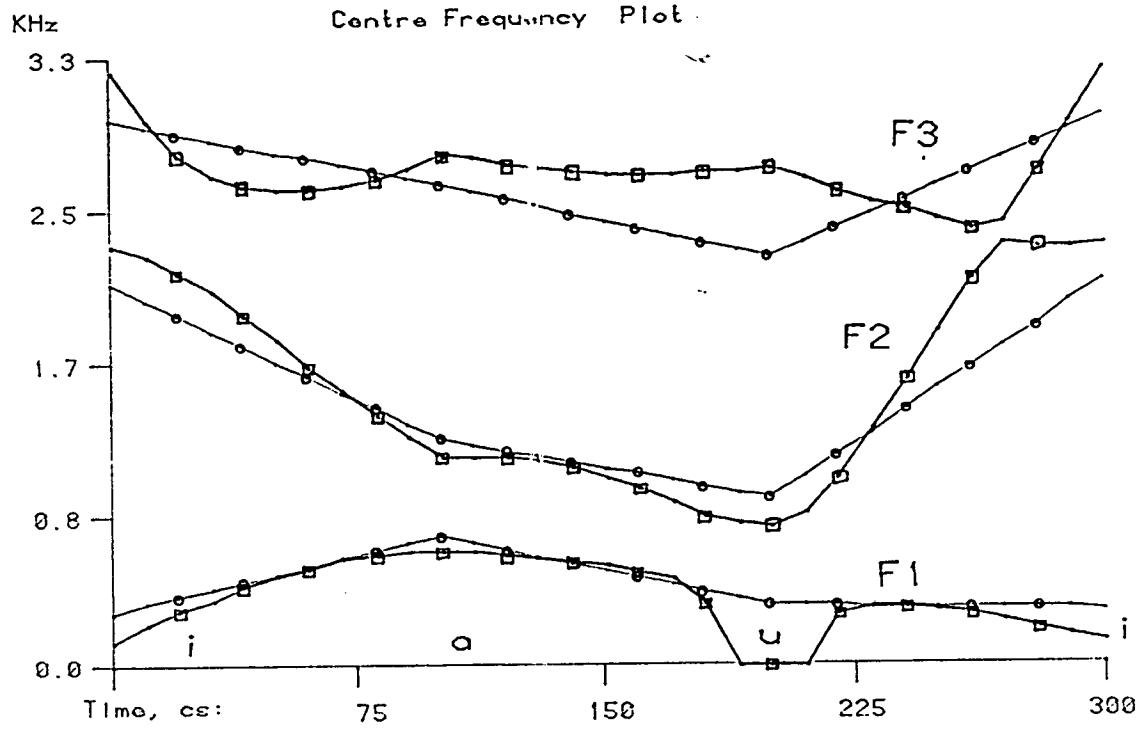
Figure 6.4. Transitions for area function synthesis , compared with series resonance synthesis, for the word /rilu/. The differences are larger for bandwidth than for centre frequency. Legend: series (•) ; area function (▪) .

Finally, the method used to produce articulatory parameters was based only on formant frequencies, not formant amplitudes or bandwidths. It is perhaps not surprising that the resultant vocal tracts had formant bandwidths which were very variable, and not a good match to the Klatt data even at the target points. Thus large overall bandwidth differences were observed, as seen in Figure 6.5.

As part of the investigation of articulatory parameters reported in Chapter Five, an attempt was made to discover the reason for the large bandwidth variations. The two-dimensional tongue parameter space (the factors labelled Front Raising vs Back Raising) was investigated for a fixed lip opening and loss. Formant values were then plotted as contours over this space (Figure 5.6), showing variation in resonant frequency and bandwidth of the first three formants. It was shown that the formant frequencies were reasonably well-behaved functions of the articulatory control parameters. Formant bandwidths, however, did not have such regular behavior. For all three formants there were areas in the Front Raising vs Back Raising space where bandwidth changed very rapidly for small changes in articulatory parameter value.

6.3.2 Quantitative Analysis

Quantitative results are presented in Table 6.3. For all eight nonsense words, the series resonance transitions were compared to transitions for the five other parameter sets. The absolute differences were computed, and the table gives the average over the eight nonsense words. Data for individual words are given in Appendix 2.2. Differences are presented as a percentage of the series resonance values. The minimum detectable change (Just Noticeable Difference =JND) for steady formant frequency is in the range 3-5% (Flanagan, 1965). The comparison of synthesis parameter variation with psychophysical JND values follows a procedure used by Shadle and Atal (unpublished).



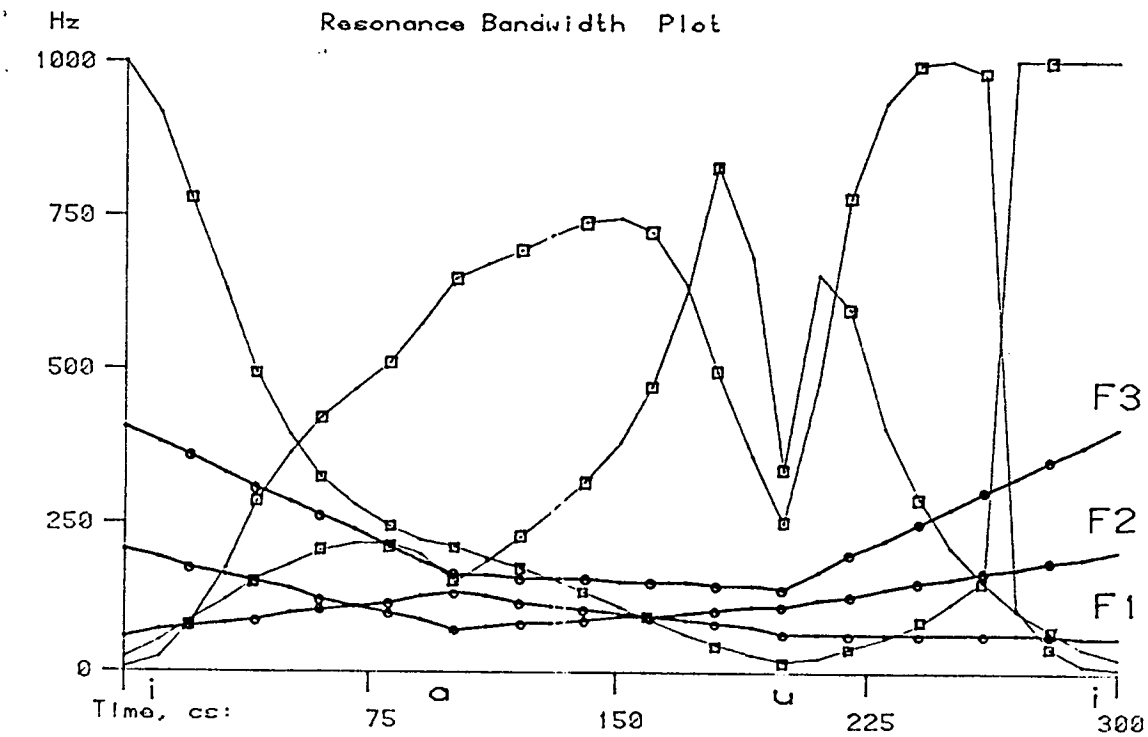


Figure 6.5. Results of linear interpolation of ladefoged et al (1978) tongue parameters, for the word /iaui/. Formant centre frequencies are reasonable (though F1 becomes real on /u/); formant bandwidths are erratic.
Legend: series (•) ; articulatory parameters (▪) .

Because the significance of the bandwidth of a resonance depends upon the centre frequency of that resonance, the bandwidth data in Table 6.3b are given in terms of Q, the ratio of centre frequency to bandwidth

The transition path comparison for series vs parallel is a special case. There are no differences in formant frequencies, because in both the series and parallel cases these values are the result of linear interpolation between identical endpoints. Further, the parallel case has fixed bandwidths so it is not reasonable to directly compare bandwidths. Instead a comparison was made of amplitude differences at resonance. Then we computed the amount of bandwidth difference required to produce the observed amplitude difference, had a series synthesizer been used. Thus the parallel vs series amplitude differences were converted to an 'equivalent bandwidth difference' and these resultant data, in percentages, form the table entries.

The greatest differences between the series and parallel cases occur between resonances, rather that at the resonance frequencies themselves (Holmes, 1982; Klatt, 1980b). These differences result from the system function zeroes introduced by the parallel configuration. The spectra in Figure 6.6 show two examples, first of a moderate spectral difference at a target value (the vowel /a/), and second a more dramatic difference along the transition path between /a/ and /u/. In this second case a parallel-case zero was near the unit circle, causing a 40 dB notch in the spectrum, and moving the DC response down nearly 25 dB. Computation of average spectral difference (ASD) was therefore made (Table 6.4), simply by summing the absolute differences (in dB) between log power spectra. The spectra were produced by evaluation of the parallel and series case system functions at frequencies which were approximately equally spaced on a perceptual scale (linear spacing with a 50 Hz interval below 1KHz; 1/15 octave spacing from 1KHz to 5KHz). Without ASD scores, a significant difference between the series and parallel case would have been neglected.

Table 6.3: Average absolute difference between series resonance transition paths and paths for the five remaining parameter types.

A: Centre frequency differences, percentage of the series value.

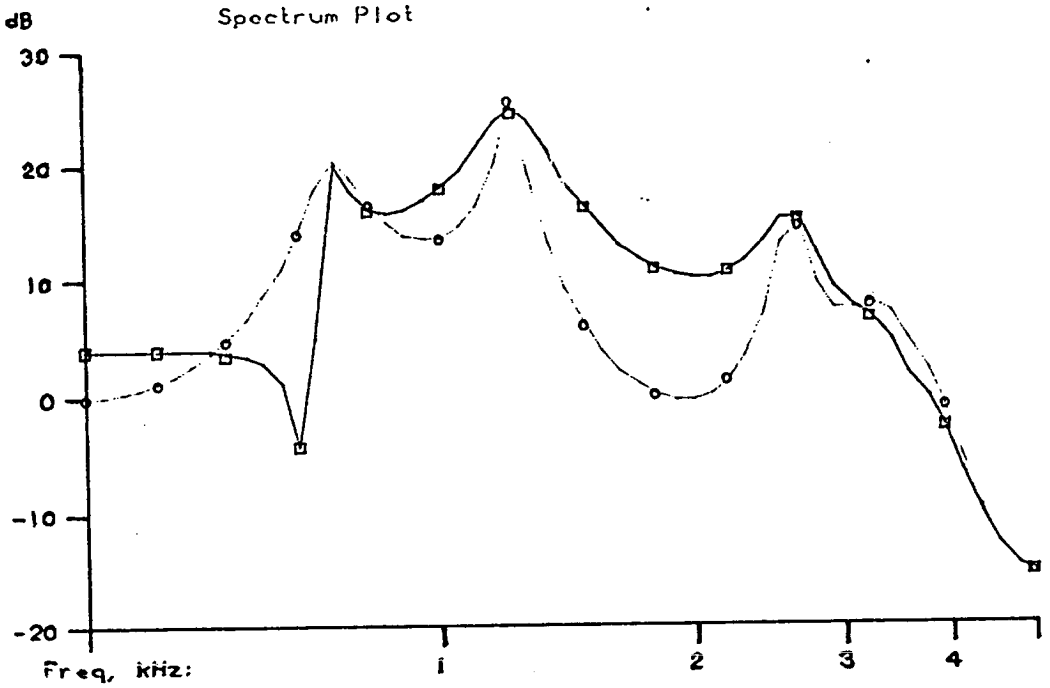
Synthesizer Type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Parallel	0.0	0.0	0.0	0.0	0.0	0.0
Direct form	8.7	7.5	3.6	6.6	0.1	4.0
Reflections	16.5	4.2	2.2	7.7	0.3	4.7
Area func.	11.8	11.0	3.7	8.8	0.7	5.6
Articulatory	42.7	22.4	18.8	28.0	20.0	24.8

B: Differences in Q, as a percentage of the series form value.

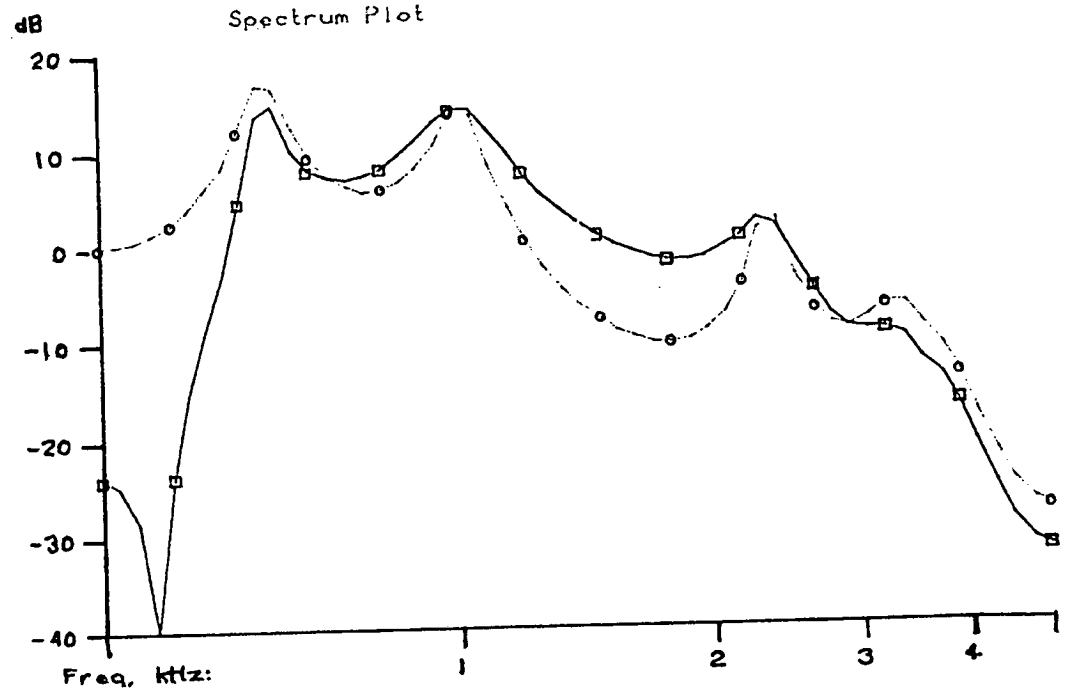
Synthesizer Type	Q1	Q2	Q3	Ave Q1-Q3	Ave Q4-Q5	Ave Q1-Q5
Parallel	25.6	83.6	85.8	65.3	34.7	52.3
Direct form	9.0	21.4	36.8	22.4	0.2	13.6
Reflections	9.5	28.8	29.8	22.7	4.8	15.5
Area function	11.0	18.2	36.5	21.9	6.6	15.8
Articulatory	69.4	2660	1950	1560	5800	3250

Table 6.4: Average spectral distance (in dB) between the series resonance synthesizer and five other types. The average was taken over all spectra for seven words (see text).

Synthesizer Type	Average Spectral Distance (ASD)
Parallel	3.60
Direct form	3.55
Reflections	2.50
Area func.	3.38
Articulatory	13.92



A. A rather minor difference for the sound /a/, caused by a low-frequency zero of moderate bandwidth.



B. A much greater difference along the transition between /a/ and /u/, caused by a narrow-bandwidth spectral zero.

Figure 6.7. Spectra for series (•) vs parallel (▪) synthesis.

6.4 DISCUSSION

A direct comparison of synthesizer parameters (against series synthesizer centre frequency and bandwidth parameters as a reference) and the effects of their linear interpolation upon formant transitions reveals the overall effects summarized in Table 6.2. The parallel form has identical centre frequency motion, but has the largest amplitude differences (except for the articulatory parameters). Although parallel formants move on exactly the same paths as series formants, the overall spectral differences are just as large as for reflections and area functions (Table 6.4), and larger than for the direct form.

There appeared to be a tendency in the graphical results for area parameters to have larger path differences than was the case for reflection coefficients. This was not borne out in Table 6.3, however, which shows only slightly more centre frequency variation (averaged over the lowest three formants) for area parameters vs the result for reflection coefficients. But when just the second formant is considered, reflection coefficients had much smaller centre frequency difference than did area functions.

As a general summary, the two parameter sets which were related to series parameters by approximations (parallel and articulatory), yielded the largest path differences, both graphically and numerically. The remaining three sets (direct form, reflection coefficients, and area function) were quite similar. All three had average formant centre frequency differences which exceeded the JND for steady formants by about 50%. There was a tendency in our data for area functions to spread the discrepancies evenly across all the first three formants, whereas reflection coefficients concentrated the errors at the low frequency end of the spectrum.

6.4.1 Problems with Vocal Tract Models

Figure 5.6 in Chapter Five showed the consequences of manipulating a lossless tube model (loss only at the lips). In that case the manipulation was in terms of the Ladefoged et al tongue description parameters, though there is no reason to expect a dissimilar result for any similar reduced-dimensionality representation. The result of small changes in tube shape is often a large change in bandwidth. This showed up in Figure 5.6 as places where the contour lines get very close together, representing a rapid rate of change.

No such problem exists for formant frequencies: their surfaces are reasonably smooth and regular. Thus the problems experienced with finding articulatory parameters which would yield reasonable bandwidths as well as centre frequencies is a problem of the lossless tube model itself: it does not model losses explicitly, and hence bandwidths fall out in an uncontrolled fashion. But the lossless tube parameters (area function or reflection coefficients) can be exactly related to formant frequencies and bandwidths (ie losses). This effect might be called 'compensatory articulation': the lossless tube can indeed be manipulated to achieve any desired loss, but only by using a shape that is (1) different from the shape required if realistic losses were present in the tube; and (2) outside the subspace of vocal tract shapes allowed by articulatory constraints.

6.4.2 Problems with Resonance Models

A more general critique of synthesizer parameters should include broader considerations, such as simplicity, independence and physical interpretation of the parameters. The comparisons made in this chapter were with respect to effects upon formants. When speech synthesis is viewed in a more general way, there are problems with a formant representation:

(1) Formant parameters are used in synthesis as independent control variables, whereas physical formants are not independent. For instance, a change in tract length affects all the resonances simultaneously. Thus formants are not a minumum dimensionality set of control parameters.

(2) Similarly, a simple change in vocal tract shape produces a complicated set of changes in formant motion. This is well known. For example, Fant (1970; p84) shows how the simple motion of a constriction from one end of a tube to the other produces a sequence of changes in five formants. Figure 6.7 shows a related result in our data. The F3 value for /i/ in the Klatt data is rather higher than the F3 for the remaining vowels, because of the high F2. The interpolation in articulatory parameters shows the 'crossing' of the formants: the large forward cavity for /u/ is gradually reduced as the system moves toward /i/, eventually producing a resonance above that of the usual value for F3. For a series resonance synthesizer to produce a similar transition, either F2 and F3 would have to cross or they would need very abrupt changes in path (of the sort not ordinarily produced by any kind of interpolation).

(3) Finally, formants cannot be made subject to dynamic contraints in any simple fashion. Formants do not physically move; it is the tract which moves. Simple constraints concerning articulator position and velocity are related in a highly non-linear way to formant positions.

The same criticisms can be made for all the synthesizers studied, with the exception of the articulatory parameters. Traditionally articulatory models have not been widely used, partly owing to their usual complexity. The simple model evaluated in this study has no such problems, as it is merely a low-dimensionality approximation to the series resonance model or any of its exact all-pole equivalents. But use of this model did show up a problem with lossless tubes: once the area function is subjected to speech-like constraints, the lossless tube has inadequate control of formant

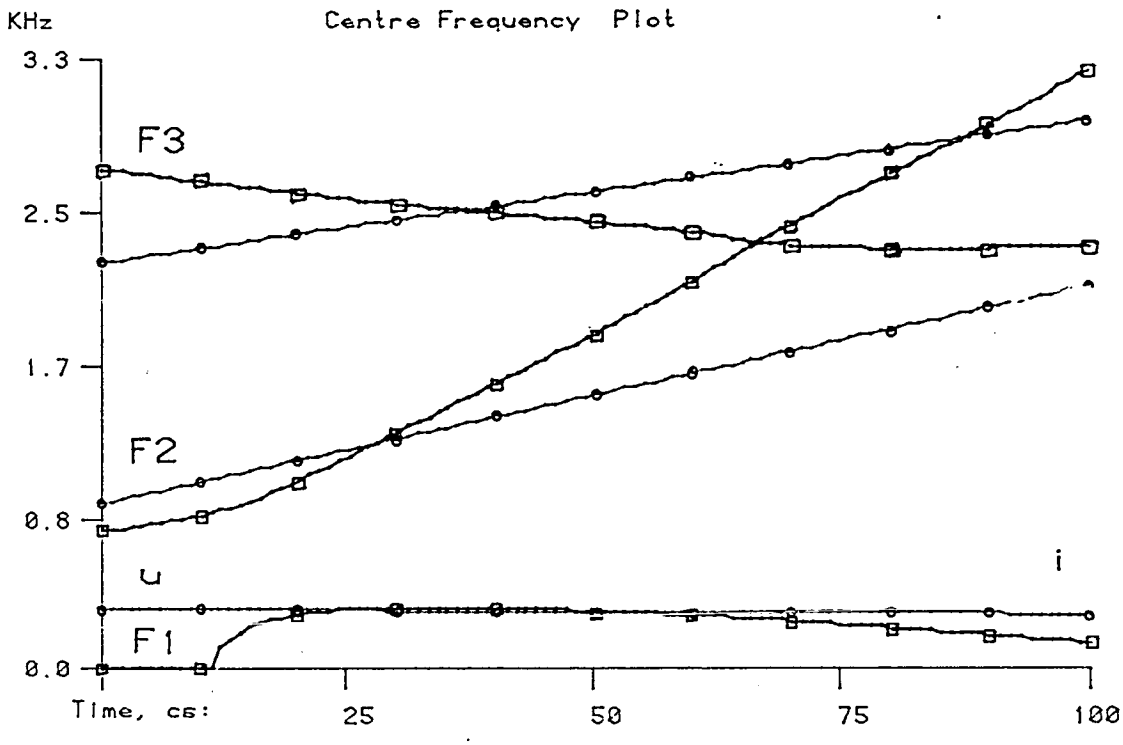


Figure 6.8. Formant 'crossover'. The transition from /u/ to /i/. The series synthesizer has rising F2 and F3 values, whereas the articulatory synthesizer keeps F3 fairly constant, and makes 'F2' take on a value above 'F3' for the /i/ sound.
Legend: series (•) ; articulatory parameters (▪) .

bandwidths. Thus articulatory constraints need to be applied to a more realistic model than that offered by the lossless tube approximation.

6.5 CONCLUSIONS

There are no target value differences for the four all-pole models; they only differ in formant motion implications of interpolation in the original parameter spaces. These differences are generally small, though informal listening tests have shown them to be (in some cases) supraliminal. Further subjective tests on more natural tokens are the subject of the experiments reported in the next four chapters.

The articulatory model is qualitatively different: it has three parameters rather than ten, it is (potentially) subject to dynamic constraints, and the two tongue factor parameters constrain the vocal tract shape. But the lossless tube model used in this study to implement a vocal tract proved inadequate for determination of resonance bandwidths.

This investigation has shown how formants actually move, given linear interpolation between target points for six sorts of speech synthesizer parameters. A related issue is the question of how formants or other parameters ought to move, given dynamic constraints on an articulatory model. That problem is considered in Experiment IV (Chapter Nine) on articulatory parameters and control.

Chapter Seven: Intelligibility comparison of synthesizer types. Experiment II

7.1 OBJECT

The first question asked about synthesizer parameters was whether they differed in ways which could be objectively measured. The question and its answers formed Experiment I, which was the subject of the previous chapter.

The second question is to determine to what extent these measurable differences affect human perception, specifically intelligibility and naturalness. Experiment II, the current chapter, is a consideration of intelligibility differences for five of the six synthesizers studied in Experiment I. The articulatory parameters are studied separately in Experiment IV, Chapter Nine. Another experiment (number III, Chapter Eight) takes up the issue of methods of interpolation of synthesizer parameters. Finally, the fifth and last of the experimental studies (Chapter Ten) examines naturalness for all the synthesizers considered in the first four experiments.

The object of Experiment II is to assess whether the interpolation path differences found in Experiment I have an effect on intelligibility. Do these differences cause some synthesizers to be easier to understand? And if so, by how much?

7.2 THEORY: WORD LISTS

Intelligibility measurement is the determination of the degree to which a speech signal is correctly understood. There are various types of intelligibility measurement, such as consonant and vowel recognition rates, or word or sentence level recognition scores. Also there are many sets of materials for the testing of intelligibility.

For the comparison of synthesizer parameters, there is little point in considering anything higher than the single word level. All the synthesizers have identical excitation signals, and can thus be expected to have indistinguisable prosodic features of stress, timing, rhythm and intonation. [These aspects of speech are described in Appendix Six: Basic Properties of Speech]. The area of potential difference is the segmental level: will the interpolation path differences affect phonemic labelling?

Sounds that are simpler than words might be used for intelligibility testing, for example nonsense syllables with a restricted structure. Consonant-plus-vowel (CV) tokens, having only one transition per test item, could be used if transitions are of most interest. Steady vowels could be used for stimuli if target values were the only consideration. But in using nonsense syllables, there is no place for the linguistic constraints on phoneme sequences which may contribute greatly to actual intelligibility of real speech. Thus measurements on nonsense syllables may be poor predictors of the utility of a speech signal for the purposes of human communication.

It is noteworthy that isolated consonants cannot be tested, as consonants cannot appear in speechlike sounds without an associated vowel. Thus CV or VC nonsense syllables form the simplest test material. Words are typically slightly more complicated in that they may have a CVC, CCVC, CVCC or even more complex structure. Words may thus have two or more transitions to be considered, and are the lowest linguistic level at which meaning can be conveyed. Sentences involve much more linguistic structure, but little more in the way of complication of phonological structure. Sentences (or just polysyllabic words) do present phoneme sequences that are impossible in single-syllable words, such as /kd/ in "backdoor". Thus there are transition possibilities which are not examined using monosyllables. However before studying transitions which cross a syllable boundary, one should look at the more integral transitions occurring within a syllable.

This study looks at responses to single-syllable words. Thus it is looking at a level where phonological constraints on phoneme sequences do apply, but below the level where prosodics are a consideration.

Having decided to use word tests of intelligibility, there remains considerable choice of test type. The tests can be divided according to the task of the listener. This task can be:

a - to 'say what word is heard', with no restrictions. (open response tests)

b - to pick a response from a small group of words on an answer sheet. (closed response tests)

7.2.1 Open Response Tests

The earliest formal test of intelligibility were of the open response type, the articulation tests of Fletcher and Steinberg (1929). These test were notoriously tedious and time-consuming, and required thoroughly trained crews of talkers and listeners (House et al, 1965, p158).

The main problems were a long learning period on the part of the listeners, and great variability in responses for the 'wrong' responses. The learning period was the time taken for the listeners to become familiar with the whole set of words being used. Only when the listeners had learned just what words might actually be presented did their performance reach a plateau. This learning period could take many hours, spread over many days.

Further, since the subject had no constraints upon selected responses, it was difficult to make sense of subject errors. The actual reported word was a combination of what the subject heard, familiarity with the word list, familiarity with a much larger list of words which sounded like the words in the word list, the relative frequency of occurrence of all

these words, and the willingness of the subject to report an unnatural or nonsense word if that was indeed closest to what was 'heard'. Thus a composite of learning and linguistic effects complicated interpretation of results.

One approach to sorting out this disorder was to at least ensure that the presented words were of uniform familiarity and perceptual difficulty, and that the distribution of speech segments followed that of general samples of English (the principle of phonetic balance). Many lists were developed over the 35 years following the original work of Fletcher and Steinberg:

Harvard PB lists (Egan, 1948)
CID W-22 lists (Hirsh et al, 1952)
CNC lists (Lehiste & Peterson, 1959)
Northwestern Univ lists (Tillman et al, 1963)

An interesting approach to constraining responses and also aiding learning of the word list was to use words in rhyming groups (Fairbanks, 1958). This procedure also focussed attention on one segment of a word, one of the consonants, because within a CVC group the vowel and one consonant were always identical. The test was called the Rhyme Test, because when the final consonant was the 'fixed' consonant the words did actually rhyme. When final consonants were to be tested, and initial consonants were fixed, the groups were just as constrained but in a sort of reverse rhyme. Subject learning time was accelerated, but not eliminated.

7.2.2 Closed Response Tests

The most important step forward in intelligibility testing was to tell the subject the answers! In actuality, to provide the subject with a short list of response possibilities (six words in the original study of House et al, 1965). This Modified Rhyme Test (MRT) virtually eliminated the problem of listener familiarity and learning time. With the answers provided, there was nothing to learn.

Also the closed set of response possibilities allowed errors to be accumulated and analysed. This was noted and performed in the original MRT. A later development was to ensure the selection of a response set according to a systematic procedure (Voiers, 1977). This approach attempted to ensure that all the 'incorrect' response options in a response set were equally attractive as choices. One means to achieve this end was to have each item differ in exactly one 'feature', such as voicing, manner or nasality. Thus response biases could be eliminated and the errors could be more readily interpreted. This Diagnostic Rhyme Test (DRT) was first developed by Voiers (1967).

The MRT used a response set of six, to maximise the efficiency of the test. Having six responses allows several perceptual decisions to be simultaneously tested. More than six items leads to difficulty in coping with simply reading the response lists. The DRT used a response set of two, which is less efficient but solves the problem of 'incorrect' alternatives of unequal attractiveness by only having a single 'incorrect' alternative. This pairwise arrangement then also allowed each word in the lists to be both a stimulus item and a response item.

The MRT and DRT have essentially replaced open-response tests for telecommunications purposes. The DRT has recently been used in the UK to evaluate synthetic speech (Pratt, 1986).

The MRT has also been used for evaluation of synthesis, notably by Pisoni (1979) as part of the large MIT project on synthesis. The MRT has recently been used in the UK by Faulkner (1987) at the IBM Science Centre, Winchester.

Another closed-response wordlist was developed at the Institute of Hearing Research in Nottingham (Foster and Haggard, 1979). This test chose a middle ground between the unbiased two alternative DRT and the efficient but unsymmetric six alternative MRT, and uses four alternatives. Very sophisticated evaluation has been made of response errors, to allow relating scores to specific parts of the

auditory spectrum. Thus the test is named FAAF, the Four Alternative Auditory Feature test. It is more efficient than the DRT, and more symmetric and less biased than the MRT. Also it is in British English; the DRT and MRT both require modifications for British.

The test is only of consonants, and a limited set of vowels is used, rather than the eight vowel environments used in the DRT. This reduction in attention to vowels allows the FAAF test to be about 60% shorter than the DRT, and still test every contrast twice as often (because of the 4x4 rather than 2x2 arrangement).

The considerations of efficiency, diagnostic potential, and use of British English led to the choice of the FAAF test for the purpose of itelligibility testing in these experiments.

7.3 SYTHESISER CONFIGURATION

In Experiment I, Chapter Six, only control data and their steady-state spectra were examined. No synthesis as such took place, and no synthesizer existed. There was only a facility for the creation and interpolation of control data.

It was decided to implement the simplest possible synthesis scheme. This was not just expedient, but also to provide a baseline system for comparison with any later sophistication. One requirement of the implementation was an unusual capability for synthesis systems: it needed to be controlled by a variety of parameter types: series and parallel resonance data, direct form, reflection coefficients, area function, and articulatory parameters.

One representation common to all six synthesizer types (and many others) was that their parameters could be converted to the polynomial coefficients of an unfactored system function, as discussed in Chapters Three to Five. Further, a system function H(z) can be immediately realised using a recurrence relationship. All that remains is

excitation and gain control, and scaling of the output to fit the D/A system to be used. Figure 7.1 shows a block diagram of the system. A 10 kHz sampling rate was used, and all calculations were in 32-bit floating point, up to the 12-bit coding of the output for digital-to-analogue conversion. Any exponential growth (from unstable filter coefficients) was limited to a factor of two.

The synthesis system is most notable for what it excludes:

No explicit zeroes; only the implicit zeres from the parallel configuration;
No nasal channel;
No shaping filters, separation filters, or pre-emphasis;
No source-system interaction; no variation of bandwidths to approximate the effects of open vs closed glottis;
No pitch-synchronous parameter updates;
No variable excitation pulse shape or duty cycle;
No variable mix of voiced and voiceless excitation.

Unlike more fully-fledged systems, where series vs parallel resonances are a striking difference, this system function implementation treats parallel resonances simply as the cause of a non-trivial numerator polynomial. All the other synthesis parameters considered have a constant in the numerator, while parallel resonances give rise to an eighth-degree polynomial. The denominator is in all cases tenth-degree. There is no possibility of introducing extra resonances or shaping or channels without changing the denominator degree, and thus destroying the interchangeability and formal equivalence between configurations. The system is 'bare-bones', but does ensure that the comparison of parameter types is not confused by extraneous issues.

7.4 STIMULI - SYNTHETIC FAAF WORDLIST

The FAAF test, its administration and scoring are documented in a manual (Foster and Haggard, 1984). The test consists of

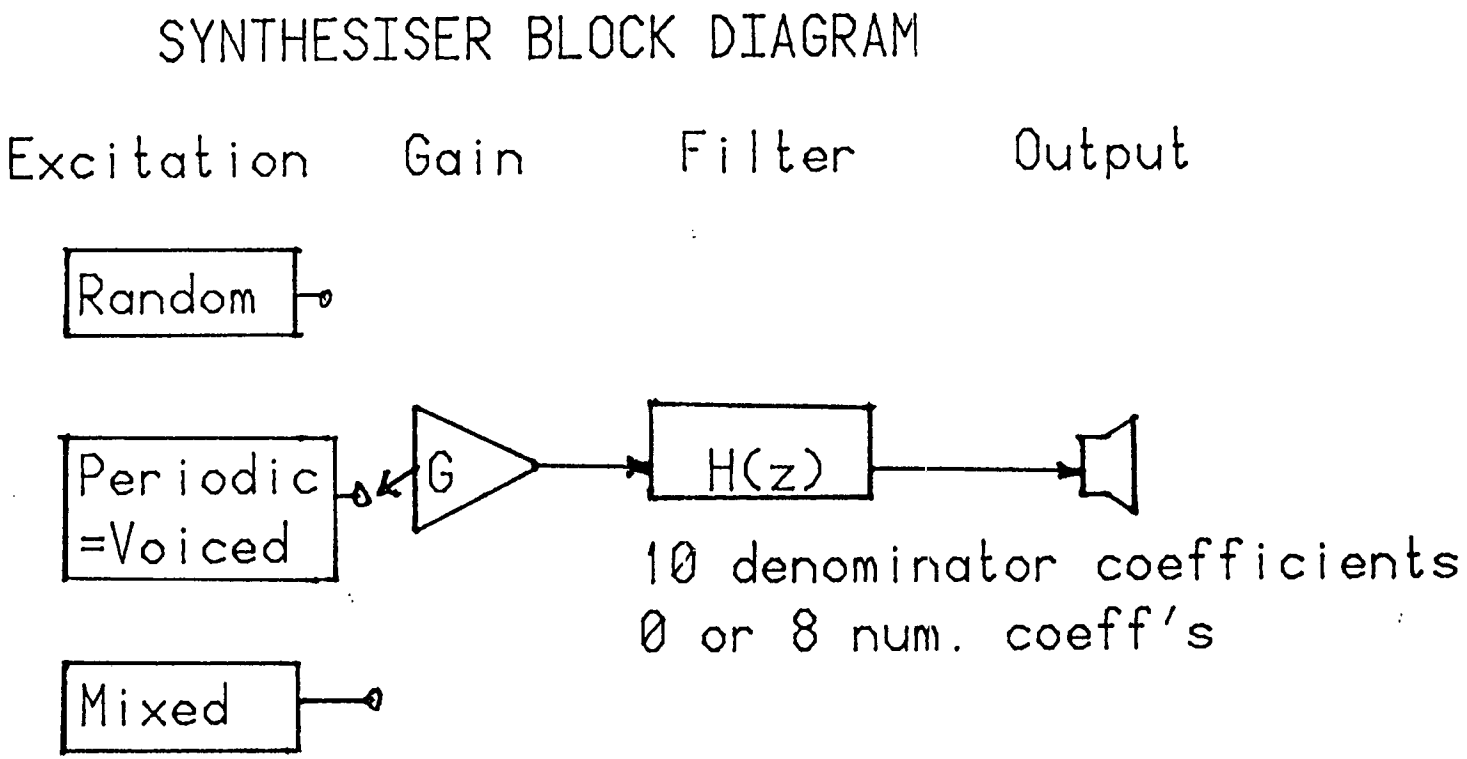


Figure 7.1: Block diagram of synthesizer.

80 test items and 25 practice items, as listed in Appendix 4. The words are all monosyllables, and mostly actual English lexical items. They have a basic CVC structure, though there are some clusters of consonants.

7.4.1 Target Data

Synthesis-by-rule of the FAAF items requires phonemic target values. A starting point are the published data of Klatt (1980). These values are reproduced in Table A4.1 of Appendix 4. Unfortunately this set of data is optimised for the Klatt series + parallel sythesiser. Only vowels and approximants (/wrlj/) are purely serial, and so the given data had to be modified to fit a 10th order all-pole system.

The modifications are discussed in the next sections, and the results are given in Table A4.2 of Appendix 4.

7.4.1.1 Approximations for fricatives and stop bursts

The Klatt synthesizer has a series path that is just the same as the standard five-resonance series synthesizer studied in Experiment One. But for stops, fricatives and nasals there is an additional set of six parallel resonances, and a path bypassing all the resonances. These six resonances (a1 to a6) plus bypass (ab) are controlled by a set of seven amplitude parameters. Klatt offers the following general guidance for use of these parameters:

a = amplitude in dB re nominal 'off' value of 0 dB.

a1 is always zero, because only front cavity effects are of interest (Stevens, 1972; Klatt 1980 p 981).

a2 is zero for consonants preceeding a front vowel, and about 60 dB for other vowels.

a3, a4 and a5 control the amplitude of the parallel resonance which has random or mixed excitation and has the same centre frequency and bandwidth as the same numbered resonance in the series path.

a6 is a special resonance at 4900 Hz (1000 Hz bandwidth) used just for /sz/.

ab is the bypass amplitude, the gain of a channel which bypasses all the parallel resonances and feeds the excitation directly through to the output. This is used for a flat (unresonated) spectrum.

The Klatt values for these parameters for unvoiced stops and fricatives are given in Table 7.1. Voiced fricatives and stops have the same parallel amplitudes as their unvoiced counterparts, but narrower series bandwidths. Klatt also singles out /d/ to have higher parallel amplitudes than for /t/, as shown in the Table 7.1.

Table 7.1: Klatt (1980) parallel control data for consonants.

Sound	a2	a3	a4	a5	a6	ab	Notes
f	0	0	0	0	0	57	Just ab = flat noise.
θ	0	0	0	0	28	48	Add noise at high end.
s	0	0	0	0	52	0	Just noise at high end.
p	0	0	0	0	0	63	Like /f/: flat only, no peaks.
t	0	30	45	57	63	0	Broad tilted spectrum, most energy at high end.
d	0	47	60	62	60	0	Less tilt, more gain than /t/.
k	0	53	43	45	45	0	Between /p/ and /t/. Broad flat spectrum, multiple resonant peaks, highest at low end.

The problem was to achieve something like the Klatt spectra within the limitations of a simple series synthesizer. The first approach was to use all the bandwidth parameters to try to achieve appropriate spectral changes. In particular, the 'ab' channel simply adds a signal with a flat spectrum. A simple approximation would be to widen all the bandwidths of the five resonances.

This approach of bandwidth widening produces spectra that look quite plausible, as shown in Figure 7.2. A whole set of similar data were generated for all the stops and fricatives, and bandwidths were manually altered until the desired spectra were obtained. Actual FAAF words were then made using this data and a trial intelligibility study was performed, using three subjects.

The resultant intelligibility was low. The chance level for raw scores on a FAAF test is 25% correct (because there are four alternatives), and results on normally-hearing subjects for natural speech are nearly 100% correct at signal-to-noise ratios (SNR) in excess of 5 dB. But these initial stimulus items with widened bandwidths had an identification rate of only 52% correct. This would correspond to the recognition rate on natural speech at a presentation SNR of only -8.5 dB. The recognition rate for FAAF wordlists changes at about 6% per dB between -10 and -2.5 dB, corresponding to 40% and 85% correct identification rates, respectively (Foster and Haggard, 1984, p19). This steep slope is typical of word identification tests, and indeed the classical Miller and Nicely tests (1955) also yielded a slope of 6% per dB in this range.

The problem with wide bandwidths for stop and fricative target data is a loss of formant definition. The standard theory of acoustic phonetics holds that formant transitions into or out of a vowel from a consonantal 'locus' are a major cue to consonant identity (Fry, 1979, p138). Especially significant are the first formant (which lowers to show an occluded vocal tract and therefore a consonant rather than a vowel), and the second formant which varies in target

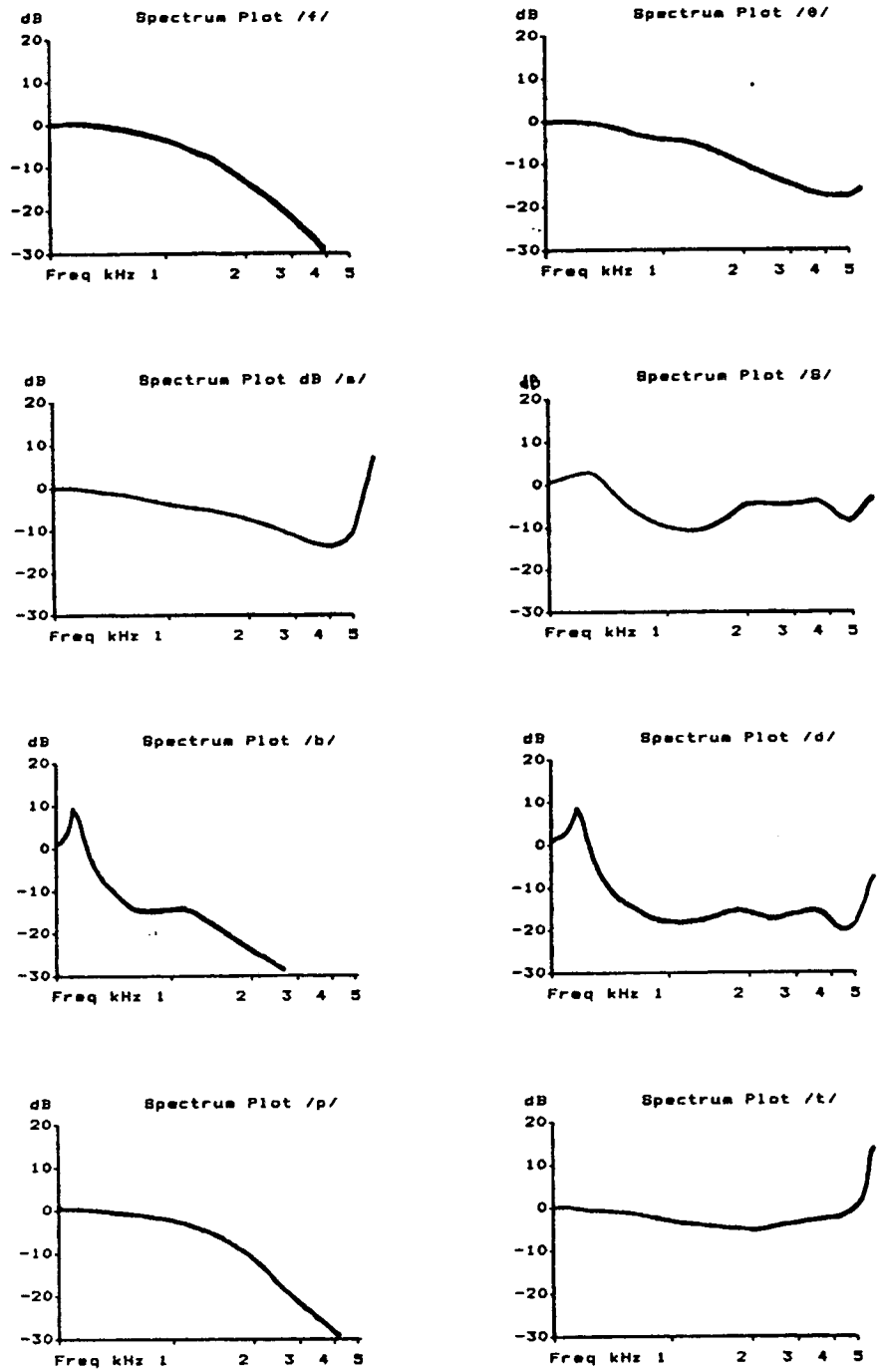


Figure 7.2: Example spectra for approximation of Klatt (1980b) serial-plus-parallel synthesis of consonants, using a tenth order serial synthesizer. All bandwidths were widened to approximate the effects introduced by Klatt's parallel set of resonances.

frequency according to where the vocal tract is occluded (place of articulation). Increasing all the formant bandwidths could be expected to reduce the place information in particular, with a consequent reduction in intelligibility.

Accordingly, a second attempt at modified target values was made, keeping all of the Klatt values for frequency and bandwidth for all three lower formants, and only using F4 and F5 to model the main effects of Klatt's parallel channels. Another set of data was produced, spectra were examined (Figure 7.3) and also informal listening tests were performed.

This process was repeated until the resultant stops and fricatives were at least plausible. Then another set of FAAF words were produced and tested. The results on two listeners jumped from 52% to 68%. In equivalent SNR terms this is just under a 3 dB improvement, but just in the critical region from -8 dB to -5 dB effective SNR, where the intelligibility vs SNR function is steepest (Foster and Haggard, 1984, p19-20).

Table 7.2: Sample all-serial control data for consonants, using only F4 and F5 modifications to the Klatt data.

sound	cf4	bw4	cf5	bw5	ampl	comments
f	3300	1000	3850	1000	45	Simulate ab with flat F4, F5
θ	3300	1000	4900	600	40	Lift high end
s	4000	2000	5000	300	50	Sibilant, hand trimmed
p	3300	2000	3850	2000	45	Simulate ab with flat F4, F5
t	3300	500	4900	500	45	Lift F4, F5
k	3300	1000	4900	1000	45	Intermediate case

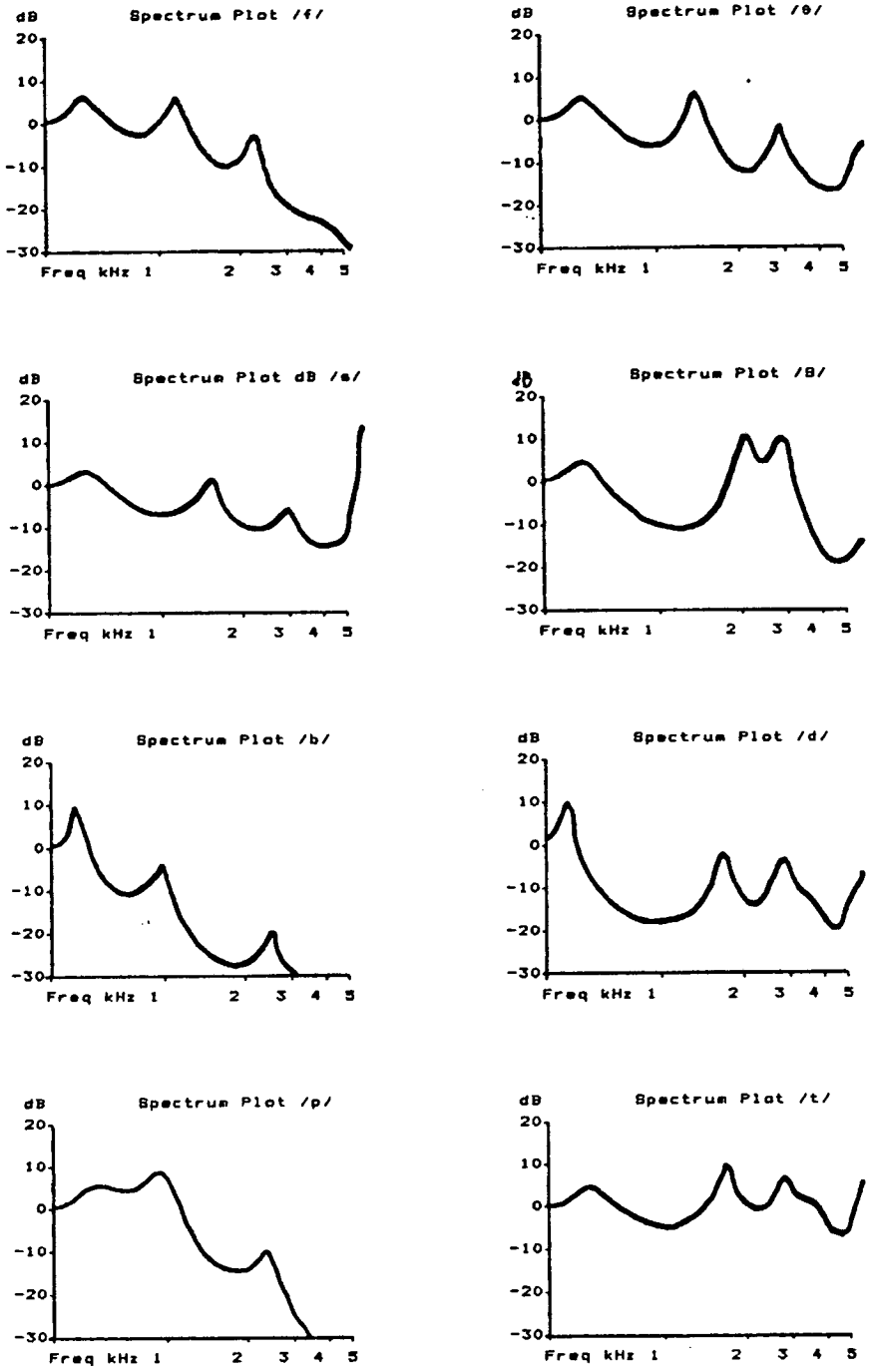


Figure 7.3: Example spectra for second attempt at approximation of Klatt (1980b) serial-plus-parallel synthesis of consonants, using a tenth order serial synthesizer. Only the fourth and fifth formants were modified to approximate the effects introduced by Klatt's parallel set of resonances.

7.4.1.2 Approximations for nasals

A final problem was the synthesis of nasals. The serial model and its all-pole equivalents have no separate nasal channel, whereas Klatt advocates a pole-zero pair, as shown in Table 7.3. A first attempt at a 'separate but still serial' nasal effect was to use F5 as a substitute for the nasal pole (ignoring the nasal zero). Unfortunately this caused transition difficulties into and out of vowels, as F5 moved across all the other formants from 270 Hz (the usual position for the nasal resonance) to 3850 Hz (the nominal F5).

Table 7.3: Klatt data for nasal consonants: three variable formant frequencies and bandwidths, and a pole+zero pair. fnp=frequency of nasal pole; fnz=frequency of nasal zero. (cf4=3300 bw4=500; cf5=3850 bw5=700; fixed resonances.) (Syllable initial only, hence no 'eng'=velar nasal)

sound	cf1	bw1	cf2	bw2	cf3	bw3	fnp	fnz
m	480	40	1270	200	2130	200	270	450
n	480	40	1340	300	2470	300	270	450

The second try was to use F1 as the nasal resonance, as it is in reality somewhat obscured. In the Klatt method a zero is introduced at nearly the F1 frequency (450 Hz zero, 480 Hz F1) for nasal consonants. The real motivation for the separate nasal resonance is to nasalise adjacent vowels when appropriate. This is an important feature for naturalness of continuous speech, probably an important feature for the identification of syllable-final nasal consonants, and a critical feature of nasality in syllable-final consonant clusters (cf Appendix 6: Basic Properties of Speech). However whether or not a synthesizer has an independent nasal channel, nasal vowels can only be synthesised in synthesis-by-rule if target values exist for nasal as well as non-nasal vowels, or if there are rules to appropriately modify the non-nasal vowel target data. As

the synthesis strategy used in this study did not extend to any form of coarticulatory phenomena (other than simple transitions between unmodified target values), there was no place for nasal vowels. Hence there was no real reason not to use F1 as a nasal resonance, as only nasal consonants were to be produced. The resultant data are shown in Table 7.4.

Table 7.4: Use of modified F1 to indicate nasality.

sound	cf1	bw1	cf2	bw2	cf3	bw3	cf4	bw4	cf5	bw5	ampl
m	300	100	900	200	2150	300	3300	500	3850	700	45
n	300	100	1600	200	2600	400	3300	500	3850	700	45
N	300	100	2400	200	2850	400	3300	500	3850	700	45

7.4.2 Durations

All the FAAF words are monosyllabic, and mainly with a CVC (consonant-vowel-consonant) structure. The words were synthesized on a basic 500 msec framework: 100 msec C, 300 msec V, and 100 msec C. This does not imply that all word durations were 500 msec, because for initial or final stops the 100 msec could include a (silent) closure portion, and for clusters some of the notional vowel portion could be actually an approximant or nasal.

One essential reason for trying as much as possible to standardise durations was to minimise the differential cues in the synthesis. The role of durations in speech synthesis and perception was not the question under study. The question being considered was the role of different synthesizer parameters, which are largely independent of durational considerations. If words can be recognised by duration rather than spectral information, then such words are of little value for examining synthesizer parameters.

Most of the frequency transitions are 50 msec. Amplitude transitions are 'instantaneous' for stops, and otherwise mainly 30 msec. Exceptions are mainly a few specific words where individual phoneme combinations were adjusted to fit the basic 100-300-100 framework. The full phonetic spellings with excitation, transition and duration data are given in Appendix 4.

7.4.3 Amplitudes

There are two sorts of amplitude control. First each phoneme has an 'intrinsic amplitude', given in the target values table. Second, amplitude changes are required to control the onset and offset of signal at the beginning and end of each word, and to produce closure gaps for stops. This second part of the amplitude control is associated with an entire word, rather than individual phonemes.

7.4.3.1 Intrinsic gains

The intrinsic amplitudes are based on values for excitation control given in Klatt (1980). His synthesizer has separate parameters to control two voiced and two voiceless sources, allowing various mixes of periodic and aperiodic excitation to simulate frication, aspiration, murmer or 'voice bar', mixed excitation for voiced fricatives and ordinary voicing for vowels and approximants.

In the simpler scheme used in this study, an attempt was made to separate gain from excitation. Vowels are typically louder than approximants, but this could be modelled by an appropriate intrinsic amplitude and a common excitation. This separation of gain from excitation allows a restricted set of excitation signals to suffice for all the FAAF words.

Table 7.5 shows the values used. The data for fricatives were adjusted after informal listening. The values for stops may seem mysterious: a stop means a silence, an amplitude of zero. But a stop is a sequence of events, closure followed

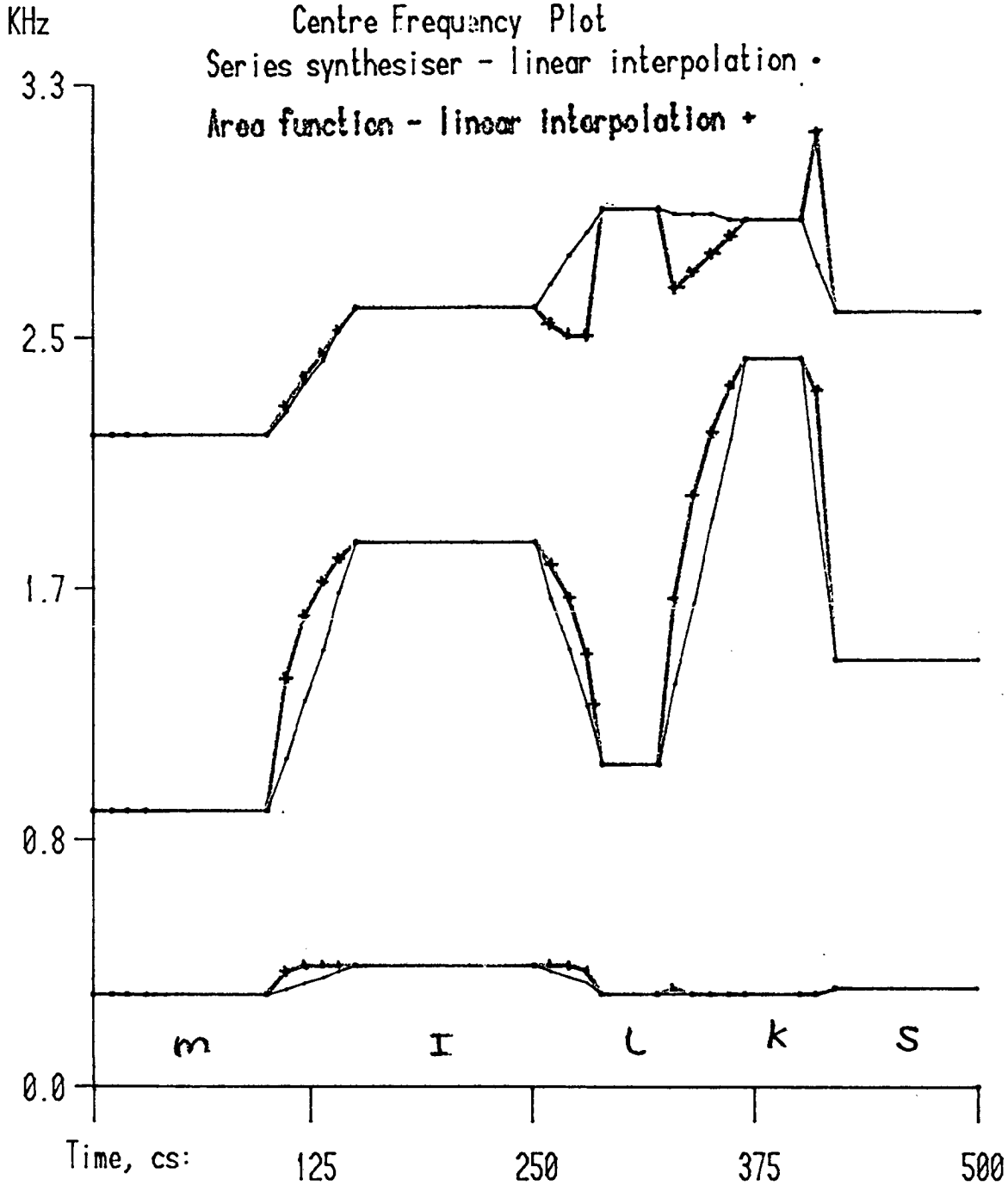


Figure 7.4: Example formant paths for the word 'milks'; series resonance parameters and area function parameters.

by release followed by optional aspiration. The intrinsic amplitudes in the table refer to target values for the release and aspiration stages of stops. The reduction to 0 dB for closure is accomplished as part of the second stage of amplitude control, the overall amplitude contours.

Table 7.5: Intrinsic amplitudes in dB re an 'off' or silence value of 0dB.

Vowels:		60		
Approximants:		50		
Fricatives:	f	45,	v	50
	θ	40,	D	40
	s	50,	z	55
	S	50,	Z	55
Stops:		45		
Nasals:		45		

7.4.3.2 Overall amplitude contours

The general amplitude shape of a word depends not on the specific phonemes involved, but on general phoneme categories. The essential distinction was between stops and continuants (consonants other than stops). Stops have closure and rapid amplitude transitions; the remaining sounds do not have closure and have slower amplitude transitions. An additional distinction was necessary between voiced and voiceless stops in syllable final position, because of rate and durational differences. However it was found that these could be adequately captured by lumping syllable-final voiced stops in with continuants, so only a two-way distinction was still adequate.

Thus for simple CVC syllables, a binary distinction for each consonant yields four contours:

1. CnVCn continuant+vowel+continuant
   examples: mail, nail; man, nan

   CnVS1 continuant+vowel+voiced stop
   treat final voiced stop as a continuant and use CVC
   examples: sub, sud

2. CnVS2 continuant+vowel+voiceless stop
   examples: fought, thought

3. SVCn stop+vowel+continuant
   examples: bail, dale; own, tone

   SVS1 stop+vowel+voiced stop
   treat final voiced stop as a continuant and use SVC
   examples: dab, gab; old cold gold

4. SVS2 stop+vowel+voiceless stop
   examples: taught, port

There is one special case: /h/. This phoneme only occurs in English in syllable-initial position, and is essentially an aspirated vowel: it has the resonances of the following vowel. Thus /h/ has no entry in the phoneme table, because it has no specific individual spectral description.

The first attempt to synthesise /h/ by simply feeding voiceless excitation into vowel formants (with the vowel's intrinsic amplitude of 60 dB) produced much too loud an output. This could have been solved by having a reduced intrinsic amplitude for /h/, but there is no /h/ in the phoneme table and so no intrinsic amplitude.

Another possibility was to permanently change the voiceless excitation signal, so that a voiceless excitation into a standard vowel specification produces a reasonable amplitude aspirated vowel. But this would have required a compensatory change upward of all the fricative and stop intrinsic amplitudes, giving them values 30 dB higher that used by Klatt. Such a change might indeed be a more principled way

to produce synthetic speech, but it seemed a drastic step just to synthesise four FAAF words (ham, high, hang, how) in which /h/ wasn't even distinctive. Therefore an ad hoc approach was used and a separate overall contour was created specifically for words beginning in /h/:

5. HVCn aspirant+vowel+continuant
   examples: ham, high, hang, how; hold

Words with consonant clusters produced the need for further contours. The simplest case was syllable-initial clusters, where a continuant followed by a stop required a gap some way into the word:

6. CnSVCn continuant+stop+vowel+continuant
   examples: stone, steam, stream

A more complicated situation arises with syllable-final clusters. English is particularly rich in these structures, which arise largely from the use of suffixes for syntactic purposes (notably possessives, participles and plurals).

Fortunately only the stop vs continuant distinction mattered. Continuant pairs and stop pairs did not occur in the list, and the two triplets (milks, lands) could be handled by allowing the first consonant of the cluster to occur during what ordinarily was the final part of the vowel portion of the word, and then treating the remaining two consonants just like any other consonant pair.

Total combinations of CVCC with continuant vs stop in the syllable-initial position, and two orderings of continuant and stop in final position, and a voicing distinction on each consonant would lead to 4x4x4 or 64 possible contours. But English is highly constrained, and the FAAF words are even more constrained. The voicing distinction doesn't affect amplitude in word-initial position, and is only considered once in final position (clusters 'agree' for voicing). This reduces the combinations to 2x2x2x2 = 16 (stop or continuant in each of the three positions times one voicing decision).

As further constraints, only voiceless syllable-final continuant+stop tokens occur in the FAAF wordlist, and the reverse order cluster only occurs on words beginning with a continuant. This removes two of the binary choices, leaving only four contours:

7. CnVCnS continuant+vowel+continuant+stop
   examples: lest, messed

8. SVCnS stop+vowel+continuant+stop
   examples: boast, ghost

9. CnVSCn1 continuant+vowel+stop+continuant (voiced)
   examples: ridge, rids

10. CnVSCn2 continuant+vowel+stop+continuant (voiceless)
    examples: match, mats

In summary, ten amplitude contours suffice for all the words in the FAAF list. It should be noted that most synthesis-by-rule schemes do not have a syllable-level gain specification; gain is more usually assigned at the segmental level. Gaps are introduced by splitting stops into multiple sub-phonemic segments (gap+burst+aspiration), and onset and offset ramps can be managed with transitions to a special 'silence' phoneme. Using these extra segmental devices allows all gains to be introduced in the phoneme table, and is probably an efficient way of handling unrestricted syllable types.

This study used gain contours for two reasons: it was at least as efficient as the method of 'sub-phonemic segments', as all the phonemic spellings were kept to minimal complexity; and gain contours do reflect linguistic regularities, such as common properties of stops vs continuants, constraints on consonant sequences, and agreement of voicing within clusters.

7.4.3.3 Interpolation of amplitude data

The amplitudes given by intrinsic values, and then modified by gain contours, are specifications at specific times along the course of the 500 msec word. For synthesis these amplitudes require interpolation to provide a control value for every frame. This interpolation was linear in dB, which is usual in synthesis (Holmes et al, 1964), and corresponds roughly with human perceptual scaling (Ladefoged, 1982, p169).

7.4.4 Excitation

The synthesis method used is based on an assumption of complete separation of excitation and filtering. Thus synthesis can then simply be a matter of passing an excitation signal through a digital filter. For general synthesis the excitation signal would be generated at the moment of synthesis, as there would be as many different kinds of excitation signals as there are utterances. For a restricted wordlist, however, it becomes possible to have a small set of precomputed excitation functions.

The excitation signal has three roles: an input signal with a flat spectrum, a distinction between periodic and aperiodic excitation (or some combination of the two), and finally provides a pattern of excitation frequency vs time over the course of the utterance which is perceived as the intonation contour.

7.4.4.1 Excitation types

Three kinds of excitation were used, as per Klatt (1980): fully voiced, fully unvoiced, and mixed.

The Voiced excitation Vx is purely pulses, a unit value followed by zeroes. No attempt was made to remove the DC component, which did in fact lead to discontinuities at the

ends of utterances. This caused no problems, because no substantial accumulation of DC occurred over 500 msec, and the output hardware was AC coupled. A unit pulse when run through the loudest sound (an /a/ with relative amplitude of 60 dB) produces numbers which just fit into a 12-bit D/A format.

The unvoiced excitation Rx is Random noise. Random numbers with a flat spectrum and flat amplitude distribution were added together (in blocks of 16) to produce a signal which still had a flat spectrum, but an approximately normal amplitude distribution. The range of amplitudes was adjusted to have a unit standard deviation.

The mixed excitation Mx is just used for voiced fricatives. Klatt (1980b) specifies using random noise which has been amplitude modulated by a square wave. The frequency of the modulating waveform is the desired fundamental frequency value, and the depth of modulation is 50%. Klatt does not discuss either varying the duty cycle or the depth of the modulation.

It is noteworthy that degree of mixing of periodic and aperiodic components is not continuously variable in the Klatt system. Nor is there any independent control of mix as a function of frequency, as in Holmes (1972).

Informal listening tests were performed on an implementation of the Klatt-style mixed excitation. The resultant signal was hard to distinguish from an unvoiced signal, except at very low fundamental frequencies (below 100 Hz). Accordingly one change was made to the Klatt specification: the depth of modulation was increased to 80% to increase the apparent periodicity of the signal.

Three excitation types in a sequence of three sounds (for CVC syllables) could generate a maximum of 27 combinations, without even considering consonant clusters. However there were many constraints on excitation combinations:

1- only Vx was allowed in the middle (vocalic) position, reducing the possibilities to 9;

2- Mx never occurred in a word which also required Rx, eliminating two combinations;

3- finally, no words both began and ended with Mixed excitation.

The result was a total of six excitation arrangements: VxVxVx, VxVxRx, RxVxVx, RxVxRx, VxVxMx and MxVxVx.

For example, VxVxVx is used for any word beginning with a voiced stop, nasal or approximant, and ending the same (eg mail). RxVxRx begins and ends with something voiceless (eg post).

7.4.4.2 Intonation contour

All the words have the same intonation contour: starting at 150 Hz, staying at 150 Hz for 200 msec, then declining to a target of 75 Hz over the remaining 300 msec. This is a one octave 'high fall', suitable for careful pronunciation of isolated single words, as in citation form.

Interpolation of voiced excitation (and the periodic component of mixed excitation) is logarithmic, corresponding to uniform. steps on a musical scale (which is appropriate to perception of intonation).

7.4.5 Parameter Updating

The interpolation of fundamental frequency and overall amplitude has already been mentioned. All the remaining parameters also required interpolation between target values to supply data for each synthesis frame. The frame duration was chosen to be 10 msec. In line with the general simplicity of the synthesis implementation, and to provide a baseline for more sophisticated methods, the updated

parameter data was not inserted pitch synchronously, but rather was put in strictly at the 10 msec intervals.

Methods of interpolation were not the subject of this experiment, but are considered in the next chapter. For this experiment on parameter types, all interpolations were linear, but linear within the individual parameter spaces. For example, synthesis from an area function used a linear interpolation of vocal tract section areas, and then the data were transformed to the standard system function representation for actual synthesis. Examples of series resonance and area function parameters for the word /milks/ are given in Figure 7.4.

7.5 PROCEDURE

The FAAF wordlist was synthesised as described in sections 7.3 and 7.4, above. A control method was available for the synthesis output programme (Sinclair and Munden, 1986) which allowed an entire 85 word test sequence (five practice items and 80 test items) to be produced and tape recorded automatically. The FAAF materials include five randomisations of the 80-word list, and all five were recorded. Appendix 4, section 5 shows the control data, including generation of pauses between pages (on the response sheet), precursive tones before each word, and interstimulus intervals. These intervals were limited by disc file access times, and resulted in a presentation interval of about seven seconds, which was just a bit slower than would be desired.

The recordings were made on a Uher CR160AV cassette tape recorder, using chromium dioxide tapes and not using Dolby or any other noise reduction technique. Tapes were monitored periodically for print-through, which did occur after six weeks (by which time the experiment had been completed).

Five synthesizer parameter types were tested rather than the full six which have been considered all along in this

research, because articulatory parameters were to be the subject of a separate experiment (Chapter Nine). One synthesizer type was used for each of the five orderings of the FAAF wordlist. This material was then presented to five subjects, with presentation order randomised according to a Latin Squares design (Steel and Torrie, 1981, p221). (Cf section 7.7.2, below.)

All subjects completed a questionairre concerning general health and hearing defects, but no direct measurement of hearing ability was performed. Subjects were mainly university students or researchers associated with the IBM (UK) Science Centre, and all had some familiarity with synthetic speech. The subjects were offered payment according to the number of correct identifications. Written instructions were used. The questionairre and instructions are in Appendix 4, Section 6.

The subjects were not trained to tune-in to the style of the synthesis. There is a large potential for learning associated with closed-response tests. The question of learning the response set is eliminated, but the stimulus material itself may allow for considerable learning, especially if it differs greatly from natural speech. The examination of this effect was not made a formal dimension of the experiment, but there was one subject available who had been exposed to the material for several hours a day for a period of three months. This subject had been engaged in informal listening experiments throughout the period of generation of the test stimuli, and so had been repeatedly exposed to all the tokens in the presense of feedback as to what the intended word actually was. This subject (the author) was included in the full knowledge that variation between subjects would inevitably be increased. The benefit would be a measure of naive performance (by the other four subjects) vs an indication of level of achievement produced by extensive learning.

The stimuli were presented binaurally through headphones in a sound treated room. The level was judged to be comfortable

by the experimenter, and was kept constant across subjects and trials. The stimuli were played on a Uher CR160AV cassette recorder (the same one they were originally recorded on) directly into Beyer type DT-100 headphones.

The stimuli were all produced with vowels having the same intrinsic gain of 60 dB. This procedure does not produce vowels or syllables of equal dB value, nor of equal loudness. However no attempt was made to adjust levels to minimise loudness differences, as any change in level would result in presentation of identical consonants at different levels, depending upon vowel context. A tone was placed on each stimulus tapes at a known level, and all the presentations were kept constant using VU-meter readings on these tones.

The subjects had a printed response sheet of the standard format used with FAAF tests: one page for the five practice items, and then a further four pages with 20 response lines per page. The task was to listen to a stimulus word, read the four response possibilities on the form, and circle one word. Each subject did all five FAAF tests in a one hour session, with short breaks between each test and about a five minute break after the third test. Each test lasted about ten minutes.

7.6 DATA

The collected data are in two parts. For each test the total number of words correct were scored, as shown in Table 7.6. These scores are the basis for any general conclusions concerning overall intelligibility as a function of synthesizer parameter type.

The data can also be scored in a more detailed (diagnostic) fashion to determine which confusions occur. The result of this analysis is many pages of printout, though averages across subjects are reasonably compact. The diagnostic analysis is more illuminating when combined with results on types of natural speech (such as unprocessed, digitised and

linear predictive coding analysis-synthesis). As these further results are not available until the end of all three intelligibility experiments, the discussion of diagnostic FAAF scores is presented in Chapter Nine.

Table 7.6: Raw scores (words correct out of 80), sums of squares, sums and average percent correct.

Subject:	1	2	3	4	5	Σjx^2	Xi.	Ave %
Synthesizer								
Series	59a	52b	61c	59d	76e	19163	307	76.8
Direct	50b	49d	56a	57e	68c	15910	280	70.0
Reflect.	51c	53e	54b	66a	71d	17723	295	73.8
Areas	57d	53a	57e	57c	71b	17597	295	73.8
Parallel	51e	38c	54d	54b	69a	14638	266	66.5
Σix^2	14432	12167	15938	17251	25243	Σ=85031		
X.j	268	245	282	293	355		Σ=1443	
Ave %	67.0	61.2	70.5	73.2	88.8			

Order of presentation is indicated by the letters abcde in the raw data. Subject 5 had prior training on the stimuli.

7.7 ANALYSIS

Two types of statistical test may be applied to the data of Table 7.6. The first type considers the question "which synthesis parameters give highest intelligibility?", and determines level of significance through t-tests on mean intelligibility scores for pairs of synthesizers, or on any synthesizer vs the group mean. The second type considers the data as a whole, and asks if overall either the differences in synthesizer parameters or test subjects were significant, and if so by how much.

7.7.1 Tests of Mean Intelligibility Differences

By inspection of the scores in Table 7.6, the serial synthesizer parameters scored highest and the parallel scored lowest. The point of statistical analysis is to determine a significance level, a probability that the observation arose by chance. The problem with the raw data is that there is more variation by subject than by synthesizer. Consider the average percentage correct scores. When averaged across subjects, to show effects according to synthesizer, the range of scores is from 66.5% to 76.8%; when averaged across synthesizers to show overall subject differences, the scores ranged from 61.2% to 88.8%. Even without subject 5, there is still more subject variation than synthesizer variation, which means it is hard to prove significance of synthesizer difference unless the subject effects can be separated. Standard t-tests were run on the data in Table 7.6, but not even the largest difference (between series and parallel synthesis) achieved significance at the 0.05 level.

Separating subject from synthesizer effects requires analysis of variance, and indeed such analysis leads to statistical significance on the Table 7.6 data, as discussed below. It might have been possible to achieve significant results without analysis of variance: the method of paired comparisons could have been used, or the subject scores could have been individually normalised before analysis. However only analysis of variance could have not only separated subject effects from synthesizer effects, but also tested for effect of order of presentation of the tests. As an analysis of variance was to be carried out to test for an order effect, no attempt was made to perform these simpler tests for significance of mean differences.

7.7.2 Analysis of Variance

The general model underlying analysis of variance is to postulate factors which could produce a spread or dispersion of data, and thus enlarge the total variance. The

explanation of total variance according to contributing sources is the essence of the analysis.

While there are many varieties of analysis of variation, according to different types of experimental design, there are certain common variance assumptions:

1) observations are samples from a normally distributed random variable;

2) standard deviations of the different classes (rows, columns, treatments) are all equal.

It is not usually possible to prove the validity of these assumptions, as at best one only has estimators of the parameters of an underlying distribution, not direct access. Further, the concept of an underlying distribution is rather more a model than a reality - for instance, subjects were picked because they were available, which may or may not be equivalent to sampling a normal distribution.

However, there are no obvious reasons why intelligibility scores on randomly chosen subjects should not be reasonable, especially when results are around the 50% to 70% region where saturation (ceiling) effects are minimised. The standard deviations of the scores can be estimated, and as the sums of squares in Table 7.6 differ by less than a factor of two, the estimated standard deviations are within a factor of the square-root of two of each other. Thus the requisite assumptions were deemed not to be obviously violated.

The simplest form of analysis for a matrix of data arranged by subject and 'treatment' is two-way analysis of variance. The data for this analysis are in Table 7.8.

With proper arrangement of trials, a notionally two-dimensional data array can be tested for a third effect of presentation order (or learning) providing the layout is such that each row and column includes exactly one example of each step (level) in the ordering. Thus for a 5x5 layout of

synthesizers times subjects, where each subject is exposed to all the synthesizers in 'random' order, the actual order of presentation must be such that each synthesizer is tested once in each of the possible ordinal positions: first, second, third, fourth and fifth.

A Latin Squares NxN matrix is a matrix whose elements are the integers one to N, with each row and column containing each number exactly once. This desired constraint for analysis of a third effect within essentially a two-dimensional experimental design is ensured by using a Latin Squares matrix of subjects by synthesizers, with matrix elements representing (in this case) test presentation order.

A Latin Squares design was used in this experiment, and the analysis data are in Table 7.9, including analysis of the effect of presentation order. Order of presentation is indicated by the letters abcde in the raw data, Table 7.6. The additional sums and sums of squares pertinent to analysis of order effect are in Table 7.7. These analysis of variance results are given in Section 7.8. Conclusions are presented in Section 7.9 and discussed in Section 7.10.

Table 7.7: Order effect calculations: sums of squares, sums and average % correct according to presentation order.

order	Σx^2	Xt	Ave % correct
a	18543	303	75.8
b	16077	281	70.2
c	15639	275	68.8
d	17088	290	72.5
e	17684	294	73.5
Σ	85031		
Σ		1443	

7.8 RESULTS

The standard analysis of variance is presented in table 7.8 for a two-way analysis (subjects and synthesizers), and in table 7.9 for a three-way analysis which also looks at an order effect. Full details of the analyses are contained in Appendix 4, Section 4.7.

The columns in tables 7.9 and 7.10 are:

1) postulated source of variance;

2) variance about the postulated source, labelled sum of squared deviations;

3) the statistical degrees of freedom (dof);

4) variance divided by degrees of freedom (a sort of normed variance);

5) F-ratio: a ratio of normed variances which has a known distribution, and can be tested for significance. Significance at the 0.05, 0.01 or 0.001 levels is indicated by one, two or three asterisks, respectively.

Table 7.8: Two-way Analysis of variance

source of variance	sum of squared deviations	degrees of freedom	mean square deviation	F
synthesizers	B = 201	I-1 = 4	B/(I-1) = 50.25	4.45*
subjects	C = 1359.4	J-1 = 4	C/(J-1) = 339.8	30.1**
residual	D=A-B-C=180.6	(I-1)(J-1)=16	D/dof = 11.29	
total	A = 1741	IJ-1 = 24		

---

table 7.9: Latin squares (three-way) analysis of variance:

source of variance	sum of squared deviations	degrees of freedom	mean square deviation	F
synthesizers	B = 201	r-1 = 4	B/(r-1) = 50.25	7.15**
subjects	C = 1359.4	r-1 = 4	C/(r-1) =339.8	48.3***
order	D = 96.2	r-1 = 4	D/(r-1) = 24.0	3.41*
residual	E=A-B-C-D=84.4	(r-1)(r-2)=12	E/dof = 7.03	
total	A = 1741	r^2-1 = 24	A/(r^2-1)= 72.54	

7.9 CONCLUSIONS

Table 7.6 shows that intelligibility differences arising from the five different synthesizer parameter types are small. The range in scores for average percentage correct word recognition is about 10%, and three of the synthesizers lie within a 4% range. This is less than the differences between subjects, where the range even for the four similar subjects (numbers 1 to 4) was 12%. Further, Table 7.7 shows a 7% range just for presentation order differences (data averaged across subjects and synthesizers). The experiment does not have a dramatic result.

The result of the two-way analysis of variance, table 7.8, is a statistically significant difference amongst the five synthesizers ($p<0.05$), and a more significant difference amongst subjects ($p<0.01$).

The symmetry of the experimental design is fully exploited in the three-way analysis of variance, table 7.9. The result is a higher level of significance on the same conclusions:

1- a very significant difference among synthesizers, $p<0.01$.

2- an extremely significant difference among subjects, $p<0.01$.

The three-way analysis also examined the effect of order of presentation of the FAAF tests, and concluded that this effect is also significant, at the 0.05 level.

As the subjects neither uniformly improved (learning) nor got worse (fatigue) as a function of serial order, but rather performed slightly worse in the middle than at the beginning or the end (Table 7.7), the order effect could be called the 'sag effect'. The performance declined on the second and third tests, and rose on the fourth. There were small breaks before the second and third tests, and a five minute break before the fourth test. Possible the 'sag effect' was fatigue over three sets followed by an improvement after a break.

Finally, the one extensively trained subject scored nearly 90% correct on this wordlist. He had the same slight preference for the series parameters as had the remaining subjects.

This high intelligibility score has two implications:

(1) there are slight acoustic cues available in the stimuli to produce a recognition rate significantly higher than that achieved at first exposure; but these are not quite like the cues in natural speech because they must be learned.

(2) the author of any synthesis system has good reason to have an inflated view of the intelligibility of his or her own synthesis method : for that one person the high intelligibility actual exists, because of longterm learning.

7.10 DISCUSSION

Experiment I was an objective analysis of differences amongst synthesizer parameters, and showed that interpolation paths differed in ways which were larger that the difference limen for steady formants (though not greatly larger). Thus there was the potential for differences which would affect intelligibility and naturalness.

Experiment II performed an intelligibility test, and found a small but statistically significant difference. Series resonance parameters scored highest, parallel resonance lowest, and the rest with nearly identical scores in the middle.

However, another 'result' of Experiment I was the development of three tools for experimentation:

1- methodology and subroutines for interchanging parameters;

2- a synthesis-by-rule scheme;

3- formal specification of the FAAF stimuli.

These synthesis capabilities are used in the remaining experiments, and could also be of more general interest. The parameter interchanging routines may be of use to any enquiry into questions of synthesizer parameters and their acoustic consequences. Pascal listings of these routines are given in Appendix 1. The synthetic FAAF specification allows synthesizers to be tested with regard to specific detail. One problem with simply using orthographic or near-orthographic spelling of wordlists as the input to synthesis schemes is that there are many levels of processing. Why does one text-to-speech system get a better intelligibility score than another? Is it the orthographic to phonetic conversions, the coarticulation, the phoneme table, or the final synthesizer structure? With a lower-level input, one can begin to point to more specific answers. A starting point of specified phonemic labels, durations and excitation

removes much of the uncertainty about whether differences arise from the orthographic processing or the final synthesizer. Thus there is a role for parametric input to synthesizers for the purposes of evaluation. This experiment has produced a parametric FAAF specification.

The results concerning series vs parallel resonance synthesis require qualification. The parallel resonance parameters examined in this study are not the equivalent of a practical parallel synthesizer. The parallel parameters were derived from the series parameters in such a way as to have nearly equal formant frequencies and amplitudes, and totally without regard for what happened as a result of the system function zeroes created by the parallel configuration. In fact, it was just the effects of these zeroes which were examined in both Experiments I and II. There were four synthesis tokens (ban, land, bang, lad) amongst the 80 FAAF words in which the effects of the zeroes were so pronounced as to cause a disturbing break in the word. Removing these tokens from the scoring improves intelligibility by about 4%, and puts the parallel results in line with the direct form, area function and reflection coefficient results.

Finally, interpolation of direct form coefficients yielded unstable filter configurations in 30 of the 80 FAAF words, specifically those words involving /bvfr/. The factor of two limit (within the synthesizer) on exponential growth resulted in clipped signals which sounded like clipped sine waves, producing 'bleeps' during the unstable portion of the word. In 22 cases the unnatural sound occurred during a portion of the word which made no difference to the choice of response, and in four cases the bleep only occurred on one word in a set. In six cases (bin/pin; feel/veal; robe/rove) the instabilities occurred during the sound which was critical for correct recognition. Although this affect was clearly audible, intelligibility of the direct form synthesis was commensurate with results using reflection coefficients or area functions. Subjects remarked on the 'bleeps', but were not prevented from making correct recognition.

Chapter Eight: Intelligibility comparison of interpolation types. Experiment III

8.1 OBJECT

Experiment II (Chapter Seven) was a comparison of synthesizer parameters, using linear interpolation in five different parameter spaces - the five different synthesizer types.

Another issue concerning parameters in speech synthesis is the question of interpolation of parameters between target values. There have historically been various approaches to this problem, as discussed in Chapter Two, but a search of the literature did not uncover a direct comparison of interpolation types, all else being equal.

Experiment III examines interpolation types, using a series resonance synthesizer and four interpolation strategies:

a linear (from Experiment II data)
b discontinuous formant paths (jumping without any interpolation to the new value)
c piece-wise linear (PWL) transitions used by the Joint Speech Research Unit (Holmes et al, 1964)
d cosine shaped transitions as used by IBM, Winchester (Sharman, 1986)

The same intelligibility testing materials, namely the FAAF wordlist, were also produced using linear prediction analysis-resynthesis of natural speech (LPC speech). The motivation for using the LPC material is that the transitions in this natural speech are as good as can possibly be expected (within the constraints of a ten parameter representation of speech), and so the material provides stimuli for an estimate of an upper bound on intelligibility (of parametrically coded monosyllables).

The synthesis with no interpolation - the discontinuous formant paths - is intended to provide a lower bound on the effects of interpolation upon intelligibility.

8.2 STIMULI

The stimuli were produced by synthesis-by-rule of the FAAF wordlist, exactly as in Experiment II. The tokens using linear interpolation were the same stimui as used in Experiment II.

Cosine transitions were used in the text-to-speech system at IBM (UK) Science Centre, Winchester (Sharman, 1986). They are one way to eliminate slope discontinuities going into and out of a transition, because 'Smooth, continuous formant transitions are generally observed on spectrograms of real speech' (Rabiner, 1968b, p25). The stimuli with cosine transitions were generated exactly as for the linear method, but with a different interpolation formula at the stage in the synthesis where parameters are interpolated between target values.

The abrupt formant transitions required another interpolation formula, though interpolation is rather a misnomer for the generation of discontinuities. A problem arises as to where this abrupt change should take place, relative to the amplitude control for word onset and offset. For words beginning with a stop consonant, the amplitude ramps up suddenly (over one 10 msec frame) by 50 dB at the point of release. The instantaneous formant transition could come at this same time, so that by the time the amplitude is fully on the formant has reached the final value. Alternatively the transition could come later, with the formant transition later than the the amplitude transition, allowing some acoustic effect of the formant prior to the discontinuity.

Informal listening tests showed that an early jump sounded just like a glottal stop, as expected. There are no audible formant transitions, because the formant starts off at the target value. The only stop consonant which does not produce formant motion is the glottal stop, which does not involve a change in the shape of the vocal tract.

A late jump (after the amplitude has come to full value) still sounds like a glottal stop, but preceeded by a sort of hum. There was little to prefer in either case and so it was decided to use the early jump, and avoid the hum.

The examination of piecewise linear transitions was an attempt to include a sophisticated interpolation scheme amongst the methods. Two state-of-the-art synthesis systems use a piecewise linear approach: that of Klatt (1980a) and Holmes et al (1964, as implemented by Wright, 1976).

The Klatt transitions are governed by many rules, and only a few exemplars have been published. The JSRU transitions, however, are based on a single rule with many phoneme-dependent parameters, and all the parameter values are published in Holmes et al (1964). Therefore it was decided to implement the Holmes JSRU scheme, to the extent that this was possible.

Surprisingly, about 65% of the JSRU-style transitions are one-piece linear. To produce the remainder of the scheme, the existing synthesis software was used to implement one-piece linear transitions, and these were then manually edited to produce the effect of the JSRU rule and parameters.

The linear prediction analysis-synthesis data were taken from a high quality recording produced by the Institute of Hearing Research in Nottingham. This tape has all the words recorded very carefully and adjusted for amplitude, but the words occur in a carrier phrase.

The IBM (Winchester) Speech Group analysis system (Alderson, et al, 1984) was used to digitise all the phrases, and excise the requisite words. These words are very carefully articulated, in a citation form style (despite the carrier phrase) including careful release of syllable-final stops. As the synthetic FAAF data was equally carefully created to specifically not include release of stops in this position, part of the editting process was to eliminate these portions of the natural utterances.

The actual LPC analysis-synthesis was performed using the standard Markel & Gray (1976) autocorrelation method, including their SIFT method of fundamental frequency determination. A tenth order model was used, with a 25 msec analysis window, 10 msec frame rate and 10 kHz sample rate. No quantisation of LPC coefficients was introduced; the coefficients and all calculations used a 32-bit floating point representation. The frame rate is the same as was used for the purely synthetic data.

8.3 PROCEDURE

The FAAF materials as described in Experiment II were synthesised using each of the interpolation types. The stimuli generation and recording were exactly as in Experiment II. Linear interpolation data had been produced for that study, so one of the four remaining standard randomisations of the 80-word list was used for each of the three other interpolation types, and the final randomisation was used for the LPC speech.

The four new stimulus types were then presented to four subjects (all of whom had previously been tested on the linear interpolation data), with presentation order again fully randomised according to a Latin Squares design.

The subjects had also participated in Experiment I, and the discussion of that experiment in the previous chapter describes subject selection, the health questionaire, and instructions for the experiment. Again, the subjects were offered payment according to the number of correct identifications. Again one subject had been extensively exposed to the stimuli, to get an indication of the effect of training.

The stimuli presentation and control of level were as described for the previous experiment. Again a written response was required, using printed response sheets. The task was to listen to a stimulus word, read the four response

possibilities on the form, and circle one word. Each subject did four FAAF tests in a one 50-minute session, with short breaks between each test and a five minute break after the second test.

8.4 DATA

The results of the intelligibility tests are shown in Table 8.1, which gives scores in words correct out of the total of 80 stimuli presented in each FAAF test. As with Experiment II, the further results obtainable by analysis of the different types of identification errors will be deferred until the next experiment, when the diagnostic FAAF results can be presented as a group and compared with scores on natural speech.

Table 8.1: Raw scores (words correct out of 80), sums of squares, sums and average % correct for four interpolation methods and LPC speech.

Subject:	1	2	3	4	Σx_j^2	$\bar{X}_i$.	Ave % correct
Interp. discont.	40	44	37	55	7930	176	55.0
PWL	56	53	54	61	12582	224	70.0
linear	67	53	61	76	16795	257	80.3
cosine	68	59	55	76	16906	258	80.6
LPC	68	72	60	75	19033	275	85.9
Ave %:	74.8	70.2	66.8	85.8			

Subject 4 had been trained before testing.

8.5 ANALYSIS

Experimental design issues of subject and order effect had been taken up in Experiment II, and so were not reconsidered in the analysis of the present data.

As in Experiment II, there is considerable subject variability which leads to problems of nonsignificance when just comparing mean results. In that case analysis of variance was used to separate the sources of variation in the results, as analysis of variance was required for subject and order effect evaluation anyway.

The data of Table 8.1 represent a wider range of effects than was the case for Experiment II. There it was a matter of considering five parameter types, with no a priori reason for any of them to be markedly different. In the interpolation data of Table 8.1 there are three sets of notionally sensible interpolation-type results, plus scores on null interpolation designed to give a minimum score, and finally the data from natural speech designed to give a maximum score.

The analysis-of-variance performed on the data of Experiment II tests the hypothesis that there are no significant differences as a function of stimulus type. In the present case there are obvious differences between the abrupt transitions and the LPC speech, and so another analysis is needed to make more specific determinations of significance.

A standard t-test of mean-value differences is possible, if the problem of subject variation can be handled. The approach used is paired comparisons: the data are differences in intelligibility between two interpolation types, rather than raw intelligibility scores. The analysis for cosine vs LPC is given in Table 8.2.

Table 8.2: paired-comparison analysis of two interpolation types.

subject	cosine	LPC	difference
	Y1	Y2	d = Y2 - Y1
1	68	68	0
2	59	72	13
3	55	60	5
4	76	75	-1
ΣY	258	275	Σd = 17 Σd² = 195
ave Y	64.5	68.75	D = ave d = 4.25

$s^2 = (\Sigma d^2 - [\Sigma d]^2/n)/(n-1)$

$= (195 - (289/4)) / 3 = (195-72.25) / 3 = 40.9$

$s^2/n = 40.9/4 = 10.2$

$s = \sqrt{(s^2/n)} = 3.2$

$t = D/s = 4.25/3.2 = 1.33$, dof=3.

one-tailed test for LPC > cosine, mean difference > 0.

8.6 RESULTS

The results of the tests for significant differences are given in Table 8.3. Only results with raw intelligibility differences of more than 11% gave rise to significance at the 0.05 level. The comparison of piece-wise linear vs cosine interpolation, with a 10% intelligibility difference, was approaching significance. The 5% difference between LPC and cosine was not significant.

No further analysis was made of the linear interpolation vs cosine function data, as the scores were so similar. Also, no test was performed on the linear interpolation results versus the PWL, LPC or abrupt data, because the comparisons in Table 8.3 involving cosine data are nearly identical. Further, conclusions regarding intelligibility of cosine interpolated synthesis apply equally to linear interpolation.

Table 8.3: Results of paired-comparisons tests. D, s and t as for Table 8.2.

pair	D	s	t	significance	
cosine vs LPC	4.25	3.2	1.33	NS	
abrupt vs LPC	24.75	1.97	12.54	p<<0.01	***
cosine vs abrupt	20.5	2.78	7.36	p<0.01	**
cosine vs PWL	8.5	3.12	2.72	p<0.1	
PWL vs LPC	12.75	2.69	4.74	p<0.02	*

8.7 Conclusions and Discussion

The data support several conclusions. The first is that there is no significant difference between linear and cosine interpolation methods. There is hardly even an insignificant difference between these two methods; there is virtually no difference at all, so far as intelligibility is concerned.

The basic reason for the similarity of linear and cosine interpolation is illustrated in Figure 8.1, which shows formant frequency paths for the synthesis of 'milks'. The transitions in this word are not faster than for the synthetic FAAF tokens in general; this token was chosen simply because it has more than the usual number of transitions and thus more illustrations of the basic point. The reason linear and cosine interpolation give near-identical word recognition scores is that they are nearly identical methods. There is very little apparent difference in the parameter paths, and hence hardly any chance for an auditory difference. Only on very long transitions indeed (over 100 msec) could linear vs cosine transitions be expected to be perceptually distinguishable.

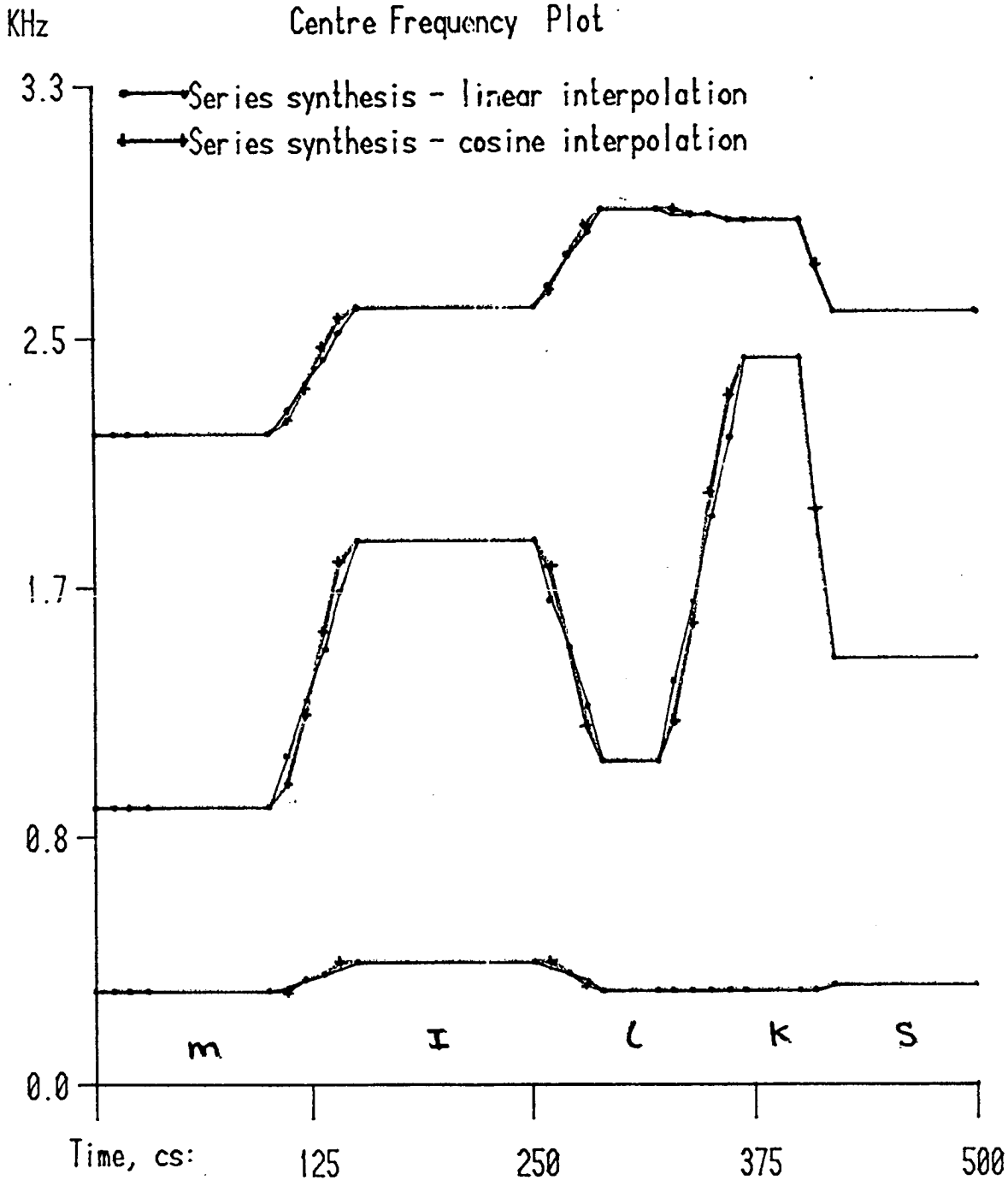


Figure 8.1. Formant frequency paths for linear and cosine interpolation, for the token /milks/.

The average difference between a linear and cosine path for a normalised transition from 0.0 to 1.0 can be easily computed as shown in Figure 8.2 and Equation 8.1.

$$\bar{d} = \frac{1}{\pi/2} \int_{t=0}^{\pi/2} \sin(t)dt - \pi/4 \approx 0.137 \qquad (8.1)$$

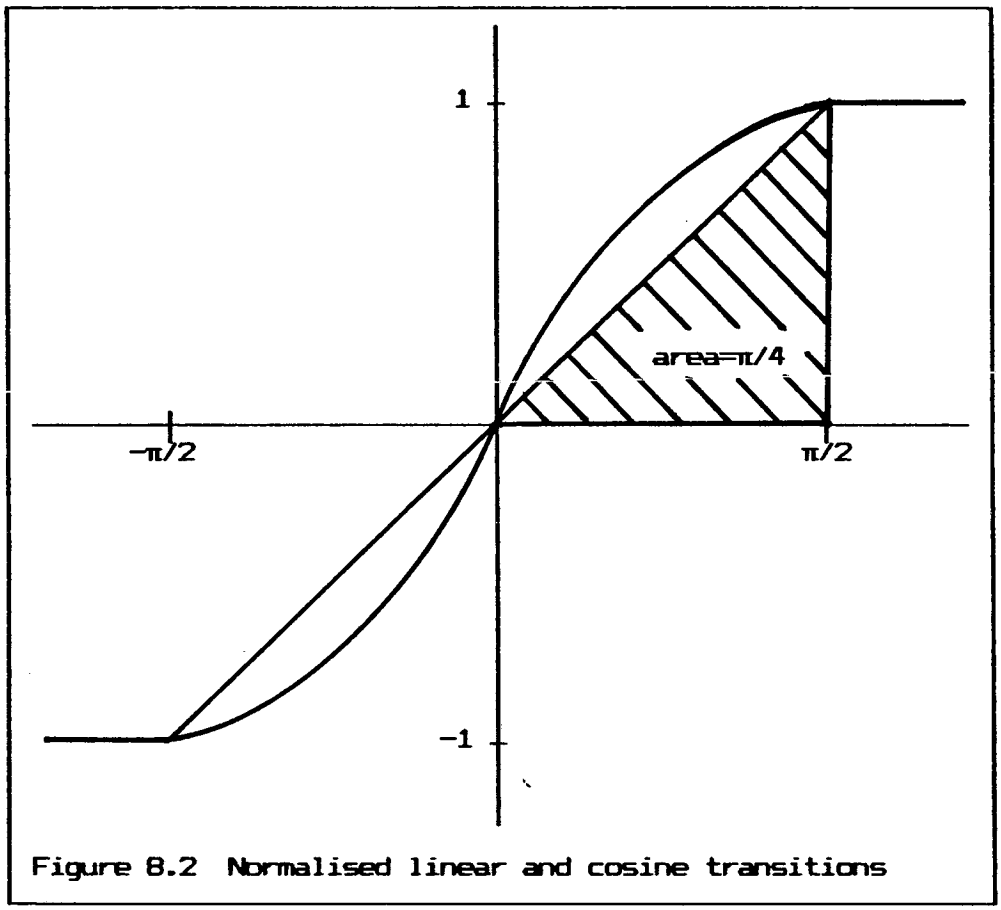


Figure 8.2 Normalised linear and cosine transitions

The average difference of 0.137 will have greater or lesser significance, depending upon the size of the transition. For the purposes of a rough estimate, a transition of half the size of the target value leads to a formant frequency transition difference (between cosine and linear) of about 7% of the steady formant values. As mentioned in Experiment I, the Just Noticeable Difference for steady formants is in the range 3-5%, so we could well expect the linear vs cosine differences to be imperceptible.

The next conclusion is that the simple single-line linear interpolation used in this study works quite well. The results were within five percentage points of the ceiling score on data obtained through linear predictive analysis-synthesis of natural speech. Further, it is an open question whether the remaining 5% could be made up by any interpolation method, as there are many differences between the purely synthetic stimuli and the LPC stimuli.

So far as the trained subject was concerned, Table 8.1 shows that there is nothing to be gained by changing from linear interpolation. This subject scored 95% correct using linear or cosine methods, just as high as for the LPC speech. It might be possible for some other interpolation method to improve acoustic cues sufficiently for a naive listener to move toward these figures. It seems rather more likely that this high performance by the trained listener indicates weak or unnatural distinctions in the target value data (the phoneme table) which can be learned through long exposure.

In the next experiment (number IV) further data on natural speech will also be obtained, and the diagnostic FAAF results will be presented. At that point the question of which issues (synthesizer parameters, interpolation type, or other factors) govern intelligibility can be considered in the light of the nature of the recognition errors.

A third conclusion is that a sophisticated method can give poor intelligibility if not properly implemented. The piecewise linear interpolation was based on parameters for a parallel-resonance synthesizer with a separate nasal channel (the JSRU synthesizer). The series synthesizer used in this experiment had a broadened first formant for nasal consonants. Applying the JSRU rules for a nasal resonance directly to the pseudo nasal resonance (created by having the first formant do double duty) leads to a gross alteration of first formant frequency and bandwidth during the vowels adjacent to nasals. The auditory effect is a slow quality change, rather than the desired coarticulation or 'spread' of nasality.

Another difficulty with implementation of the JSRU rules involves specification of the second formant for the velars /kg/. These rules were intended to model the allophonic change for velars as a function of whether the adjoining vowel was front or back: a palatal sound before front vowels, and moving towards a true velar as the vowel moves back. In the implementation tested in this study, the rules produce a second formant frequency transition which changes direction in the middle of the transition. The rapid motion and 'doubling-back' makes a very unnatural effect on all the tokens involving velars; this problem affects a sizeable portion of the FAAF words.

The result of these problems with nasal and velar consonants is that any benefit of the use of two linear segments rather than one is overshadowed by the anomolies on these particular sounds. In principle two lines should be better than one, simply because the paths possible for single line interpolation are a subset of the two line case. In practice, at least in this study, the attempt to model coarticulatory effects in an inappropriate way led to paths that were clearly worse that in the single line case.

Single line linear interpolation is already within five percentage points of the score for LPC speech. Even ideal interpolation (whatever that may be) may well not score the full five points higher, as there are all the other differences between the synthetic and the LPC speech tokens to consider. The result is that any anomolies introduced by a method more ambitious than the simplest single-line linear interpolation may well (as in this study) cause a net reduction in intelligibility. There simply is little room to do better, and lots of potential for doing worse.

The final conclusion is that, although very simple interpolation does quite well, this is not because interpolation is unimportant. Complete lack of interpolation has a disastrous effect on intelligibility, a drop of 25 to 30 percentage points.

The uninterpolated (discontinuous) formant paths differ from transitions produced by all the other methods primarily in duration. The other methods have transition durations of 30 msec or more (out to 100 msec), but discontinuities in parameter paths are transitions of zero duration. It is well established by experiments on speech perception that formant transition duration differences lead to differences in phonemic categorisation. Short transitions cue the perception of stop consonants, longer duration transitions (above about 50 msec) are heard as approximants. Examples are the continuum from /b/ to /w/, and also from /g/ to /j/, first studied in the classic Haskins Laboratories synthetic speech experiments using the Pattern Playback (Lehiste, 1967, pp159-169). In a study of word concatenation using coded natural speech, Young and Fallside (1979, p690) reported that "modification of the formant parameters (or equivalent) at word boundaries was of only secondary importance", but that "syllable duration should be modified to give good intonation and natural rhythm".

The point is that even the simplest interpolation, the linear method, allows transitions to be of appropriate duration. Eliminating interpolation by using abrupt transitions can be interpreted as reducing transition duration to zero, causing problems at both the phonetic and prosodic levels. One might conclude that what really matters, when going from A to B, is not how you get there but how long the trip takes.

Chapter Nine: Intelligibility assessment of an articulatory model. Experiment IV

9.1 OBJECT

The object of Experiment IV is to determine the intelligibility of synthesis using articulatory constraints. A three parameter model (based on Harshman et al, 1977) is used to determine tongue and lip positions from which a vocal tract area function may be derived. The results are compared with intelligibility of natural speech items, and with results on more conventional approaches to synthesis as studied in the previous experiments.

From the beginning of this investigation of synthesis parameters, one method of representation has been distinct: the articulatory parameters. An articulatory representation, specifically tongue and lip control factors, differs in important respects from the remaining speech representation possibilities:

1- dimensionality: three tongue and lip factors make a plausible model, whereas the other representations considered in this study all have ten parameters.

2- approximation: the four all-pole models considered in Experiment I are formally equivalent. There is an exact conversion relationship between the parameters. The parallel resonance model also can be exactly converted to a series representation if bandwidths are not constrained, and even with fixed bandwidths the formant frequencies and amplitudes can be closely matched. But the articulatory representation (because of reduced dimensionality) severely constrains possible vocal tract shapes.

3- target data: there have been many studies of formant positions in natural speech, providing data for series and parallel synthesis. Target data for the direct

form, reflection coefficients, and area function may be obtained through the exact conversion from series parameters to the other all-pole 10th-order representations. But obtaining plausible and useful articulatory targets from resonance data is difficult, as discussed in Chapter Five. Conversion from an acoustic representation to articulatory targets is the notorious problem of the acoustic-to-articulatory inversion.

4- control strategy: acoustic parameters do not control physical objects with standard properties such as mass and velocity. Formants do not have physical properties, and therefore it is awkward to try to determine how formant motion should be controlled. Even the reflection coefficients and area function do not refer to individual physical entities, but to the shape resulting from the motion of the articulators. But tongue and lip controls do refer to specific articulators which have physical properties and are subject to constraints. Not all the constraints are known, however, and this is an area of active research. The use of an articulatory representation does allow hypotheses about such constraints to be tested.

Because of these differences, articulatory synthesis was not tested in Experiment II, but was made the subject of this separate experiment. Once acceptable articulatory target values had been established for the phoneme inventory (by the adaptive search method of Chapter Five), two types of articulatory synthesis were produced: conventional linear interpolation between target values, and parameter motion subject to a phase-plane constraint on position vs velocity (Kelso et al, 1985).

9.2 ARTICULATORY TARGET VALUES

As discussed in Chapter Five, an attempt was made to derive useful articulatory target values from the standard Klatt

(1980b) resonance data, as this dataset had been used for all the other synthesis.

Four methods were tried:

1- the equations of Ladefoged et al (1978) relating tongue and lip parameters to formant frequencies;

2- a related approach using the projections of area functions onto the Harshman et al (1977) tongue parameter basis vectors. The area functions are really lossless-tube pseudo area functions determined from the series synthesis parameters by an exact relationship.

3- hand-picked target values, using contour plots of formant frequency vs tongue parameters.

4- an automated gradient search.

Test items were synthesised using the first two methods, and the results were not judged usable. Even after correcting for bandwidth problems by simply using the articulatory parameters to only determine formant frequencies (and taking the bandwidths from the original Klatt table) there were still problems. Fully half the FAAF words had bleeps - portions of words with pronounced and non-speechlike oscillations in the waveform. These effects arose mainly from the first formant frequency going to zero (complex roots becoming real). The majority of the words were markedly unnatural, for both the first and second methods of attempting to obtain articulatory parameter values.

The contour plots were adequate for determining acceptable formant frequencies in the two-dimensionsal tongue factor space, but the third dimension of lip opening had been neglected. Adding lip opening as a parameter meant further plots for each value of lip opening to be considered. Ten such values and five formant frequencies yield 50 plots, at which point the hand-picked approach became intractable.

The result of the automated search (with manual intervention) is a table of tongue and lip parameters which give a reasonable approximation to the Klatt data for most sounds. This data is presented in Table 5.3 (Chapter Five).

9.3 ARTICULATORY CONTROL

The question of temporal or dynamic constraints upon speech production is a large issue. There are considerations of neurological processes, muscle properties, force and mass of the articulators, feedback processes (both acoustic and proprioceptive), and linguistic questions of timing and duration, to name a few obvious aspects.

Because there is no accepted and available model of articulatory dynamics, part of this experiment was to test a partial implementation of a simple model which is currently receiving general attention: the model of intrinsic timing and phase-plane constraints of Kelso et al, 1985.

The papers on the question of speech timing deal principally in measurements of durations on the speech waveform (Lehiste, 1967). Another large set of physiological measurements (for example Harris, 1971 and many more Haskins Labs papers) concentrate on relative timings for various channels of activity, such as muscle groups or lip vs jaw motion, but usually for single articulatory gestures or single syllables.

The Kelso et al (1985) paper measures lip and jaw motion on continuous speech, with particular attention to timing of stressed vs unstressed syllables. Results are presented on a displacement vs velocity phase-plane for lip and jaw motion. The data tend to form ellipses, as shown in Figure 9.1a.

These results suggest a position vs velocity constraint upon articulatory motion that could be used in synthesis. Although the constraint of articulatory parameter motion to elliptical orbits in the phase plane is initially attractive, there are problems with implementation:

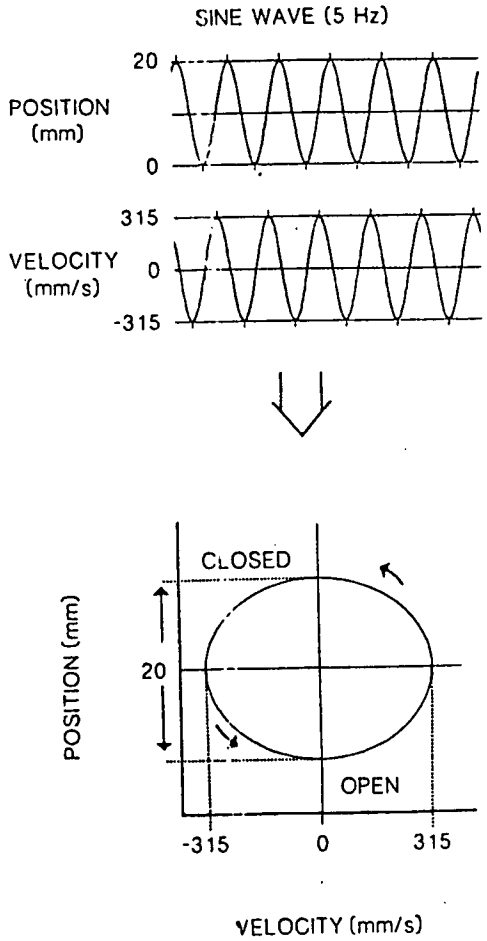
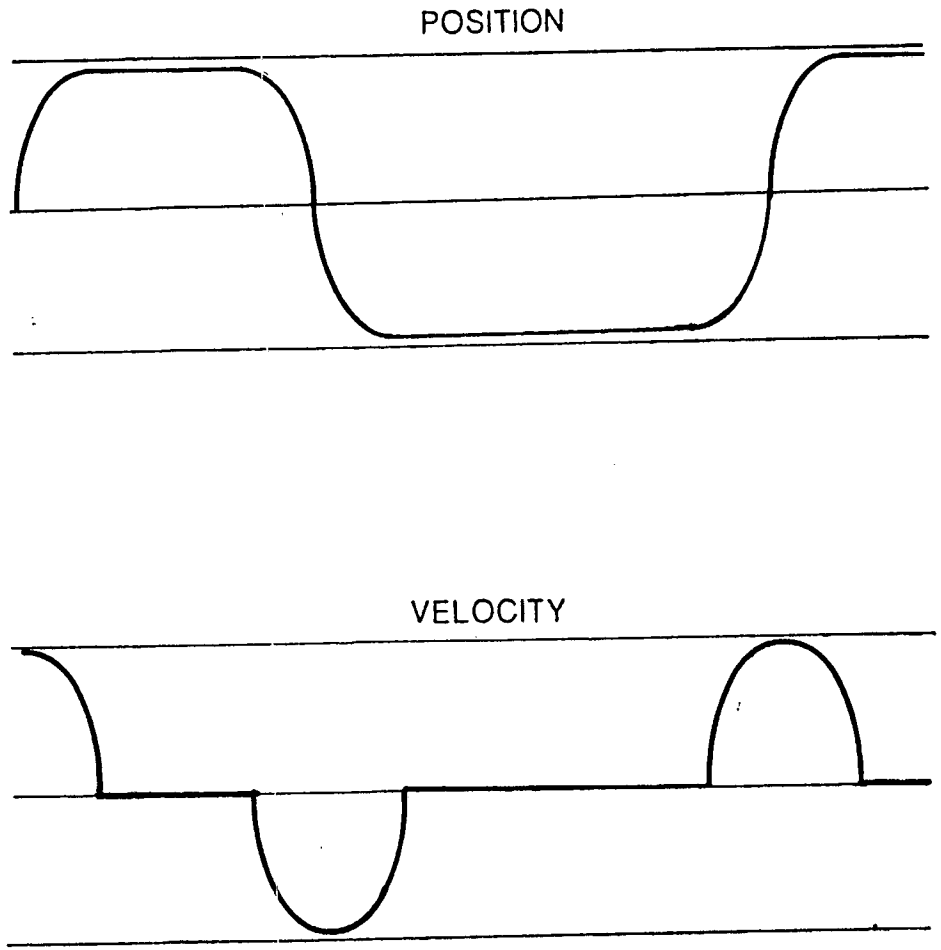


Figure 9.1-A: Original phase-plane constraint of Kelso et al (1985). Sinusoidal motion produces an elliptical trace in the plot of position vs velocity.



9.1-B: Cosine transitions with steady-state positions of arbitrary duration produce the same phase-plane trace as for the purely sinusoidal motion of 9.1-A.

1- The phase-plane constraint does not uniquely specify parameter motion. Sinusiodal motion has this property, but so also does motion which pauses indefinitely when velocity is zero, as shown in Figure 9.1b.

2- The constraint strictly applies only to motion which returns (after one cycle) to the starting position. The phase-plane method shows ellipses for "buh-buh-BUH-buh-bub" utterances (from which the Kelso et al data were measured), but the syllables of the FAAF wordlist do not, in general, return to the starting position.

A partial implementation of the phase-plane constraint was used: the constraint was applied only to speech transitions. Rather than having the complete time dimension determined by 'intrinsic' dynamics (which are incompletely specified and therefore cannot be implemented), the conventional steady-state durations were used when target values were reached.

The Kelso et al constraint was used to determine parameter motion during transitions. During a transition, position vs velocity must traverse half of an elliptical locus in the phase-plane. Only half of the ellipse is traced out because the path is from one displacement with zero velocity to a different displacement with zero velocity. The next transition then follows another semi-ellipse, and so on through the word.

The requisite relationship between displacement and velocity is implemented by using sinusoidal variation of position vs time. Another way to describe the constraint, as implemented, is that the synthesis requires cosine-function interpolation between target values.

The final test as performed in this experiment is linear vs cosine interpolation, but in the context of control of articulatory parameters rather than acoustic ones (as in Experiment III).

9.4 STIMULI

There are two approaches to the use of articulatory parameters: the targets can be treated exactly like the target data for the other five synthesizers studied in these experiments, or some consideration of articulatory dynamics can be attempted.

The FAAF words were synthesised in both of these ways, providing two sets of stimuli:

1- In the first case linear interpolation of the articulatory parameter data was used, an area function was derived, interpreted as a lossless tube, and converted to formant frequencies and bandwidths. The bandwidths were then corrected using data from the original Klatt table, and the result was converted to a discrete system function for use as a synthesizer.

2- In the second case the limited application of consideration of dynamics as discussed in the previous section leads to the simple result that transitions should follow a cosine interpolation function, and the remaining stages are as described for the first set of stimuli.

As the experimental subjects are quite capable of four or five FAAF tests in a session, two sorts of natural speech were added. Results on this material were meant to help put all the remaining results into perspective, by providing an upper bound on intelligibility. The two sets were:

3- The original, undigitised FAAF materials, including the carrier phrase.

4- The excised isolated FAAF words, digitised at 10 kHz sampling rate, 12-bit representation, but not subjected to LPC analysis-resynthsis.

Thus Experiment IV is a comparison of:

(a) articulatory parameters, linear transitions;
(b) articulatory parameters, cosine transitions;
(c) digitised natural speech;
(d) natural speech in a carrier phrase.

Additionally, it was of interest to compare results on these materials with at least two other sets of FAAF stimuli from previous experiments:

(e) LPC words;
(f) series resonance synthesizer, linear interpolation.

Experiment III showed the intelligibility of LPC words. A comparison with the results on digitised natural speech (c) allows an assessment of how much of the reduced intelligibility of the LPC data is caused by the analysis-resynthesis, and how much comes from digitising and elimination of the carrier phrase.

Finally, the whole point of investigating the articulatory parameters is to compare results with a more conventional approach. The series synthesizer with linear interpolation was shown in Experiment II to be (marginally) the best of the five 10-parameter synthesizers. Experiment III showed that cosine interpolation had the same intelligibility as had the linear method. Thus result (f) represents the best of the higher-dimesional approaches, for comparison with articulatory methods (a) and (b).

9.5 PROCEDURE

The FAAF materials as described in Experiment II were synthesised using the two articulatory approaches described above (Section 9.3). Stimulus generation and recording were as in the previous experiments.

Four standard randomisations of the 80-word list were used

for each of the stimulus types (a)-(d) as described in Section 9.4. The randomisations for stimulus types (e) and (f) had been fixed in the previous experiments. As there were only five standard randomisations and six sets of stimuli, the same randomisation was used for both the series linear synthesis (f) and the natural speech in a carrier phrase (a).

The four new stimulus types were then presented to four subjects, all of whom had previously participated in Experiment I or II, or both.

The test instructions were as for the previous experiments. Again, the subjects were offered payment according to the number of correct identifications. Again one subject had been extensively exposed to the stimuli, to get an indication of the effect of training.

The presentation of stimuli and control of level were exactly as described for the previous experiment. Again a written response was required, using printed response sheets. The task was to listen to a stimulus word, read the four response possibilities on the form, and circle one word. Each subject did five FAAF tests in a one 50-minute session, with short breaks between each test and a five minute break after the third test. The five tests consisted of the four new stimulus types a-d, and one extra test of type (e) or (f).

As not all of the Experiment IV subjects had participated in both Experiments II and III (though all had done at least one of them), the subjects took a fifth FAAF test in the same session. This ensured that for each subject there were in fact results on all six types of speech material of interest, as shown in Table 9.1.

9.6 DATA

The results in Table 9.1 give raw scores in words correct

out of the total of 80 stimuli presented in each FAAF test. Additional columns give sums of squares, sums and average percent correct. Averages by subject are also given.

Table 9.1: Intelligibility scores for articulatory parameters and control. Comparison data for series synthesis, LPC speech, and two types of natural speech. Raw scores (words correct out of 80), sums of squares, sums and average percent correct.

Subject:	1	2	3	4	Σx^2_j	$\bar{X}i.$	Ave % correct
stimuli							
carrier	79	79	78	78	24650	314	98.1
digit'd	77	72	77	76	22818	302	94.4
LPC	68	67	75	72	19922	282	88.1
series lin	67	65	76	59	17971	267	83.4
cosine artic	56	58	63	56	13605	233	72.8
linear artic	56	57	62	54	13145	229	71.6
Ave% correct	84.0	82.9	89.8	82.3			

Subject 3 had been trained before testing.

carrier = natural speech in a carrier phrase
digit'd = digitised natural speech, isolated words
LPC = linear prediction analysis/synthesis
series lin = linear interpolation, series resonance synth.
cosine artic = cosine interpolation, articulatory parameters
linear artic = linear interpolation, articulatory parameters

9.7 ANALYSIS

9.7.1 Significance of Mean Value Differences

The analysis of results for this experiment is the same as for Experiment III. The question is not whether the four new

kinds of stimuli tested exhibit as a group sufficient differences to declare them non-identical. Therefore the analysis of variance is not carried out. Instead, a standard t-test of mean-value differences is made, using the method of paired comparisons to reduce the effect of subject variability. The analysis for the series resonance synthesis (using linear interpolation) vs articulatory parameter synthesis (also with linear interpolation) is given in Table 9.2.

Table 9.2: Paired comparison data, series vs articulatory synthesis, both with linear interpolation.

	series	artic	
	Y1	Y2	d = Y1 − Y2
subject			
1	67	56	11
2	65	57	8
3	76	62	14
4	59	54	5
ΣY	267	229	$\Sigma d = 38$ $\Sigma d^2 = 406$
ave Y	66.75	57.25	D = ave d = 9.5

$S^2 = (\Sigma d^2 - [\Sigma d]^2/n)/(n-1)$

$= (406 - (1444/4)) / 3 = (406 - 361) / 3 = 15$

$S^2/n = 15/4 = 3.75$

$s = \sqrt{(S^2/n)} = 1.94$

$t = D/s = 9.5 / 1.94 = 4.9$, dof=3.

one-tailed test for series > artic, mean diff > 0.

9.7.2 Detailed FAAF Analysis

A closed-response intelligibility test allows investigation of types of error. The data in Table 9.1 give overall intelligibility scores, but the fact that errors are limited to a closed set, and a small set at that (three erroneous possibilities in the case of the FAAF test) allows more detailed analysis of just what types of error occur.

The standard analysis for FAAF tests (Foster and Haggard, 1984) provides data in several major categories. These categories are defined below. In the definitions, the phrase '% of total words' means simply that the figure is the number of tokens in this category divided by the total number of stimuli (and multiplied by 100 to make a percentage). The phrase '% of possible words' means that the number of responses falling in this category are divided by the number of stimuli for which the category is applicable (and multiplied by 100).

Correct = % of total words correct
  Init. = % of possible words correct in word-initial position
  Final = % of possible words correct in word-final position

Errors:

SF = % of total words with a Single Feature error; the correct word and the chosen word differed in only one respect. If the stimulus was 'man' and the response was 'nan' or 'van', it was a single feature error of place or manner.

DF = % of total words with a Double Feature error; a stimulus of 'man' and a response of 'than' is wrong in both place and manner.

Place = % of possible words with a error corresponding to a difference in articulatory place of contact or constriction;
  Voiced = % of possible voiced words with a place error;
  Unvoiced = % of possible unvoiced words with a place error.

Manner = % of possible words with a Manner error, a confusion between the categories stop, fricative, approximant or nasal.

Voicing = % of possible words with an error associated with periodic vs aperiodic excitation.

In/Om = % of possible words with an Intrusion or Omission of a sound. Thus responding 'milks' to 'mix' is an intrusion error, responding 'Mick' is an omission.

A more detailed analysis is possible (Foster and Haggard, 1984), breaking down the place, manner and voicing categories into 30 error types. This analysis was also completed for all the FAAF tests in all three intelligibility experiments. The results were not judged to be pertinent to these experiments, and are not presented here (but are available from the author).

The total amount of diagnostic data available is rather large. Table 9.3 gives scores in just the major diagnostic categories, averaged across the four subjects of this experiment. Scores are given for each of the six stimulus types whose intelligibility scores were given in Table 9.1. Tables 9.4 and 9.5 go on to list all the remaining major category results for the stimulus types used in all three intelligibility experiments, again averaged across subjects.

Table 9.3: Diagnostic FAAF data for the six stimulus types of Experiment IV, averaged across subjects.

	natural carrier	natural digit'd	LPC speech	series linear	artic. cosine	artic. linear
Correct	98	94	87	82	73	72
Init.	97	94	86	75	66	68
Final	99	95	88	88	78	74
Errors						
SF	2	6	13	16	24	26
DF	0	0	2	3	4	3
Place	2	6	10	6	13	16
Voiced	4	6	10	5	11	17
Unvoiced	0	5	9	8	16	16
Manner	0	0	4	10	10	8
Voicing	0	1	6	12	17	14
In/Om	0	0	0	3	4	8

Table 9.4: Diagnostic FAAF data for the five synthesizers of Experiment II, averaged across subjects.

	series resonance	direct form	reflection coeff's	area function	parallel resonance
Correct	78	70	74	74	67
Init.	74	66	74	71	72
Final	81	73	74	76	62
Errors:					
SF	20	25	24	23	29
DF	3	5	2	2	4
Place	6	9	8	10	12
Voiced	5	7	4	9	8
Unvoiced	8	11	13	12	17
Manner	12	12	13	14	18
Voicing	17	23	24	15	19
In/Om	5	8	6	8	10

Table 9.5: Diagnostic FAAF data for the interpolation methods stimuli of Experiment III, averaged across subjects. All but the LPC data are for series synthesis.

	abrupt	PWL	linear	cosine	LPC
Correct	55	67	78	81	87
Init.	44	67	74	78	86
Final	64	67	81	83	88
Errors:					
SF	38	27	20	17	13
DF	8	6	3	3	2
Place	25	13	6	6	10
Voiced	29	8	5	3	10
Unvoiced	18	20	8	10	9
Manner	12	18	12	13	4
Voicing	20	8	17	11	6
In/Om	8	10	5	3	0

abrupt = instantaneous transitions (path discontinuities)
PWL = JSRU-style piecewise linear interpolation (2-piece)
linear = simple one-piece linear interpolation
cosine = cosine function interpolation
LPC = linear predictive coding analysis/synthesis

9.8 DISCUSSION OF RESULTS

9.8.1 Overall Intelligibility Differences

The results of the tests for significant differences in mean intelligibility scores are given in Table 9.6. Only results with average intelligibility differences of more than 6% were significant (at the 0.05 level). The comparison of natural speech in a carrier phrase vs digitised natural words in isolation, with a 4% intelligibility difference, approached significance, with a probability of the difference arising by chance being less than 0.2. The 5% difference between LPC and series was not significant.

Table 9.6: Results of paired-comparisons tests on intellibility scores for various types of natural and synthetic speech (scores are in Table 9.1; D, s and t as in Table 9.2)

pair	D	s	t	significance
carrier vs digitised	3.0	1.35	2.2	NS, $p<0.2$
digitised vs LPC	5.0	1.47	3.4	$p<0.05$ *
LPC vs series	3.75	2.89	1.3	NS, $p<0.3$
series vs articulatory	9.5	1.94	4.9	$p<0.02$ *

The data in Table 9.1 are in rank order according to average recognition rate. The results in Table 9.6 are just for pairs which are next to each other in the Table 9.1 ranking. The drop in intelligibility from natural speech in a carrier phrase to digitised words to LPC words to series synthesis are of similar size, from 4% to 6%. The value of the significance test is to show that the step from digitised words to LPC speech shows a statistically significant effect on intelligibility.

The next step down is from series synthesis to articulatory synthesis. This is a 9% or 10% drop, and is the most significant step in the ranking. As to the two sorts of articulatory transitions, their results are virtually the same. This is the same situation as was obtained for linear vs cosine interpolation in Experiment III, using a series synthesizer.

There is no significant difference between the intelligibility results for LPC words and for serial resonance synthesis, indicating that the synthesis has reasonable intelligibility.

9.8.2 Implications of the Diagnostic FAAF Results

The results of the detailed FAAF analysis are spread over three tables, but certain features emerge, which will be presented in the next three subsections.

9.8.2.1 Natural vs synthetic speech

Table 9.3 gives natural speech and synthetic speech results. From the first column, natural speech (undigitised; original analogue recordings) has only place errors; no manner or voicing errors. The digitised isolated words (reduced bandwidth, reduced signal-to-noise ratio, presented in isolation) maintain this pattern, with the exception of a small voicing error (1%). The parametrically coded and resynthesised LPC speech has place, manner and voicing errors, but has about twice the percentage of place errors as for the other two categories.

For synthetic speech with a series resonance synthesizer this pattern is reversed: twice as many manner and voicing errors (in percentages) as place errors. In Table 9.4 for the other all-pole models and the parallel resonance data the same pattern is maintained.

Results on articulatory synthesis (in the final two columns

of Table 9.3) show a different pattern: similar numbers for all three types of error. But the overall intelligibility of the articulatory synthesis is below that for the other types, and so the number of errors is up overall as well as having a different distribution into place vs manner vs voicing. Thus it cannot necessarily be concluded that the pattern has changed. It may just be that when there are lots of errors rather than a few, they tend to spread more evenly.

The final notable result in Table 9.3 is that the level of place errors in the series resonance synthesizer with linear interpolation is comparable to that of the digitised natural words: about 6%. The dramatic difference between the natural speech and the synthesis is in the proportion of manner and voicing errors: natural speech has almost none, the series synthesis has 10% to 12%.

The data in Table 9.4 are for five sets of synthesis parameters, all of the same dimensionality. The distribution of errors into the major categories given in the table is similar for all five columns. The main difference between columns is that those synthesizers with a lower overall intelligibility have more errors in general, and all the column entries tend to be higher.

One difference between all five Table 9.4 synthesizers and the natural speech results of Table 9.3 is in the relationship between place and voicing. The digitised words and the LPC speech have about equal numbers of place errors for voiced sounds as for voiceless sounds. For the 10-th order synthesizers there is a preponderance of place errors for unvoiced sounds.

9.8.2.2 Effects of interpolation

Table 9.5 gives results using four kinds of interpolation on a series resonance synthesizer, with LPC speech for comparison. The largest effect is for the null or abrupt interpolation: this method mainly causes place errors. The

level of manner and voicing errors is about the same as for linear interpolation, even though that synthesis had much higher overall intelligibility. Place errors are about four times as frequent when there are no sensible transitions at all.

Experiment III described the problems with the piecewise-linear synthesis. This data had obviously unnatural nasal and velar consonants, so it perhaps does not warrant detailed consideration. It has the highest amount of manner errors, which could be accounted for by the known problems with nasals and velars. But it has fewer voicing errors than any of the synthesizers. Thus the cues used in the PWL data for differentiating voiced and voiceless sounds are more perceptible than those used for the other cases. Unfortunately this is not simply a matter of transition paths, as the JSRU method involves specification of durations as well.

Finally, there is an effect of syllable position and transition type. The abrupt transitions affected the initial positions most: recognition was 44% in initial position as compared with 64% in final positions. This can perhaps be accounted for by the fact that there are additional durational cues in syllable final position.

9.8.2.3 Articulatory synthesis

The detailed FAAF-test results using the articulatory synthesizer were presented in Table 9.3. When series resonance parameters are compared to the articulatory method (with linear transitions in both cases), the result is about the same level of manner and voicing errors, but more than twice as many place errors using the articulatory model.

The Experiment IV data in Table 9.5 show that place errors are related particularly to transitions: the abrupt transitions mainly caused an increase in place errors. The articulatory stimuli differ in two ways from the series

resonance stimuli: the target values differ (because of the approximation involved in reducing from ten to three parameters), and the transitions differ (because the interpolation is in terms of articulatory parameters rather than formants). The increase in just place errors leads to a suspicion that the reduction in intelligibility is because of the transition differences, not the target differences.

9.9 CONCLUSIONS

9.9.1 Articulatory Synthesis

The simplest conclusion is that the articulatory approach did not improve the intelligibility of the synthesis, and in fact significantly lowered the word-level intelligibility. This does not rule out an important role for articulatory considerations in the synthesis of continuous speech, or for implementing coarticulatory effects. Also the particular articulatory implementation used in this experiment is not very sophisticated (though neither were the acoustic parameter synthesizers).

Why did the articulatory approach yield a lower word recognition rate than for the series synthesizer (and indeed lower than all the other five synthesizers)? Considering just the difference in number of available parameters (three vs ten), perhaps the answer is obvious. Perhaps getting 70% intelligibility when the ten parameter approach gets 80% is not doing poorly, but well. The equivalent effect from additive noise is only a 2 dB difference in signal to noise ratio.

Finally, the articulatory synthesis was aiming at targets which were originally optimised (by Klatt) for resonance synthesis. It is possible that different targets would work better for articulatory synthesis.

9.9.2 Articulatory Dynamics

We can also conclude that the limited implementation of a constraint upon articulatory motion was of no value. The ordinary linear interpolation between targets was just as intelligible as were the transitions having an elliptical locus in the phase plane. However Experiment III showed that linear and cosine interpolation give virtually indistinguishable intelligibility results for resonance synthesis. Further, this can be accounted for by the observation that the linear and cosine parameter paths are very similar. Hence the linear transitions for articulatory parameters will also have paths which (when plotted vs time) are very similar to cosine paths.

One might say that it wasn't so much that the phase-plane constraint didn't add anything, but rather that linear interpolation was already close t+ satisfying the posited constraint. In other words, linear is already nearly the same as cosine, so they are both constrained. It is then interesting to observe that this similarity of linear and cosine transitions (when plotted vs time) does NOT extend to their phase-plane representations. Linear transitions have a constant velocity, and thus form a square in the phase-plane!

Given:

a) the similar intelligibility;
b) the similar paths as time functions;
c) the dissimilar phase-planes;

then the evidence of this experiment supports the conclusion that the phase-plane is irrelevant, at least for the effect of transition paths on word-level intelligibility of synthetic speech.

9.9.3 Intelligibility

The synthesis of all the ten-parameter synthesizers is less intelligible than for the LPC speech, though not by very much

in the case of series resonance synthesis. But the difference is wholly owing to manner and voicing errors. The series resonance synthesis has no more place errors than for the digitised natural words. The remaining ten-parameter synthesizers have about as many place errors as for LPC speech.

An increase in intelligibility must come from a decrease in manner and voicing errors. Consideration of the nature of the acoustic cues to manner and voicing (cf Appendix 6) leads to the conclusion that the synthesizer type and the interpolation type are not especially relevant. Experiment III showed most clearly that interpolation type strongly effects place, and not much else. Experiment II showed that synthesizer type had hardly any effect on manner and voicing.

What would improve intelligibility? The clearest answer is for voicing. Although there are many acoustic cues to this contrast, depending upon whether the sound is a stop or a fricative, and syllable initial or final, one major acoustic cue is the periodicity of the excitation. Thus more attention must be paid to excitation to improve on the intelligibility of the synthesis used in this study.

The manner distinctions arise from a range of phenomena. Timing affects approximant vs fricative vs stop. Acoustic parameters (including the amplitude) affect nasality. There is an excitation difference between nasals and fricatives.

The three intelligibility experiments concentrated on parametric representations of speech. All of the 10-parameter models had a property of adequacy: they could achieve a set of spectral targets. To bring any (or all) of these synthesizers to a higher level of intelligibility requires attention to matters beyond that of synthesis parameters: in particular, excitation and duration.

This investigation of parametric representations for speech synthesis ends with the realisation that the parameters are by no means the whole story. We have seen the effects of

synthesizer type and interpolation type. We have shown how simple formant transitions yield highest intelligibility. These transitions can be most easily obtained by linear interpolation of formant parameters of a series resonance synthesizer. Finally, the intelligibility of the resultant speech is as good as that of digitised natural speech, as far as place distinctions are concerned.

The manner and voicing errors require particular attention to duration and excitation, matters outside the scope of this investigation.

Chapter Ten: Naturalness comparison of ten natural and synthetic types of speech. Experiment V

10.1 OBJECT

The object of this experiment was to measure the extent to which the various types of synthetic speech were judged to be similar to natural speech.

Experiments II, III and IV involved word recognition measurements, and for this purpose the standard FAAF test procedure was used. For naturalness there is no standard test procedure, and so the description of the experiment involves a preliminary discussion of development of an appropriate test.

The background to naturalness tests is presented in section 10.2. Section 10.3 examines selection of speech material for presentation to the listeners, and Section 10.4 describes the final format of the test.

The remaining two sections present the experimental results, analysis and conclusions.

10.2 THEORY

The previous three experiments have measured intelligibility as a function of synthesizer type and interpolation method. Intelligibility at the level of recognition of individual words is obviously an important aspect of the evaluation of any speech system. Speech communication is a set of decisions at the source and at the receiver. Intelligibility tests measure how accurately the listener can make these decisions.

But speech also has an aesthetic aspect: listeners judge how much they like to listen to a particular sort of voice. These judgements are important when speech has been processed

in some fashion, because there is then much greater scope for the final signal to contain undesirable factors or characteristics.

All aesthetic factors could be combined under the heading naturalness. This does not mean that there is only a single dimension, however. There may well be a variety of factors contributing to an aggragate judgement of naturalness, just as in the subjective evaluation of the quality of concert hall acoustics.

There is less uniformity of approach in determination of naturalness than is the case for intelligibility. Evaluation of speech transmission systems has included formal procedures for obtaining human judgements on all the acceptance factors beyond intelligibility since the 1920's (Fletcher, 1929), but the problems of definition of what to measure and how to measure it remain an area of active research (Rothauser, et al, 1971).

One approach to measurement of subjective factors is the use of 'semantic differential scales'. In essence an experimental subject simply picks a number (often from one to ten) on a scale whose endpoints are labelled with descriptive terms, usually opposites (good-bad, loud-soft, rough-smooth, black-white). The words chosen do not need to correspond to any physical properties of the phenomena under examination: one can ask for loud-quiet judgements on visual patterns, and bright-dull judgements on sounds.

The resultant data are numerical, and can then be treated as quantitative measurements following the scaling procedures of ordinary psychophysics (Stevens, 1947; Torgerson, 1958). In particular, the scales are ordinal and so the experimental observations have a median value. If one can also assume equal intervals on the scale, means and standard deviations are meaningful, and lead to the possible use of the full range of parametric tests.

One difficulty with rating scales is the question of how many to use. There is a tendency in the early literature for a scattered approach, using as many scales as the experimentor could think of, or the subject could cope with. This problem was treated systematically with the introduction of multidimensional scaling (MDS,Shepard, 1962; Kruskal, 1964). The ratings on a multiplicity of semantic dimensions may be treated as a point in a space of high dimensionality. The procedure of multidimensional scaling determines how well the space can be represented in fewer dimensions.

Rather than use a proliferation of semantic scales and an MDS procedure, an examination of the literature revealed a simpler strategy. Pratt (1986) investigated intelligibility of nine kinds of synthetic speech, plus natural speech for comparison. A semantic differential rating procedure was also used, with four dimensions: intelligibility, effort, pleasantness and naturalness. A main result was a high correlation for intelligibility, effort and pleasantness. All three were also highly correlated with the actual intelligibility scores. The one largely independent factor was naturalness, as shown in Table 10.1. The minus signs on the correlation arise from the assignment of numbers on the rating scales: from one for 'completely intelligible', 'no special effort required', 'completely natural' and 'very pleasant' up to ten for 'totally unintelligible' and so forth.

Table 10.1: Correlation of word intelligibility scores (Diagnostic Rhyme Test) with semantic rating scales. The significance is the probablility that samples from two independent distributions would yield the observed correlation. (Based on Pratt, 1986)

Rating scale	Correlation	Significance
Intelligibility	-0.88	<0.01
Effort	-0.90	<0.01
Naturalness	-0.31	NS
Pleasantness	-0.79	<0.05

The correlation scores were sufficient evidence to decide not to use all four scales in the present experiment. The three which were highly correllated with intelligibility were discarded, leaving the single semantic dimension of natural vs unnatural.

10.3 STIMULI

The next issue is the actual speech material to be rated. One could test phrases or individual words; the material could be of high or low intelligibility. Individual items could be rated or a group of items could be presented and then rated collectively.

Individual FAAF words, rather than phrases, were used as stimuli. The differences between FAAF words and complete phrases are mainly prosodic: factors of stress, timing and intonation. These dimensions are identical in our materials. The varieties of synthesizer and interpolation strategy studied in these experiments do not have an effect on the time dimension or excitation signal, which form the acoustic basis for these prosodic dimensions (Wright, 1987; Appendix 6).

The issue of hard or easy words (in terms of recognition scores) seemed an open question. There might be no difference in naturalness judgements. Easy words might provide greater discriminatory power, if the subjects were able to be more critical of naturalness when confident of word identity. Hard words might make the best test, because they would be most generally demanding.

A test with both hard and easy words could be used to try to resolve the uncertainty about which were best. Recognition difficulty was obtained simply by combining all the results from the FAAF tests of the previous experiments. Appendix 5 gives the recognition difficulty rank orders as determined from the intelligibility results of Experiments II, III and IV combined (61 FAAF tests in all). The appendix also gives

the rank order data published by the Institute of Hearing Research (IHR) (Foster and Haggard, 1984, p22), based on natural speech in noise.

There is a problem with the definition of words which are hard or easy to recognise. If the recognition is measured with a closed response test (such as the FAAF procedure), the recognition scores depend not just on the individual words but on the whole response set.

A clear example from the FAAF data is the word BAG. In the set:

BAG BACK BAT BAD

the word BAG scores rather well: one of the eight easiest in the IHR data, one of the 20 easiest for synthesis.

But in the set:

BAG BANG BAN BAD

the same word BAG is much more likely to be unintelligible. It drops from a rank of 73 down to 60 in difficulty (where one equals most difficult) for the IHR natural speech, and is now one of the 20 hardest for synthesis.

Thus difficulty involves consideration of the entire response set. These set rankings are also listed in Appendix 5. The hardest set for both synthetic and natural speech is:

MAN VAN NAN THAN

The easiest set for synthesis (and second easiest for natural speech is:

MIX MICK MILK MILKS

Table 10.2 shows how these words scored in the recognition experiments (II, III and IV). Although these are the easiest

and hardest sets, the individual words in each set are not of uniform difficulty. In particular, the word MAN is well recognised, and in fact scores better than does the hardest word in the easy set (MICK).

Table 10.2: Intelligibility of individual words and word groups, based upon results of Experiments II, III and IV.

word	% intell.	word	% intell.
milk	100.0	man	85.2
mix	100.0	than	57.4
milks	95.1	nan	32.8
mick	68.8	van	24.6
ave	91.0	ave	50.0

The next section will discuss how these hard and easy FAAF words were used to make up a naturalness test.

10.4 PROCEDURE

The considerations in sections 10.2 and 10.3 above led to the choice of isolated FAAF words as stimulus items in a semantic differential rating test. Only one test dimension, labelled "completely natural - totally unnatural" was to be used. The ratings would be on the scale one to ten.

A subset of tokens from the FAAF wordlist would be used, consisting of easy words and hard words (in terms of their recognitions scores as obtained in Experiments II, III and IV).

These choices do not define a test procedure; they give a framework. There still remain decisions about how many words, how many types of synthesis to test, how many

naturalness judgements to ask for, training and instructions for the subjects, and what the subjects should be doing besides listening for naturalness.

There are two general possibilities for administering a subjective rating procedure for speech materials, which can be denoted the block method and the item method. The more usual procedure is the block approach: an entire passage or wordlist is heard, and one rating (per scale) is made. The study of Pratt (1986) followed this approach. Subjects performed an entire Diagnostic Rhyme Test for one type of synthesizer, and then made one rating on each of the four semantic scales being tested.

The block method requires attention to the problem of consistency of ratings across trials. In the case just cited samples of all the synthesizer types were played before each test, and subjects were specifically instructed to use the whole numerical range in their responses.

In the present experiment it was of interest to test the discriminatory power of hard vs easy words. This question could have been tested with a block approach, but could also be tested with the item method. The item method is simply to require a rating for every word. The main difference is that whereas a block approach would require two test sessions, one with hard words and one with easy words, the item approach allows all the words to be mixed in one session.

Once the strategy of 'one rating per stimulus' was chosen, it also became possible to consider mixing the types of synthesis within the one test. Providing the number of test words times the number of systhesis types is not too large, all the naturalness ratings could be obtained in one test. Problems of consistency across sessions are eliminated, or at least redefined to a problem of consistency within a test session. The multiple sessions are removed, though there is still no guarantee that subjects will perform consistently during the one session.

In the present case, the whole variety of speech types could be contained within the one test. It was still necessary to encourage subjects to use the full range of responses, and this was made part of the written instructions (Appendix 5).

In order to keep the subjects occupied with something beyond naturalness, they were also required to identify the words as in an ordinary FAAF test. This was meant to make the test situation more representative of the real uses of speech. We rarely only judge speech quality. Speech is usually heard for some purpose, and so listening for recognition as well as naturalness is a more realistic task.

Low and high intelligibility words with full sets of four words each implies a minimum of eight test items. Using all eight words just once for each of N types of speech to be rated makes a total test of 8N items. Experience with the 80-long FAAF tests led to the conclusions that an N of around ten would be practical.

The tests of experiments II, III and IV had involved 13 separate FAAF tests. Some of these were of more interest than others:

1- There were two sorts of natural speech, the original IHR analogue recordings using complete phrases, and the extracted, digitised single words. Just the digitised words would suffice for comparison with the synthetic materials.

2- Two of the four kinds of interpolation lead to poor recognition and obviously unnatural quality, namely the abrupt or discontinuous data, and the piecewise linear data. They could thus be eliminated, leaving only cosine and linear as interpolation types.

3- There seemed little point in testing linear vs cosine interpolation for both the series and articulatory synthesis, so one of these could be dropped.

These considerations reduced the list of relevant speech types to nine: the six synthesizer parameter sets (with linear interpolation), one extra interpolation type, natural words, and LPC words. As ten synthesis types would make an 80-item test, it was decided that one new synthesis type would be added: deemphasised (integrated) speech, meaning tilting the spectrum to have a -6 dB/octave average slope.

Speech is usually modelled as having a 6 dB/octave rolloff with frequency, as discussed in Chapter Two. The synthesis used in Experiments II, III and IV had a flat excitation, and no deemphasis was applied. This excitation gives less amplitude to the low-frequency components in the speech spectrum. The subjective effect is a lack of bass, a 'small' voice.

There are two reasons to add deemphasised speech to the list of synthesis types to be tested. First the lack of deemphasis is an obvious difference between the simplified synthesis strategy employed in these experiments, and the more complex practical synthesizers. Secondly, this difference in spectrum shaping is known to affect naturalness (Witten, 1982, p96), and so should lead to a difference in the ratings obtained in this experiment.

The ten synthesis types used, and the four types not used, are listed in Table 10.3.

The final format for the experiment was to test ten speech types (natural, LPC and eight varieties of synthesis), with eight test words for each speech type (four hard and four easy words), making an 80-item test just as in a FAAF intelligibility test.

Table 10.3: Types of synthetic and natural speech tested in Experiment V, and the four synthesis types from previous experiments not tested in Experiment V.

Tested	Not tested
1 natural words, isolated and digitised	11 natural words in phrases
2 Linear Predictive Coded words	12 series synthesizer, abrupt transitions
3 Series synthesizer, cosine transitions	13 series synthesizer, piecewise linear interp.
4 Series synthesizer	14 articulatory parameters, cosine interpolation
5 Series synthesizer, deemphasised (-6 dB/octave)	
6 Direct form	Synthesis types 4-10 all used
7 Reflection coefficients	linear interpolation
8 Area function	
9 Articulatory parameters	
10 Parallel	

Although the test appeared to be a concise way to test many varieties of speech, and did not involve multiple sessions and their associated problems of consistency, the procedure adopted was not without its own pitfalls. The use of the same eight words ten times over in one test made the procedure much more repetitive than for a conventional FAAF test. Also it would be possible to lose track of the correct place on the response sheet, as there were only two response sets repeated 40 times. [In fact none of the subjects did get lost.]

As there were ten kinds of speech to be tested, it seemed appropriate to have ten subjects. Three had participated in one or more of the previous experiments. Seven had no experience of synthetic speech. As in all the other subjective experiments, one subject had extensive training on all the stimuli before testing.

All 80 stimuli were fully randomised. The recording and presentation of the materials were exactly as described in the previous experiments. The only real differences between this test and a standard FAAF test were:

1 Only eight words repeated ten times, rather than 80 words once each.

2 Two response sets rather than twenty.

3 Subjects had two parts to their task: circle the word heard (as in a FAAF test), and circle a number between one and ten to indicate place on the scale from 'completely natural' = 1 to 'totally unnatural' = 10.

The response sheets included a definition of the naturalness scale, to prevent subjects from reversing the scoring by mistake. This information appeared at the top and again halfway down each page of the four pages of the response sheets.

The tests lasted about ten minutes, and proceeded with no apparent difficulties. The next sections will give the resultant data and discuss results.

10.5 DATA AND RESULTS

There are two sets of scores for the procedure used in this experiment: word intelligibility scores, and naturalness ratings.

10.5.1 Intelligibility

Table 10.4 gives word recognition scores for each of the ten speech types. The scores are given as a percentage correct. Because there were ten subjects, each score in Table 10.4 is based on 80 trials (8 words, ten subjects). Note that these scores can NOT be compared directly with the intelligibility

scores of the previous experiments, because the Experiment V scores are only based on eight of the original 80 FAAF words.

Table 10.4: Overall average percent intelligibility (all eight words) for all ten speech types. N=80 (8 words, 10 subjects).

type of speech material	%correct	significance, n=4	
		t	probability
natural (digitised)	77.5	6.55	p<0.005 **
LPC words	76.2	6.24	p<0.005 **
Series cosine	46.2	0.42	NS
Series linear	43.8	(reference condition)	
Series lin, deemph.	47.5	0.50	NS
Direct form	45.0	0.12	NS
Reflections	47.5	0.88	NS
Area function	48.8	0.39	NS
Articulatory	46.2	0.25	NS
Parallel	53.8	1.85	p<0.1

significance: *=0.05 level; **=0.01 level; ***=0.001 level

As expected, natural speech has the highest intelligibility. The LPC speech is virtually the same. All the synthesis is of much lower intelligibility, about 30 percentage points lower. This would represent a Signal to Noise ratio equivalent reduction of about 5 dB, as the standard FAAF material has a slope of 6% per dB for scores in the 40% to 80% range (Foster & Haggard, 1984, p20).

Significance levels for the word recognition could be computed from the differences in the mean value across subjects, just as for the data in the previous experiments. However in this experiment there are only eight stimuli per speech type per subject, and in the other experiments there were always 80 stimuli. This reduction in responses makes the subject scores much more uneven, and the effect is only partly offset by the fact that the present experiment uses twice as many subjects.

Another way to get some division of the data (to provide an estimate of the mean and variance of the recognition scores for each speech type) is to average across subjects, and use the scores for each word. This would provide ten data points per recognition score. A way to increase the data points (at the expense of number of scores) is to combine words into groups, a sort of averaging across words and subjects. The method chosen was to take pairs of words, selected according to similar overall recognition level.

There are four pairs, and 20 trials (ten subjects times two words) per pair for each synthesizer type. Significance levels in Table 10.4 are computed from a t-test of the differences in the mean recognition rates. The mean in this case is the mean of the four wordpair scores, with each score based upon 20 responses. The details are in Appendix 5. A t-test is performed for each speech type vs series linear synthesis, because this synthesis type has been a reference throughout the experiments, and because (in this case) it represents one end of the range of recognition scores.

Intelligibility can also be averaged across the speech materials, yielding an average percent correct for each word. This result is presented in Table 10.5. Here, each percentage in the table is based on 100 trials: ten speech types and ten subjects. The results can be compared with Table 10.2. Note that the Table 10.2 data comes from 61 trials per word, with a different combination of stimuli than is the case for Experiment V, and with different subject groups.

Table 10.5: Intelligibility of individual words and word groups used in Exp V. (Compare with Table 10.2)

word	% intell.	word	% intell.
milk	80	man	73
mix	65	than	31
milks	67	nan	39
mick	42	van	29
mean	63.5	mean	43.0

The data in Table 10.4 and Table 10.5 are peripheral in this experiment; recognition has been measured in the earlier experiments, and here the main interest is naturalness. The scores in Table 10.5 provide a check as to whether the words really were hard and easy. Also the rank order can be compared with the results in Table 10.2, and are seen to be in near agreement. Finally, the data of Table 10.5 were used to pick similarly scoring wordpairs for use in the significance tests for Table 10.4.

10.5.2 Naturalness

The point of Experiment V is to obtain naturalness ratings, and the results averaged over all eight words (hard set and easy set) are given in Table 10.6. Each rating in the table is an average of 80 responses (eight words, ten subjects). These data will be discussed in Section 10.6.

Table 10.6: Overall average naturalness (all eight words) according to type of speech. N=80 (8 words, 10 subjects). 1=completely natural, 10=totally unnatural.

type of speech material	rating	significance z	prob	
natural words, digitised	2.41	-17.9	<0.0001	***
LPC words	3.54	-12.4	<0.0001	***
Series cosine	5.98	-0.69	0.25	
Series linear	6.20	0.37	0.36	
Series linear, deemphasised	5.74	-1.84	0.033	*
Direct form	5.95	-0.84	0.20	
Reflection coefficients	6.34	1.04	0.15	
Area function	6.11	-0.07	0.47	
Articulatory parameters	6.28	0.75	0.23	
Parallel	6.39	1.28	0.10	

overall: mean 5.49 synthesis only: mean 6.12
sd 1.30 sd 0.21

significance: *=0.05 level; **=0.01 level; ***=0.001 level

The significance computation for Tables 10.6, 10.7 and 10.8 is an estimate rather than an actual t-test, because of difficulties in applying an exact test. The procedure is given in Appendix 5.

Results obtained from easy vs hard sets of words were also examined, and the data are in Table 10.7. The format is as in Table 10.6, except there are two lists of results, for the hard and easy sets of words. Each part of the results has an overall mean, and a mean just computed for the eight types of synthetic speech (used for significance scores).

In addition there is an actual t-test (not a Z-score approximation) for the difference between the hard and easy set means. The test is performed first on the overall means

and then on the 'synthesis only' means. The method is the paired comparisons procedure as used for t-tests in the earlier experiments. The interpretation of these results is presented in Section 10.6.3.

Table 10.7: Naturalness as a function of word difficulty.

Stimuli	Hard set	significance		Easy set	significance	
		z	prob		z	prob
natural (digitised)	2.40	-9.37	<0.0001***	2.42	-15.4	<0.0001***
LPC words	4.65	-2.94	0.0016**	2.42	-15.4	<0.0001***
Series cosine	5.80	0.34	0.37	6.15	-1.56	0.059
Series linear	6.02	0.97	0.17	6.38	-0.70	0.24
Series lin, deemph.	5.10	-1.66	0.048 *	6.38	-0.70	0.24
Direct form	5.38	-0.86	0.19	6.52	-0.19	0.42
Reflections	5.75	0.20	0.42	6.92	1.30	0.097
Area function	5.48	-0.57	0.28	6.75	0.67	0.25
Articulatory	5.58	-0.29	0.39	6.98	1.52	0.064
Parallel	6.30	1.77	0.038 *	6.48	-0.33	0.37
overall mean	5.25			overall mean	5.74	t signif.
sd	1.05			sd	1.68	1.40 0.2 NS
synthesis mean	5.68			synthesis mean	6.57	t signif.
sd	0.35			sd	0.27	2.61 <0.05 *

significance: *=0.05 level; **=0.01 level; ***=0.001 level

Naturalness might vary according to whether or not the word was heard correctly. Table 10.8 shows the naturalness ratings divided according to correct or false recognition. Further discussion of the implications of the data are deferred to Section 10.6.3.

Finally, the interaction of recognition with word difficulty can be examined. The basic data consist of 2x2 contingency tables, as shown in Table 10.9 for several types of speech. All ten contingency tables are in Appendix 5.

Table 10.8: Naturalness as a function of word recognition.

Stimuli	Correct	significance		False	significance	
		z	prob		z	prob
natural (digitised)	2.21	-14.4	<0.0001***	3.11	-9.81	<0.0001***
LPC words	3.25	-10.5	<0.0001***	4.47	-5.42	<0.0001***
Series cosine	5.87	-0.81	0.21	6.07	-0.26	0.40
Series linear	6.37	1.04	0.15	6.07	-0.26	0.40
Series lin, deemph.	5.87	-0.81	0.21	5.62	-1.71	0.044 *
Direct form	5.86	-0.81	0.21	6.03	-0.26	0.40
Reflections	6.55	1.70	0.045 *	6.14	-0.032	0.48
Area function	6.15	0.22	0.41	6.07	-0.26	0.40
Articulatory	5.78	-1.15	0.13	6.70	1.77	0.038 *
Parallel	6.28	0.70	0.24	6.51	1.16	0.12

overall mean	5.42	overall mean	5.67	t	signif	
sd	1.38	sd	1.03	1.75	0.15	NS
synthesis mean	6.09	synthesis mean	6.15	t	signif	
sd	0.27	sd	0.31	1.16	0.3	NS

significance: *=0.05 level; **=0.01 level; ***=0.001 level

Table 10.9: Interaction of word recognition and word difficulty with respect to naturalness scores.

natural speech		hard	easy
	correct	2.08	2.29
	false	2.88	5.00
LPC		hard	easy
	correct	4.76	2.19
	false	4.19	4.50
Series cosine		hard	easy
	correct	5.36	6.17
	false	6.04	6.12

Only three 2x2 tables are shown in Table 10.9, because all the types of synthetic results were similar to those for the series synthesizer (cosine interpolation) case. These results are discussed in the next section.

10.6 CONCLUSIONS AND DISCUSSION

10.6.1 Naturalness

The main result for this experiment is Table 10.6, giving naturalness scores for all eight stimulus words taken together. The conclusions to be drawn from these data are:

1) The natural speech samples (digitised words and LPC words) are very different from all the synthesis types.

2) Deemphasis made a larger difference to the synthesis than did any of the other factors tested.

All the synthesis was very unnatural. The different types of synthesis had scores of about 6 on a scale from 1 to 10 (1=completely natural, 10=totally unnatural), with a standard deviation of only 0.21. There was only one significant difference among all the synthesis types, and that was from the use of deemphasis (-6 dB/octave spectrum rolloff).

Deemphasis did not significantly affect intelligibility (Table 10.4), but did improve naturalness. It did not begin to make the synthesis anything like as natural as LPC speech.

Another conclusion is that naturalness is more sensitive than intelligibility. There is a substantial difference between the naturalness of LPC speech and the digitised natural words, but no significant difference in the intelligibility of these two speech types. The LPC speech has reduced naturalness, which is to be expected as it has had much more processing than was the case for the digitised words.

10.6.2 Intelligibility

Intelligibility tests were not the focus of this experiment. The recognition task was performed mainly to keep subjects occupied and attending, and help keep them in the appropriate place on the response sheets. However the recognition performance was in line with previous tests, as can be seen by a comparison of Table 10.5 with Table 10.2.

In both tables the worst-scoring word had a recognition rate of just above 25%. As it is a four alternative test, chance performance is 25%. It is comforting that the lowest scores approached chance, and from above rather than from below.

Overall recognition rates in Table 10.5 are lower that those reported in Table 10.2. There are many differences between the data sets upon which the two tables are based: different task (naturalness judgements as well as word recognition task), different combination of synthetic and natural speech types, different number of responses, different subject groups, as well as the different test format: 80 items vs 'eight items repeated 10 times'. Nevertheless the scores in Table 10.5 are well down on the scores in Table 10.2. If the other factors were not present, we might conclude that performance had suffered because listening to eight words ten times each is remarkably boring.

10.6.3 Efficient Stimuli for Naturalness tests

A procedural question which the data from this experiment can begin to answer is: what kinds of words should be used to test for naturalness? In particular, hard vs easy words (in terms of recognition), and whether or not it matters if the word is correctly recognised.

These two divisions of the total data are presented in Tables 10.7 and 10.8. The t-tests for significance of a difference in mean value showed that it did not matter whether words

were heard correctly or not (Table 10.8). However it does significantly affect naturalness if 'hard' rather than 'easy' words are used (Table 10.7), but only for the eight varieties of synthetic speech. These findings need fuller examination.

10.6.3.1 Hard vs easy sets

In general, the difficult set MAN VAN NAN THAN was judged more natural than were the easy words MIX MILKS MILK MICK, so far as synthetic speech was concerned. But when natural and synthetic speech were lumped together, the difference became non-significant. Inspection of the data reveals that this is because the result for LPC speech is markedly in the reverse direction: the MAN VAN NAN THAN words were much more unnatural than were the MIX MILKS MILK MICK words.

This isolated result on just the LPC speech may reflect the fact that the hard set involves voiced fricatives, which in natural speech are a mixture of random and periodic excitation. The type of LPC synthesis used in this study does not allow for a mixed excitation, and so the LPC may be particularly unnatural for the hard set words.

One clear effect in Table 10.7 is that the easy set was inadequate for distinguishing natural speech from LPC, whereas the hard set worked very well. The data do not allow determination of whether this result is because of the recognition rate difference between the hard and easy sets, or the difference in excitation types. The result is still both clear and important: a practical tool for testing naturalness should be expected to separate LPC speech from ordinary digitised speech.

One might be hesitant to note anything about the hard vs easy sets if their difference only affected natural speech vs LPC, but as shown in Table 10.7 there was a significant difference for all the types of synthetic speech, taken as a group. The difficult words were judged to be significantly more natural than were the easy words.

10.6.3.2 Naturalness and recognition

The problem with testing hard vs easy and correct vs false is that they are not separate questions. By definition, the hard words were those most likely to be falsely identified. A way to examine the interaction is through the 2x2 contingency data of Table 10.9.

The hard set vs easy set effect is apparent in the two-way tables of Table 10.9: the difference between natural (digitised) words and the LPC words is only found in the hard set results. The easy set results cannot distinguish natural speech from the LPC speech.

The pattern shown in Table 10.9 (and the remaining 2x2 contingency tables in Appendix 5) is that correct vs false does not matter for synthetic speech. This is corroborated by the 'synthesis mean' results in Table 10.8: only a difference of 0.06 in the means, and a t-test value of 0.3, which is not at all significant. Correct vs false only matters for natural speech and LPC speech, and then only for the words in the easy set.

The conclusion is that for the purposes of naturalness testing it is immaterial whether the words are correctly or incorrectly identified. The exception is the case of MIX MILKS MILK MICK for natural and LPC speech. Here and only here the distinguishing factor was correct vs incorrect, and it didn't matter (for naturalness) whether the words were natural or LPC speech.

Why should the 'false, easy' words be so unnatural in the case of natural and LPC speech? One factor is that 'easy' natural and LPC words are very unlikely to be falsely identified. The actual number of responses underlying the 'easy word wrongly identified' data in Table 10.9 are very small: two for natural speech, four for LPC. Statistically the category 'easy+false' should be avoided: it represents a tiny number of responses and may be aberrent.

10.6.4 Summary of Conclusions

The main conclusions supported by the results of Experiment V can be summarised as follows:

1- All the synthesis was very unnatural.

2- The digitised words and LPC words were much more natural than all the synthetic types.

3- Deemphasis made a larger difference to the naturalness of the synthesis than did any of the other factors tested, and was the only significant factor.

4- The differences in parameter type and interpolation type were insignificant, so far as naturalness is concerned.

5- Although deemphasis significantly affected naturalness, it did not significantly affect intelligibility.

6- LPC speech was significantly less natural than was the digitised human speech, but only so far as the hard set words MAN VAN NAN THAN were concerned.

7- Naturalness is more sensitive than intelligibility. Digitised and LPC speech differed significantly in naturalness, but not in recognition rate. Deemphasis affected naturalness but not intelligibility.

8- It was immaterial whether the stimulus words were correctly or incorrectly identified.

9- The test format was successful: it allowed natural vs LPC vs synthetic speech discrimination, and was sensitive to the effects of deemphasis. Also it allowed ten types of speech to be rated in a single ten minute test. But the test format was probably boring, it was possible to lose one's place on the response sheet, and overall recognition rates were probably lower than for standard FAAF tests.

Chapter Eleven: Conclusions

11.1 GENERAL CONCLUSIONS

This research investigated six types of speech synthesizer parameters:

1 Series resonance
2 Parallel resonance
3 Direct recursive form
4 Lattice form using reflection coefficients
5 Area functions
6 Articulatory (tongue factors, lip opening, loss)

Conclusions based upon formal properties of the parametric representations are:

1- The first four synthesiser types are all-pole models. They are thus formally equivalent. However for synthesis-by-rule the interpolation between phoneme target endpoints of these different parameters produces different formant paths: different motion of resonance centre frequency and bandwidth during the interpolation.

2- The parallel resonance model is theoretically equivalent to the serial resonance model, but only for:

(a) identical bandwidths;
(b) parallel amplitudes constrained to be the coefficients of a partial-fractions expansion of the serial system function.

In practice it is usual for parallel resonance synthesisers to have fixed bandwidths. During interpolation the amplitudes of the resonances will commonly be altered, breaking constraint (b). In general the parallel connection will not be equivalent to the serial one.

The sixth model, using articulatory parameters, was investigated through the mapping of acoustic consequences of articulatory parameters. Vocal tract shape is determined by two tongue factors, a two-dimensional space. Any function of the tongue factors can plotted as a contour over the space. This process was performed for five formant frequencies and bandwidths, and repeated ten times for ten values of lip opening.

Acoustic parameters were computed using the lossless tube model of a vocal tract. The resultant plots showed a smooth relation between tongue position and formant frequency, but there were areas of rapid change of formant bandwidth. It was concluded that the articulatory model was adequate for control of formant frequencies for speech synthesis, but that the lossless tube was inadequate for control of bandwidths.

The experimental work with the synthesisers consisted of an objective assessment of parameter transitions, and the subjective measurement of intelligibility and naturalness. The results of these experiments will be described in the next two sections.

11.2 OBJECTIVE RESULTS

Experiment I measured formant frequency and bandwidth differences for linear interpolation of the different types of synthesizer parameters, and showed that the paths differed in ways which were larger than the difference limen for steady formants (though not greatly larger).

The main conclusions were:

1- All all-pole models (the four studied in this work, and any others) are equivalent in the steady state. There are no phoneme target value differences for these models; they only differ in formant-motion implications of interpolation in the original parameter spaces. These differences are generally small. Formant

bandwidths are more affected than are centre frequencies.

2- The lossless tube model (loss only at the lips) was used in this study to implement a vocal tract derived from a set of articulatory parameters. The model proved inadequate for determination of reasonable resonance bandwidths. Bandwidths in the lossless tube change very rapidly for small changes in articulatory setting.

3- If control of resonance parameters is to be maintained throughout the interpolated path (transition) between targets, then only a resonance synthesizer should be used. All other all-pole representations move the resonances in unpredictable ways during transitions.

4- The parallel resonance arrangement has implicit zeroes. Their location is unpredictable. The zeroes can have a sizeable effect upon the steady-state spectrum, particulary in the region below the first formant.

5- Only the series resonance set of parameters gives complete control of the resonance parameters and steady-state spectrum throughout the synthesis process.

11.3 SUBJECTIVE RESULTS

Two types of subjective experiment were performed: measurement of single-word intelligibility using the Institute of Hearing Research FAAF test, and measurement of naturalness using a procedure devised as part of this research.

11.3.1 Intelligibility

Experiment II performed an intelligibility test, and found a small but statistically significant difference. Series

resonance parameters scored highest, parallel resonance lowest, and the rest with nearly identical scores in the middle.

Experiment III studied the effect upon the intelligibility of synthesis using series resonance parameters and different types of interpolation:

a) linear
b) cosine
c) piecewise linear
c) abrupt (a jump from one target value to the next)

The main conclusions were:

1- There is no significant difference between linear and cosine interpolation methods.

2- Simple single-line linear interpolation used in this study works quite well: within five percentage points of the score on the linear predictive coded speech.

3- A sophisticated interpolation method can give poor intelligibility if not properly implemented. Use of a piecewise-linear method (designed for the JSRU parallel resonance synthesizer) on the modified Klatt data, and cotrolling a serial configuration with no nasal channel, simply reduced intelligibility. Because simple linear interpolation is already within five percentage points of the score for LPC speech, there is little room for improvement, and great potential for degredation.

4- Although simple interpolation does well, this is not because interpolation is irrelevant. A complete lack of interpolation has a significant effect on intelligibility, causing a drop of 25 to 30 percentage points.

5- Finally, interpolation cannot be considered apart from duration. Even simple linear interpolation may

allow a transition to have an appropriate duration. Eliminating interpolation by using abrupt transitions can equivalently be interpreted as eliminating durations. One might conclude that what really matters is an appropriate transition duration.

Experiment IV was a study of articulatory synthesis. The conclusions were:

1- The articulatory approach lowered the word-level intelligibility from about 80% to about 70%, equivalent to a 2 dB difference in signal to noise ratio. Perhaps this is quite reasonable performance considering the reduction in parameters from ten to four.

2- The limited implementation of a phase-plane constraint upon articulatory motion did not affect intelligibility. The ordinary linear interpolation between targets was just as intelligible as were the transitions having an elliptical locus in the phase plane. This can be accounted for by the fact that linear and cosine transitions are similar.

3- All the ten-parameter synthesizers are less intelligible than the LPC speech, though not by very much in the case of series resonance synthesis. The difference is wholly owing to manner and voicing errors. The series resonance synthesis has no more place errors than for the digitised natural words. The remaining ten-parameter synthesizers have about as many place errors as for LPC speech.

4- An increase in intelligibility must come from a decrease in manner and voicing errors, for which the synthesizer type and the interpolation type are not especially relevant.

5- All of the ten-parameter models were adequate in that they could achieve a set of spectral targets. Higher intelligibility requires improved excitation and duration.

11.3.2 Naturalness

Experiment V was a test of the naturalness of ten types of speech material:

1 Natural words, isolated and digitised at 10 kHz, 12 bits
2 Linear Predictive Coded words (10th order LPC)
3 Series synthesizer, cosine transitions
4 Series synthesizer, linear transitions
5 Series synthesizer, deemphasised (-6 dB/oct)
6 Direct form
7 Reflection coefficients
8 Area function
9 Articulatory parameters (two tongue factors, lips, loss)
10 Parallel resonance synthesizer

Synthesis types 4-10 all used linear interpolation.

The main conclusions were:

1- All the synthesis was very unnatural.

2- The natural speech samples (digitised words and LPC words) were very different from all the synthetic types.

3- Deemphasis was the only factor to significantly affect naturalness.

4- The differences in parameter type and interpolation type were insignificant.

5- Deemphasis did not significantly affect intelligibility.

6- LPC speech was significantly less natural than was the digitised human speech, but only for words involving voiceless fricative and nasal sounds.

7- Naturalness was more sensitive than intelligibility.

8- It was immaterial whether the stimulus words were correctly or incorrectly identified.

9- The test format was successful: it allowed natural vs LPC vs synthetic speech discrimination, and was sensitive to the effects of deemphasis.

11.4 CONCLUDING REMARKS

This investigation of synthesizer parameters (representations of the filter within a source/filter model) ends with the realisation that choice of parameters and interpolation methods is by no means the end of the story. We have shown the relatively minor differences arising from synthesizer parameter type and interpolation type, and have shown that simple formant transitions yielded highest intelligibility. These transitions can be most easily obtained by linear interpolation of formant parameters of a series resonance synthesizer. Finally, the intelligibility of the resultant speech was as good as that of digitised natural speech, so far as place distinctions were concerned.

However the synthetic speech is markedly unnatural, and intelligibility is limited by the high level of manner and voicing errors. These factors cannot be expected to be greatly altered by further attention to synthesizer parameters and their interpolation paths, but rather depend largely upon excitation, duration, overall spectrum shape, and more appropriate phoneme target values.

The FAAF test wordlist was very helpful in the development of the synthesis used in this research. Speech was always generated within an environment of attention to phonemic contrasts, rather than generating isolated words or phrases and making only qualitative observations on the result. The synthesis development proceeded with continuous feedback of implications for intelligibility. Then use of the FAAF test itself gave further information concerning factors

limiting intelligibility. This approach to synthesis, though also advocated by Klatt (1980b, p988), is not in general use.

The point of this research was to compare parameters, not obtain the best possible synthesis. Yet one would like to see just how much improvement could be made to the type of synthesis studied in this research, as a principled and controlled way of assessing which factors contribute to intelligibility and naturalness, and by how much. The approach used in this study, synthesis of an intelligibility test wordlist, is felt to be an excellent tool for any such further investigation.

References

Titles of the following collections of papers are abbreviated:

BK = Flanagan, J L and Rabiner, L J (Eds), Benchmark Papers in Acoustics: Speech Synthesis. Stroudsburg, PA: Dowden, Hutchinson & Ross, 1973

GALF = Carre, R, Descout, R and Wajskop, M (Eds). Articulatory Modeling and Phonetics (Proc of Symposium) Grenoble: GALF Groupe de la Communication Parlee, 1977

FF = Lindblom, B and Ohman, S (Eds). Frontiers of Speech Communication Research. London: Academic Press, 1979 (Fant Festschrift)

Ainsworth, W A (1974). Performance of a speech synthesis system. Int J Man-Machine Studies, 6, 493-511.

Alderson, P R, Kaye, G, Lawrence, S G C, and Sinclair, D (1984). A speech analyser based on the IBM Personal Computer. IOA Autumn Conference, Windermere. Proc Inst of Acoust, 236-239.

Allen, J (Ed) (1980) Notes on the MITalk-79 System for Conversion of Unrestricted English Text to Speech. MIT.

Atal, B S (1985). Linear predictive coding of speech. In Fallside, F and Woods, W A (eds), Computer Speech Processing: London: Prentice-Hall.

Atal, B S and S L Hanauer (1971). Speech analysis and synthesis by linear prediction of the speech wave. J Acoust Soc Amer 50, 637-655.

Atal, B S, Chang, J J, Mathews, M V and Tukey, J W (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J Acoust Soc Amer 63, 5, 1535-1555.

Atal, B S and Remde, J R (1982). A new model for LPC excitation for producing natural sounding speech at low bit rates. Proc ICASSP, 614-617.

Atashroo, M A (1976). Pole zero modelling and its applications to speech processing. 108pp. Dept Comp Sci, Univ Utah, Salt Lake City, Utah.

Beddoes, M P (1982). Direct sample interpolation speech synthesis. IEEE T-ASSP Dec 82, 825-32.

Bladen, R A W and Lindblom, B (1981). Modelling the judgement of vowel quality differences. J Acoust Soc Amer 69, 1414-1422.

Brantingham, G L (1980). Parameter interpolation for speech synthesis systems; US patent 4,189,779.

Bridle, J S and Chamberlain, P M (1983). Automatic labelling of speech using synthesis-by-rule and non-linear time alignment. Speech Communication 2, 187-9.

Brooks, S, Fallside, F, Gullian, E and Hinds, P (1981) Teaching vowel articulation with the computer vowel trainer: methodology and results. Br Jour Audiology, 15, 151-163.

Claasen, T A C M and Mecklenbrauker, W F G (1980). The Wigner distribution - a tool for time-frequency signal analysis. Philips J Res 35, 217-250; 276-300; 372-389.

Coker, C H (1968). Speech synthesis with a parametric articulatory model. Speech Symposium, Kyoto, Paper A-4. BK

Coker, C H (1976). A model of articulatory dynamics and control. Proc IEEE, 64, 4, 452-460.

Cooper, F S (1950). Spectrum analysis. J Acoust Soc Amer 22, 761-762.

Costello, J (1981). Time domain speech synthesis. WESCON 1981 Conference Record, San Francisco.

Crichton, RG and Fallside, F (1974). Linear prediction model of speech production with applications to deaf speech training. Proc IEE, V121, N8.

De Russo, PM, Roy, RJ and Close, CM (1965). State Variables for Engineers : Wiley, New York.

Dudley, H (1939). Remaking speech. J Acoust Soc Amer, 11, pp 169-177.

Dudley, H, Riesz, R R and Watkins, S S A (1939). A synthetic speaker. J Franklin Inst, 227, 739-764. BK

Dunn, H K (1950). The calculation of vowel resonances, and an electrical vocal tract. J Acoust Soc Amer 22, 740-753. BK

Dunn, H K (1961). Methods of measuring vowel formant bandwidths. J Acoust Soc Amer 33, 1737-1746.

Dupree, B (1984). Formant coding of speech using dynamic programming. Electronics Letters 20, p279.

Egan, J P (1948). Articulation testing methods. Laryngoscope 58, 955-991.

Estes S E, Kerby, H R, Maxey, H D and Walker, R M (1964). Speech synthesis from stored data. IBM Jr Research, 8, 2-12.

Fairbanks, G (1958). Test of phonemic differentiation: the rhyme test. J Acoust Soc Amer 30, 596-600.

Fallside, F and Young, S J (1978). Speech output from a computer-controlled water-supple network. Proc IEEE 125(2), 157-161.

Fant, C G M. (1950) MIT Acoustics Laboratory Quarterly Progress Report, July-Sept, p20.

Fant, G. (1970; Second edition) Acoustic Theory of Speech Production. Mouton: the Hague.

Fant, G (1975). A note on vocal tract size factors and nonuniform F-pattern scalings. In G Fant, Speech Sounds and Features. MIT Press: Cambridge, MA.

Fant, G (1978). Formant damping and excitation. Speech Trans Lab - Quart Prog and Status Report 2-3, Stockhnlm.

Fant, G and Martony, J (1962a). Speech Synthesis. Speech Trans Lab - Quart Prog and Status Report (July), Stockholm.

Fant, G and Martony, J (1962b). OVE II synthesis strategy. Paper F5, Sp Comm Seminar, Stockholm.

Faulkner, A (1987) A test for the intelligibility of synthetic speech. IBM UKSC Report, IBM (UK) Science Centre, Winchester.

Flanagan, J L (1957). Note on the design of 'terminal analog' speech synthesizers. J Acoust Soc Amer 29, 306-10. BK

Flanagan, J L (1965). Recent studies in speech research at Bell Telephone Labs (II). Proc Fifth Intern Congr Acoust, Liege, Belgium.

Flanagan, J L (1972). Speech Analysis Synthesis and Perception: Berlin; Springer Verlag. (Second edition)

Flanagan, J L (1972). Voices of Men and Machines. J Acoust Soc Amer, 51, 1275-1397. BK

Flanagan, J L, Coker, C H and Bird, C M (1962). Digital computer simulation of a formant-vocoder speech synthesizer. 15th Ann Meeting Audio Engr Soc, Preprint 307. BK

Flanagan, J L, Ishizaki, K and Shipley, K L (1975). Synthesis of speech from a dynamic model of the vocal cords and vocal tract. Bell System Tech J, 54, 3, 485-566.

Flanagan, J L, Ishizaki, K and Shipley, K L (1980). Signal models for low bit-rate coding of speech. J Acoust Soc Amer 68, 780-791.

Flanagan, J L and Landgraf, L (1968). Self-oscillating source for vocal-tract synthesizers. IEEE Trans Audio Electroacoust, AU-16, 57-64. BK

Flanagan, J L and L R Rabiner (eds, 1973). Benchmark Papers in Acoustics: Speech Synthesis: Dowden, Hutchinson & Ross, Stroudsburg, PA.

Fletcher, H (1929) Speech and Hearing. Van Norstrand: NY.

Fletcher, H and J C Steinberg (1929). Articulation testing methods, Bell System Tech J 8, 806-854.

Foster, J R and Haggard, M P (1979). FAAF - An efficient analytical test of speech perception. Proc Inst of Acoust, paper 1A3, 9-12.

Fujimura, O. Analysis of nasal consonants (1962). J Acoust Soc Amer, 34, 1865-1875.

Fujisaki, H (1977). Functional models of articulatory and phonatory dynamics. GALF

GALF Symposium (1977). Carre, R, Descout, R and Wajskop, M (Eds). Articulatory Modeling and Phonetics (Proceedings of Symposium). Grenoble: GALF Groupe de la Communication Parlee.

Gill, J S (1961) perception of aperiodicity of voice fundamental frequency; (cited in Holmes, 1972).

Gold, B and Rabiner, L R (1968). Analysis of digital and analog formant synthesizers. IEEE Trans Audio Electroacoust 16, 81-95.

Green, N (1976). Analysis synthesis using a pole-zero approximation to speech spectra. Proc IEEE-ICASSP, 306-309.

Haggard, M P (1970). Articulatory synthesis by rule II. Speech Syn and Percep. 3, 1-21. Psych Lab Univ Cambridge.

Haggard, M P (1979). Experience and perspectives in articulatory synthesis. FF

Harris, K S (1971). Vowel stress and articulatory reorganisation. Haskins Labs Status Report SR-28, 167-178.

Harshman, R (1970). Foundations of the PARAFAC procedure. UCLA Working Papers in Phonetics 16; University microfilms No. 10,085.

Harshman, R, Ladefoged, P and Goldstein, L (1977). Factor analysis of tongue shapes. J Acoust Soc Amer 62(3), 693-707.

Heinz, J M and K N Stevens (1961). On the properties of voiceless fricative consonants. J Acoust Soc Amer, 33, 589-596.

Henke, W (1967). Preliminaries to speech synthesis based upon an articulatory model. Proc Conf on Speech Commun and Proc, AFCRL and IEEE Audio Group, Cambridge, Mass.

Hirsh, I J et al (1952). Development of materials for speech audiometry. J Speech Hearing Disorders 17, 321-337.

Holmes, J N, Mattingly, I G and Shearme, J N (1964). Speech synthesis by rule. Language and Speech 7, 127-143. BK

Holmes, J N (1972). Speech Synthesis. London: Mills & Boon.

Holmes, J N (1973). The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. IEEE Trans Audio Electroacoust, AU-21, 298-305.

Holmes, J N (1979). Synthesis of natural sounding speech using a formant synthesizer. FF, 275-285.

Holmes, J N (1980). Avoiding unwanted low-frequency level variations on the output of a parallel-formant synthesizer. J Acoust Soc Amer 68, p S18.

Holmes, J N (1982). Formant synthesizers: cascade of parallel? JSRU Research Report 1017.

Holmes J N, Mattingly, I G and Shearme, J N (1964). Speech synthesis by rule. Lang and Sp, 7(3), 127-43.

House, A S, Williams, C E, Hecker, H L and Kryter, K D (1965). Articulation testing methods: consonantal differentiation with a closed response set. J Acoust Soc Amer 37, 158-66.

Ishizaka, K and J L Flanagan (1972). Synthesis of voiced speech from a two-mass model of the vocal cords. BSTJ 51, 1233-1268.

Itakura, F and Saito, S (1968). Analysis synthesis telephony based on the maximum liklihood method. Proc Sixth Intern Congr Acoust, Paper C-5-5, C17-20. BK

Itakura, F and Saito, S (1970). A statistical method for estimation of speech spectral density and formant frequencies, Elect and Commun, Japan, 53-A, 1, 36-43.

Jackson, PH (1984). IAX - A high level language for image processing. IBM (UK) Science Centre UKSC126, Winchester.

Jeffrey, A (1979). Mathematics for Engineers and Scientists. 2nd Ed. Van Norstrand Reinhold, UK.

Joos, M. Acoustic Phonetics (1948). Lang Monographs No 23, Ling Soc Amer 1948

Jospa, P (1977). Consequences Acoustiques des deformations dynamiques des conduit vocal. GALF

Kaiser, J F (1983). Some observations of vocal tract operation from a fluid flow point of view. In Titze, I and Scherer, R (Eds) Proceedings of Conference on Physiology and Biophysics of Voice, Univ of Iowa.

Kaiser, J F (1959). Digital Filters. In System Analysis by Digital Computer, Kuo and Kaiser, Eds, pp 218-285: New York, McGraw Hill.

Kelly, J L Jr and Lochbaum, C (1962). Speech Synthesis. Proc Stockholm Speech Comm Seminar, RIT, Stockholm. BK

Kelso, J A S, Vatikiotis-Bateson, E, Saltzman, E L and Kay, B (1985). A qualitative dynamic analysis of reiteritant speech production: phase portraits, kinematics and dynamic modelling. J Acoust Soc Amer, 77, 266-280

Kinsler, L E, Frey, A R, Coppens, A B, and Sanders, J V (1982). Fundamentals of Acoustics 3rd ed, John Wiley & Sons, New York.

Klatt, D H (1972). Acoustic theory of TASS. Proc ICSCP, IEEE T-AE, 131-135.

Klatt, D H (1976). Structure of a phonological rule component for a synthesis by rule program. IEEE T-ASSP, 24, 391-398.

Klatt, D H (1980a). The PHONET Component. Chapter 11 in Allen (1980).

Klatt D H (1980b). Software for a cascade/parallel formant synthesizer. J Acoust Soc Amer 67(3), 971-995.

Klatt, D H (1982). The Klattalk text-to-speech system. IEEE ICASSP, Paris, 1589-1592.

Kruskal, J B (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1-27.

Ladefoged, P (1982). A Course in Phonetics. 2nd Ed. Harcourt Brace Jovanovich, New York.

Ladefoged, P, Harshman, R, Goldstein, L and Rice, L (1978). Generating vocal tract shapes from formant frequencies. J Acoust Soc Amer 64(4), 1027-1035.

Ladefoged, P and Harshman, R (1979). Formant frequencies and movement of the tongue. FF

Laver, J (1982). The Phonetic Description of Voice Quality. Cambridge Univ Press.

Lawrence, W (1953). The synthesis of speech signals which have a low information rate. Communication Theory. Butterworth & Co, Ltd, 460-469. BK

Lawrence, W (1974). The phoneme, the syllable and the parameter track. Proc Speech Comm Seminar, Stockholm, Aug 1-3.

Lehiste, I (Ed) (1967). Readings in Acoustic Phonetics. MIT Press, Cambridge, Mass.

Lehiste, I and Peterson, G E (1959). Linguistic considerations in the study of speech intelligibility. J Acoust Soc Amer 31, 280-286.

Liljenkrants, J (1971). A Fourier series description of the tongue profile. Speech Trans Lab - Quart Prog and Status Report 4, 9-18, Stockholm.

Liljenkrants, J (1985). Dynamic line analogs for speech synthesis. Speech Trans Lab - Quart Prog and Status Report 1, 1-14, Stockholm.

Lindblom, B (1963). Spectrographic study of vowel reduction. J Acoust Soc Amer 35, 11, 1773-1781.

Lindblom, B and Sundberg, J (1971). Acoustical consequences of lip, tongue, jaw and larynx movement. J Acoust Soc Amer 50, 1166-1179.

Linggard, R (1985) Electronic Synthesis of Speech. Cambridge Univ Press.

Maeda, S (1977). On a simulation method of a dynamically varying vocal tract: reconsideration of the Kelly-Lochbaum model. GALF

Makhoul, J (1975). Linear prediction: a tutorial review. Proc IEEE, 63, 561-580.

Markel, J D and A H Gray (1976). Linear Prediction of Speech: Springer Verlag, New York.

Mermelstein, P (1967). Determination of vocal tract shape from measured formant frequencies. J Acoust Soc Amer 41, 1283-1294.

Mermelstein, P (1971). Calculation of the vocal-tract transfer function for speech synthesis applications. Proc 7th Intern Congr Acoust, Paper 23 C 13, 173-176. BK

Mermelstein, P (1973). Articulatory model for the study of speech production. J Acoust Soc Amer 53, 1070-1082.

Moller, W, Strube, H and Kretschmar, J (1977). Measurement and acoustic estimation of articulatory parameters. GALF

Ohman, S E G (1967). Numerical model of coarticulation. J Acoust Soc Amer 41, 2, 310-320.

Perkell, J (1969). Physiology of Speech Production. MIT Press: Cambridge, MA.

Perkell, J (1977). Articulatory modeling and phonetic description (summary remarks). GALF, pp 105-113.

Petersen, G E and Barney, H L (1952). Control methods used in a study of the vowels. J Acoust Soc Amer, 24, 175-184.

Peterson, G E, Wang, S Y and Silvertsen, E (1958). Segmentation techniques in speech synthesis. J Acoust Soc Amer 30, 739-742.

Pickering, B (1986). Cosegmentation in the IBM text-to-speech system. Proc Inst of Acoust, V8, Pt7, 385-92.

Pinson, E N (1963). Pitch-synchronous time-domain estimation of formant frequencies and bandwidths. J Acoust Soc Amer, 35, 1264-1273.

Pisoni, D (1979). Evaluation of intelligibility. In Allen (1980).

Potter, R K, A G Kopp and H C Green (1947). Visible Speech. Van Norstrand: NY.

Pratt, R L (1986). On the intelligibility of synthetic speech. Proc Inst of Acoust, Vol 8 Part 7, 183-192.

Quarmby, DJ and Holmes, JN (1984). Implementation of a parallel-formant speech synthesiser using a single-chip programmable signal processor. IEE Proc. V131F No6 p563-69.

Rabiner, L R (1968a). Digital-formant synthesizer for speech-synthesis studies. J Acoust Soc Amer, 43, 822-28. BK

Rabiner, L R (1968b). Speech synthesis by rule: an acoustic domain approach. Bell System Tech J 47, 17-37.

Rabiner, L R (1969). A model for synthesizing speech by rule. IEEE Trans Audio Electroacoust 17, 7-13.

Rabiner, L R and R W Schafer (1978). Digital Processing of Speech Signals: Prentice-Hall, New Jersey.

Rosen, G (1958). Dynamic analog speech synthesizer. J Acoust Soc Amer 30, 201-209.

Rosenburg, A E (1971). Effect of glottal pulse shape on the quality of natural vowels. J Acoust Soc Amer 49,2,2, 583-590.

Rothauser, E H, Urbanek, G E and Pachl, W P (1971). A comparison of Preference Measurement Methods. J Acoust Soc Amer 49(4), 1279-1308.

Ruiz, P M (1971). A digital simulation of a time-varying vocal tract. J Acoust Soc Amer 49, 123 (A).

Rye, JM and Holmes, JN (1982). A versatile software parallel-formant speech synthesiser. JSRU Research Report No 1016.

Sambur, M R (1975). Efficient LPC vocoder. J Acoust Soc Amer 57, S34(A).

Schott, L O (1948). A playback for visible speech. Bell Lab Record 26, 333-339.

Schouten, M E H and Pols, L C W (1977, 1978, 1980). Spectral study of coarticulation. Jr of Phonetics; Part I: 7, 1-23; Part II: 7, 205-224; Part III: 9, 225-231.

Scully, C (1979). Model prediction and real speech: fricative dynamics. FF

Seeviour, P M, Holmes, J N and Judd, M W (1976). Automatic generation of control signals for a parallel formant speech synthesizer. Proc IEEE-ICASSP.

Shadle, C (1986). Models of fricative consonants involving sound generation along the wall of a tube. Proc Intern Congr Acoust: V1, A4; Toronto.

Shadle, C H and Atal, B S (unpublished). On the use of pseudo-area parameters for speech synthesis by rule. Available from the first author, EE Dept, Univ of Southampton, UK.

Sharman, R A (1986). Designing an experimental text to speech synthesizer. Proc Inst of Acoust, V8, Pt7, 355-362.

Shepard, R N (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrica 27:125-140; 219-246.

Sondhi, M M (1977). Estimation of vocal-tract areas: the need for acoustical measurements. GALF

Sondhi, M M and Resnick, J R (1983). The inverse problem for the vocal tract: numerical methods, acoustical experiments, and speech synthesis. J Acoust Soc Amer 73, 985-1002.

Sondhi, M M and Schroeter, J (1987). A hybrid time-frequency domain articulatory speech synthesizer. IEEE T-ASSP, 35, 7, 955-967.

Song, K E (1983). A pole-zero model of speech. IEEE T-ASSP, 1556-65.

Sonoda, Y (1977). Estimation of dynamic characteristics of articulatory movements. GALF

Speech Research Group (1986). Speech Processing at the UKSC. IBM (UK) Scientific Centre, Winchester.

Steel, R G D and Torrie, J H (1981). Principles and Procedures of Statistics. McGraw-Hill.

Steiglitz, K (1977). Simultaneous estimation of poles and zeroes. IEEE T-ASSP, 229-34.

Stella, M (1985). Speech Synthesis. In Fallside, F and Woods, W A (eds), Computer Speech Processing: London: Prentice-Hall.

Stevens, K N, Bastide, R P and Smith, C P (1955). Electrical synthesizer of continuous speech. J Acoust Soc Amer 27, 207(A).

Stevens, K N, Bastide, R P and Smith, C P (1960). Electrical synthesizer of continuous speech. J Acoust Soc Amer 32, 47-55.

Stevens, K N and House, A S (1955). Development of a quantitative description of vowel articulation. J Acoust Soc Amer 27, 484-493. BK

Stevens, K N, House, A S and Paul, A P (1966). Acoustical description of syllabic nuclei: an interpretation in terms of a dynamic model of articulation. J Acoust Soc Amer 40, 1, 123-132.

Stevens, K N, Kasowski, S and Fant, G (1953). An electrical analog of the vocal tract. J Acoust Soc Amer 25, 734-742. BK

Stevens, S S (1947). Handbook of Experimental Psychology. John Wiley and Sons, New York.

Stewart, J Q (1922). An electrical analogue of the vocal organs. Nature, 110, 311-312. BK

Strube, H W (1982). Time-varying wave digital filters for modelling analog systems. IEEE T-ASSP, 30, 6, 864-8.

Sundberg, J and Gauffin, J (1979). Waveform and spectrum of the glottal source. FF

Teager, H M and Teager, S M (1983). The effects of separated air flow on vocalization. In Bless, D M and Abbs, J H (Eds), Vocal Fold Physiology, College Hill Press: San Diego CA.

Terepin, S (1979). A vocal tract model for speech synthesis. PhD dissertation, Cambridge Univ Engineering Dept.

Tillman, T W, Carhart, R and Wilber, L (1963). A test of speech discrimination composed of CNC monosyllabic words. (N.U. Auditory Test No 4). US Air Force School Aerospace Med, Brooks AFB, SAM-TDR-62-135. AD No 403275.

Titze, I R (1973, 1974), The human vocal cords: a mathematical model, Phonetica, Part 1, V28, 129-170; Part 2, V29, 1-21.

Torgerson, WS (1958). Theory and Method of Scaling. John Wiley & Sons, NY

Umeda, N (1975). Vowel duration in American English. J Acoust Soc Amer 58(2), 434-445.

Viswanathan, R and J Makhoul (1975). Quantisation properties of transmission parameters in linear predictive systems. IEEE T-ASSP, 23, 309-321.

Voiers, W D (1967). Performance evaluation of speech processing devices, III. Diagnostic evaluation of speech intelligibility. Final Report, Contract AF19(628)4987, Air Force Cambridge Research Labs.

Voiers, W D (1977). Diagnostic evaluation of speech intelligibility. In M Hawley (Ed), Benchmark Papers in Acoustics, Vol 11: Speech Intelligibility and Speaker Recognition. Stroudsburg, PA: Dowden, Hutchinson & Ross.

Wakita, H (1973). Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. IEEE Trans Audio Electroacoust, AU-21, pp 417-427.

Weibel, ES (1955) Vowel synthesis by resonant circuits. J Acoust Soc Amer 27, 858-65.

Wiggins, F and Brantingham, L (1978). Three-chip system synthesizes human speech. Electronics, August 31, 109-116.

Witten, I H (1982). Principles of Computer Speech. London, Academic Press.

Witten, I H and Madams, P H C (1978). The Chatterbox - a speech toy. Wireless World 84, 36-41 & 85, 77-80.

Wright, R D (1972). Speech synthesis from stored waveforms. NSA Computer Conf, Ft Meade MD.

Wright, R D (1973). Orthogonal linear prediction. Proc IEEE Conf on Linear Prediction, Carnegie-Mellon Univ.

Wright, R D (1976). A system for implementing rule synthesis of speech. Proc Inst of Acoust Autumn Conf, 2/20/1-2/20/4, Edinburgh.

Wright, R D (1987). Basic properties of speech. In Martin, M C (Ed) Speech Audiometry, Taylor and Francis, London.

Wright, R D and Elliott, S J (1985). Parameter motion in speech synthesis. Inst of Acoust Speech Group, Winchester, December.

Wright, R D and Elliott, S J (1986). Control of speech synthesis parameters. 12th Int Cong Acoust, A4-3.

Wright, R D and Elliott, S J (in preparation). Parameter motion in speech synthesis. Accepted (Jan 88) pending revision, Journ Acoust Soc Amer.

Young, S and Fallside, F (1979). Speech synthesis from concept: a method for speech output from information systems. Journ Acoust Soc Amer 66(3), 685-695.

Appendix 1: Pascal subroutines for parameter conversion

1	series to direct form:	zedmake
2	direct form to reflections:	reflection
3	reflections to areas:	areas
4	resonance to artic. params:	articmake
5	areas to artic. params:	areatoartic
6	series to parallel form:	parallel
7	artic. params to areas:	artictoarea
8	areas to reflections:	areatorefl
9	reflections to direct form:	prediction
10	direct form to series:	(polynomial rooting)

(complete code given for 1-9; the conversion in 10 uses a lengthy programme, available from the author; standard polynomial-rooting routines are available in mainframe scientific subroutine packages)

```
CONST P=10;           {10th order}
      articareas=8;   {8 computed areas in artictoarea}

TYPE CVEC = ARRAY[0..P] OF REAL;    {general form of params}
     avec = array[0..3] of real;    {artic params}
     complex=record
          r: real;
          i: real;
     adat = array[1..3, 0..8] of real;  {for artic-to-area}

procedure eval2(var R: complex; POLY: cvec; A: complex;
 dgr: integer); extern; {evaluate polynomial POLY of degree
 DGR with complex argument A; the result is R}
function cmod(A: complex): real; extern; {modulus of A}
```


1 series to direct form: zedmake

```
procedure zedmake(factor: cvec; var direct: cvec; order:
                  integer; fr: real);
{takes entire set of cf,bw pairs and makes direct form coeffs}
var   z:                  cvec;
      i,pairs:            integer;
      cf,bw,pif,tpif,rho,dummy: real;
const pi=3.14159 ;
begin
    z[0] := 1.0; direct[0] := 1.0;      {initial values}
    pif  := -pi/fr;   tpif := pif+pif;  {π/freq; 2*π/freq}
    pairs:= order div 2;
    for i:= 1 to pairs do
    begin
        cf   :=factor[i*2-2];
        bw:=factor[i*2-1];    {steps by two for pairs}
        dummy:=pif*bw;
        rho  := exp(dummy);
        z[2] := rho*rho;
        polymult(z,direct,2,i*2-2);     {result in direct}
    end;
end {zedmake};

PROCEDURE POLYMULT (A: CVEC; VAR S: CVEC; NA, NS: INTEGER);
{MULTIPLIES A*S WITH RESULT IN S}

VAR I,J: 0..P;
   TEMP: CVEC;

BEGIN
{DEBUG writeln(' In polymult'); }
   FOR I := 0 TO NA+NS DO TEMP[I]:=0.0;
   FOR I := 0 TO NS DO
      FOR J:=0 TO NA DO TEMP[I+J]:= TEMP[I+J] + A[J]*S[I];
   FOR I := 0 TO NA+NS DO S[I] := TEMP[I];
{DEBUG writeln(' Leaving polymult'); }
END {POLYMULT};
```

2 direct form to reflections: reflection

```
PROCEDURE REFLECTION (COEFF: CVEC; DGR: INTEGER;
VAR REFL: CVEC); {COMPUTES REFL FROM LPC IN COEFF}

VAR TEMP:     CVEC;
    I, M:     0..P;
    k,DENOM:  REAL;

BEGIN
   FOR M:=DGR DOWNTO 1 DO
   begin
      k := COEFF[M];
      REFL[M-1] := k; {refl goes 0..p-1, coeff goes 0..p)
{ref Makhoul, '75, p130 and Witten, '82, p137.  Note that I
don't exactly follow either, because of my convention that
all parameter sets begin with subscript zero.  Thus
refl[p-1]:=prediction_coeff[p], etc }

      DENOM := 1.0 - k*k;

      FOR I:=1 TO M-1 DO TEMP[I] := COEFF[I];
      if abs(DENOM) <0.000001 then writeln('Denom of zero!')
      else FOR I:=1 TO M-1 DO
           COEFF[I]:=(COEFF[I] - k*TEMP[M-I]) / DENOM;
   END
END {REFLECTION};
```

3 reflections to areas: areas

```
procedure areas (refl: cvec; var area: cvec; order: integer);
{takes in reflections and computes areas; A0=1}
var i: integer;
begin
   area[0] := 1.0;
   for i:=1 to order do
      area[i] := area[i-1]*(1.0+refl[i-1])/(1.0-refl[i-1])
{Using reflection convention of Makhoul, Rabiner & Schafer}
end;
```


4 resonance to articulatory params: articmake

```
procedure articmake(cfbw,zed: cvec; var artic: avec;
order: integer); {uses Ladefoged (1978) equations to
compute tongue and lip parameters from formant data}

{uses F1, F2, F3 (from cfbw) for tongue control params
 w1, w2 = artic 0,1}
{also uses F1 to F3 for lip opening = artic 2}
{uses last direct form coeff (from zed) for loss = artic 3}

const   c1= 2.309; c2= 2.105; c3= 0.117; c4=-2.446;
        c5=-1.913; c6=-0.245; c7= 0.188; c8= 0.584;
        l1=3.0E-3; l2=-3.43E-7; l3=4.143; l4=-2.865;

var f1,f2,f3:   real;

begin

{set loss term = last polynom coeff = last refl coeff}
artic[3]:=zed[order];

{pick up formants from cfbw}
f1:=cfbw[0];    f2:=cfbw[2];    f3:=cfbw[4];

{set lip opening}
artic[2]:=l1*f2 + l2*f2*f3 + l3*(f1/f2) + l4;

{set w1 = front raising = artic 0}
artic[0]:=c1*(f2/f3) + c2*(f1/f3) + c3*(f3/f1) + c4;

{set w2 = back raising = artic 1}
artic[1]:=c5*(f1/f2) + c6*(f2/f1) + c7*(f3/f1) + c8;

end ;
```

```
5       areas to articulatory params:          areatoartic

procedure areatoartic(area: cvec;var artic: avec; order:
integer; t: adat);
{projects areas 1-8 to get tongue control params w1,w2=FR,BR=
artic 0,1; uses area 9 for lip opening = artic 2; uses areas
10 and 9 to compute reflection coeff for loss = artic 3}

const minarea = 0.01;
      pi = 3.14159;

var   denom,dotprod,detT:  real;
      i,j,k:               integer;
      ttT:    array[1..2,1..2] of real; {t times t-transpose}
      diam,fourdpi:        real;
      articp:              avec; {unscaled projection}

begin
   fourdpi:=4.0/pi; {factor to convert areas to diameters}

   {compute ttT, the 2x2 array which gets implicitly inverted}
   for i:=1 to 2 do
      for j:=1 to 2 do
         begin
            dotprod:=0.0;
            for k:=1 to articareas do
               dotprod:=dotprod + t[i,k]*t[j,k];
            ttT[i,j]:=dotprod;
            end;

{also compute the determinant of ttT for use in inversion}
   detT:=ttT[1,1]*ttT[2,2] - ttT[1,2]*ttT[2,1];

{set loss term = last polynom coeff = last refl coeff}
   denom:=area[order] + area[order-1];
   if denom < minarea then denom:=minarea;
   artic[3]:= (area[order] - area[order-1]) / denom;

{set lip opening}
   artic[2]:=sqrt(area[order-1]*1.188*fourdpi);
{!!!  Ladef areas start at 1.188 at larynx, not 1.00; the
larynx diam is 1.23, and pi*1.23*1.23/4 => 1.188}

   for j:=0 to 1 do
      begin
         {set w1 = front raising = artic 0;
          set w2 = back raising = artic 1;}
         articp[j]:=0.0; {initialise unscaled projection}
```



```
         for i:=1 to articareas do
            begin
               diam:=sqrt(area[i]*1.188*fourdpi);
               {compute diameters and project}
               {need to subtract t[3,i], the neutral diam}
               articp[j]:=articp[j] + (diam-t[3,i])*t[j+1,i];
               end;
         end;
   {now use equations for 2x2 matrix solution}
   artic[0]:=(articp[0]*ttT[2,2] - articp[1]*ttT[2,1]) /detT;
   artic[1]:=(articp[1]*ttT[1,1] - articp[0]*ttT[1,2]) /detT;
   end;  {of procedure}
```

```
6       series to parallel form:       parallel

procedure parallel(var cfbw: cvec; dgr: integer; SR: real);
     {given centre frequency and bandwidth data in CFBW,
     replace bandwidth with amplitude at centre frequency}
const tpi=6.2831853 ;          {2*π}

var denom:              cvec;
    i,pairs:            integer;
    db110,db0,theta:    real;
    z,result:           complex;

begin
    {first make polynom in zed = denom of all-pole case}
    zedmake(cfbw,denom,dgr,SR);
    pairs:=dgr div 2;       {number of pole-pairs}
    db110:=20.0/ln(10.0);  {constant; conversion to decibels}
 {Find response at DC for dB reference}
    z.r:=1.0; z.i:=0.0;
    eval2(result,denom,z,dgr);    {response (complex) at DC}
    db0:=db110*ln(cmod(result));  {modulus, convert to dB}
 {Now do all the resonances}
    for i:=1 to pairs do
    begin
        theta:=tpi*cfbw[i*2-2]/SR;
        z.r:=cos(theta);
        z.i:=sin(theta);
        eval2(result,denom,z,dgr);
                  {get complex response at resonance}
        cfbw[i*2-1]:=db0 - db110*ln(cmod(result));
                  {get dB re resp @DC}
{evaluating polynom rather than 1/polynom; thus need to take
 ratio of ref/resp rather than resp/ref to correct.
 Thus 20log(ref/resp) comes from log ref - log resp}
{and we save the result in the BW position in cfbw}
    end;    {of FOR}
end;    {of PROC}
```

7 artic. params to areas: artictoarea

```
procedure artictoarea(artic: avec; var area: cvec;
                      order: integer; t: adat);
{implement Ladefoged, et al tongue factors}
{use artic 0,1 and eqns 1,2 for diams 0..P-2 & make areas;
 area P-1 from lip opening diam = artic2;
 area P set by loss = final area ratio = artic3}

const minarea=0.0001;   {Minimum area}
      pi = 3.14159;

var i:   0..P;
    pd4: real;

begin
pd4:=pi/4.0;    {factor to convert diameters to areas}
area[order-1]:=artic[2]*artic[2]*pd4;
    {lip opening; straight from input}

    {artic3 is loss = final refl coeff}
    {compute final area from refl}
if artic[3]>0.99 then artic[3]:=0.99; {prevent overflow}
area[order]:=area[order-1] * (1.0+artic[3]) / (1.0-artic[3]);

{compute order-2 areas = SAMPLED areas in Ladef. formula}
{t[3] is the 'n' in Ladef, p 1028, = neutral tongue position}
for i:=0 to order-2 do
    begin
    area[i]:=t[1,i]*artic[0] + t[2,i]*artic[1] + t[3,i];
    area[i]:=area[i]*area[i]*pd4;   {convert DIAM to AREA!!!}
    end;
{use indices 1 to 8 to refer to Ladef's 2 to 16 step 2}
{index 0 has t1=t2=0, and n is glottal section = 1.23 cm}

for i:=0 to order do if area[i]<minarea then area[i]:=minarea;
{formula does 8 areas; area 0 a constant;loss is preserved}

end;
```

```
{t[1,*]  t[2,*] n=t[3,*]     A file of coefficients for
 0.0000  0.0000  1.23    1   converting tongue parameters
 0.1300  0.6000  1.67    2   into areas.  Based on Ladefoged,
 0.5949  1.1900  1.72    4   et al, JASA 64-4,pp 1027-1035.
 1.0500  0.9461  1.31    6   Table I, p 1028.
 1.1370  0.0056  1.46    8   Using alternate sections (even
 0.5536 -0.9281  1.45   10   numbered) for 10th order system,
-0.8719 -1.4180  1.43   12   instead of 18 areas of
-1.4060 -1.2420  1.89   14   Ladefoged et al.
-0.9063 -0.4905  1.75   16 )
```


8 areas to reflections: areatorefl

```
procedure areatorefl(area:cvec; var refl:cvec; order:integer);
const minareas=0.0001;
{minimum areas; can't convert zero areas to refl's}

var i:  0..p;  num,denom:  real;

begin   {proc}
for i:=0 to order-1 do
    begin   {loop}
    denom:= (area[i+1]+area[i]);
    num:= (area[i+1]-area[i]);
    if denom<minareas then refl[i]:=0.0
       else refl[i] := num / denom;
    {if both areas are zero, so is the refl;
       else use formula}
    end;    {loop}
end;    {proc}
```

9 reflections to direct form: prediction

```
PROCEDURE prediction(var COEFF: CVEC; DGR: INTEGER;
                     REFL: CVEC);
{COMPUTES prediction coeff's FROM refl's}

{NOT using Witten format for prediction coeffs;
using polynomial coeffs in form
coeff[0]*x^dgr + coeff[1]*x^dgr-1 + ... + coeff[dgr]*x^0 )
{This is the Makhoul and Rabiner&Schafer refl coeff standard}

VAR TEMP:  CVEC;
     i,j:  integer;

begin
coeff[0]:=1.0; {standard form}
for i:=1 to dgr do
    begin
    coeff[i]:=refl[i-1];  {Makhuol refl's have same sign}
    for j:=1 to i-1 do temp[j]:=coeff[j]+refl[i-1]*coeff[i-j];
    {need temporary array to prevent updated coeff from
     confusing the iteration}
    for j:=1 to i-1 do coeff[j]:=temp[j];
    end;
{Makhoul algorithm, LPC tutorial}
end;    {of prediction}
```

Appendix Two: Formant trajectories

This appendix contains data supporting the conclusions of Chapter Six, the objective study of the consequences (in terms of formant motion) of linear interpolation of six types of synthesizer parameter.

Part One of the appendix (Section A2.1) gives all the formant trajectories for all eight nonsense words. The first eight figures (one per nonsense word) contain three graphs per figure, and present data on direct form, reflection coefficients and area function interpolation. Each graph also includes the series linear (straight line) results for comparison. The graphs are each in two parts: resonance centre frequency, and resonance bandwidth. Only data on the lowest three formants are presented, except for the nonsense word /ima/. The nasal /m/ has a low frequency resonance in the first formant region. To prevent problems of formant path continuity, the vowels for this nonsense word also include a nasal resonance

The remaining two figures of Part One (Figures A2.9 and A2.10) give results for the linear interpolation of articulatory parameters. Articulatory target values were determined using the Ladefoged et al (1978) method of estimating tongue and lip parameters from formant frequency data. Only the data for formant frequency is presented, allowing four graphs per figure. Thus all eight nonsense words are presented in these final two figures of Part One.

No plots are given for parallel resonance data. There is no difference between the formant centre frequency interpolation paths of a serial vs a parallel resonance configuration. Further, the notional parallel resonance synthesiser considered in this study has fixed bandwidths, making a plot of bandwidths vs time rather uninteresting.

Part Two gives detailed quantitative results of path differences between parameters for each nonsense word.

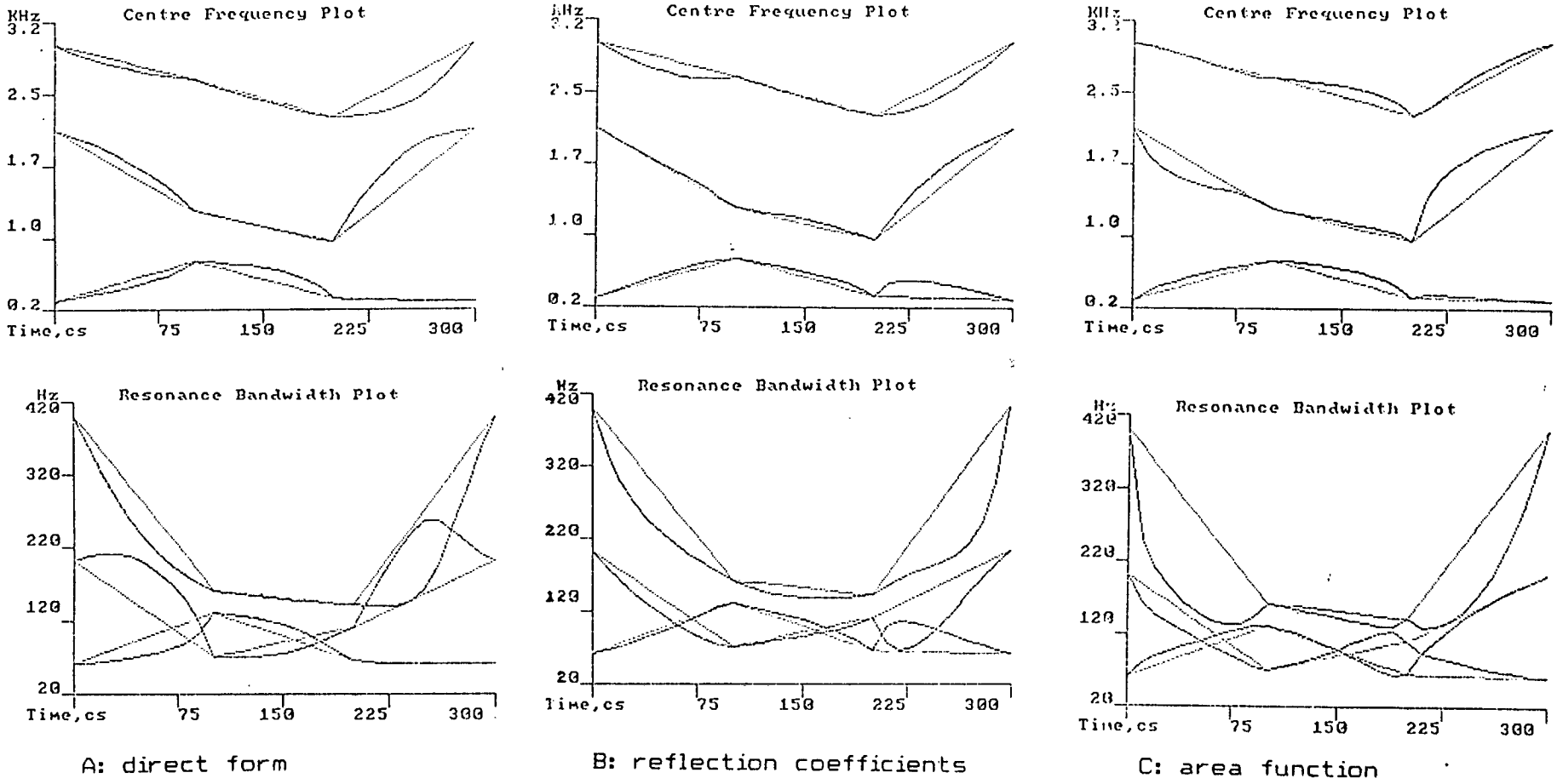


Figure A2.1: Parameters for /iaui/. Formant motion for linear interpolation of direct form, reflection coefficient, and area function synthesizer parameters. The straight lines represent comparison paths for series resonance synthesis.

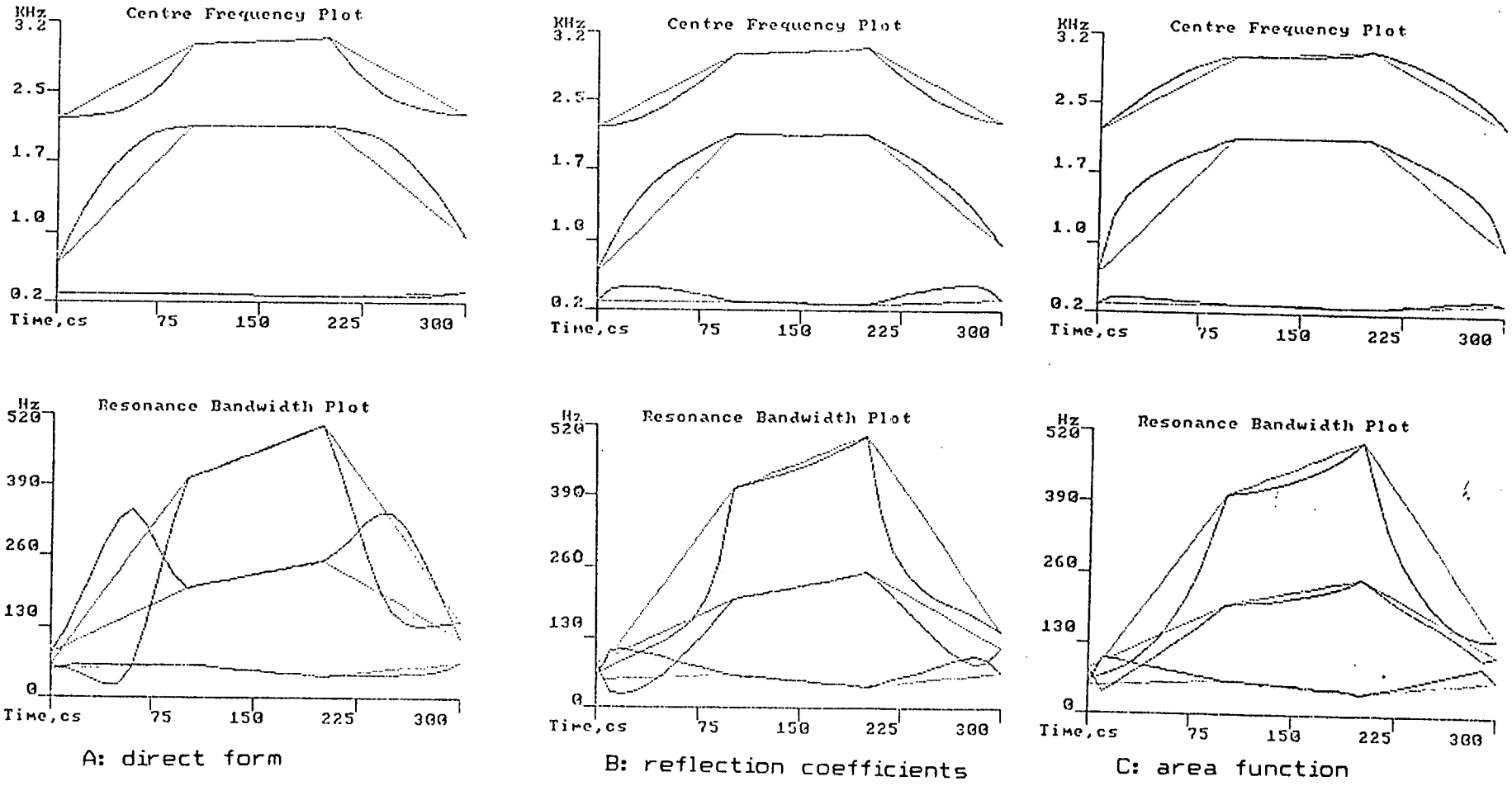


Figure A2.2: Parameters for /wiju/. Formant motion for linear interpolation of direct form, reflection coefficient, and area function synthesizer parameters. The straight lines represent comparison paths for series resonance synthesis.

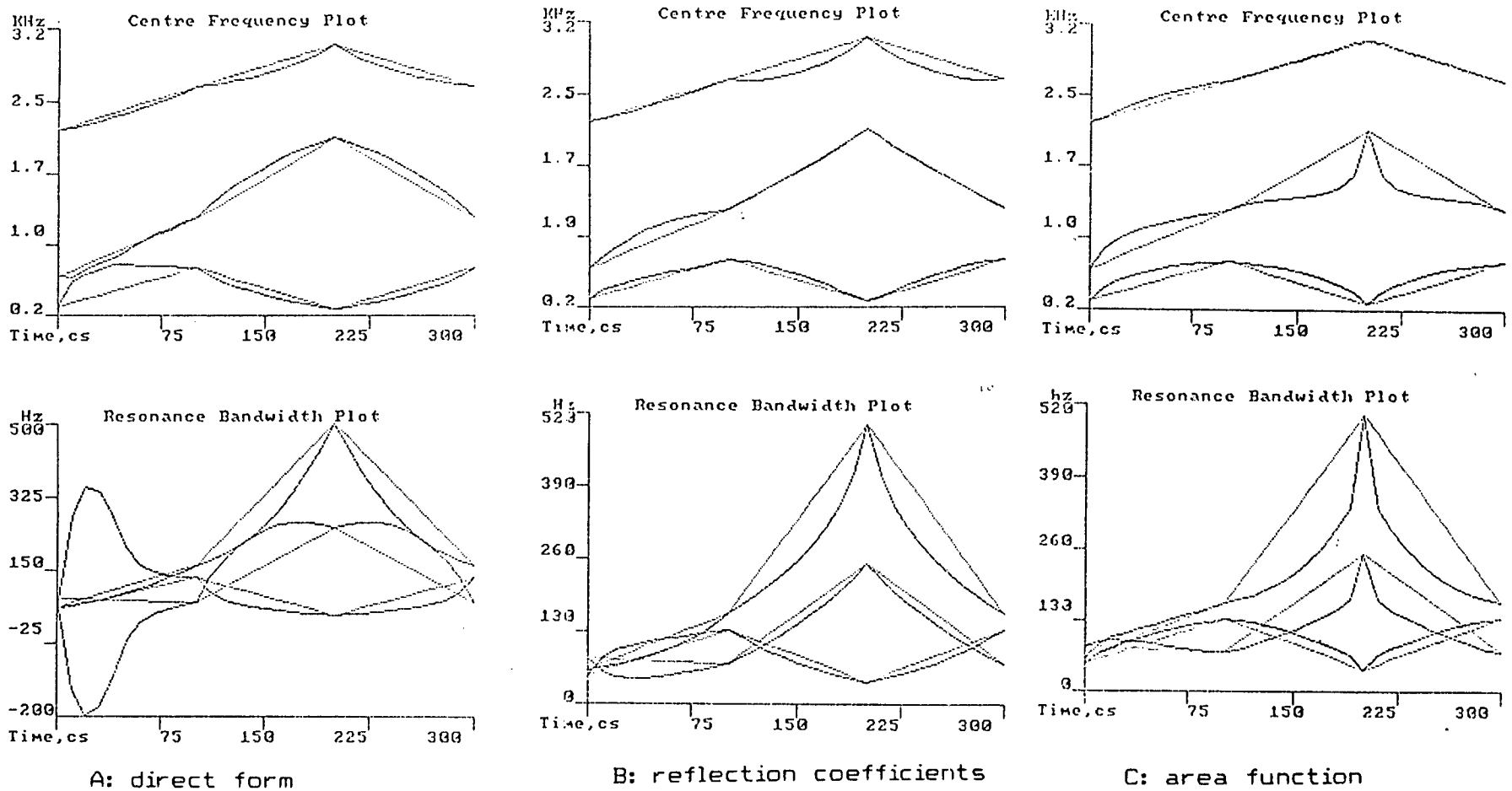


Figure A2.3: Parameters for /waja/. Formant motion for linear interpolation of direct form, reflection coefficient, and area function synthesizer parameters. The straight lines represent comparison paths for series resonance synthesis.

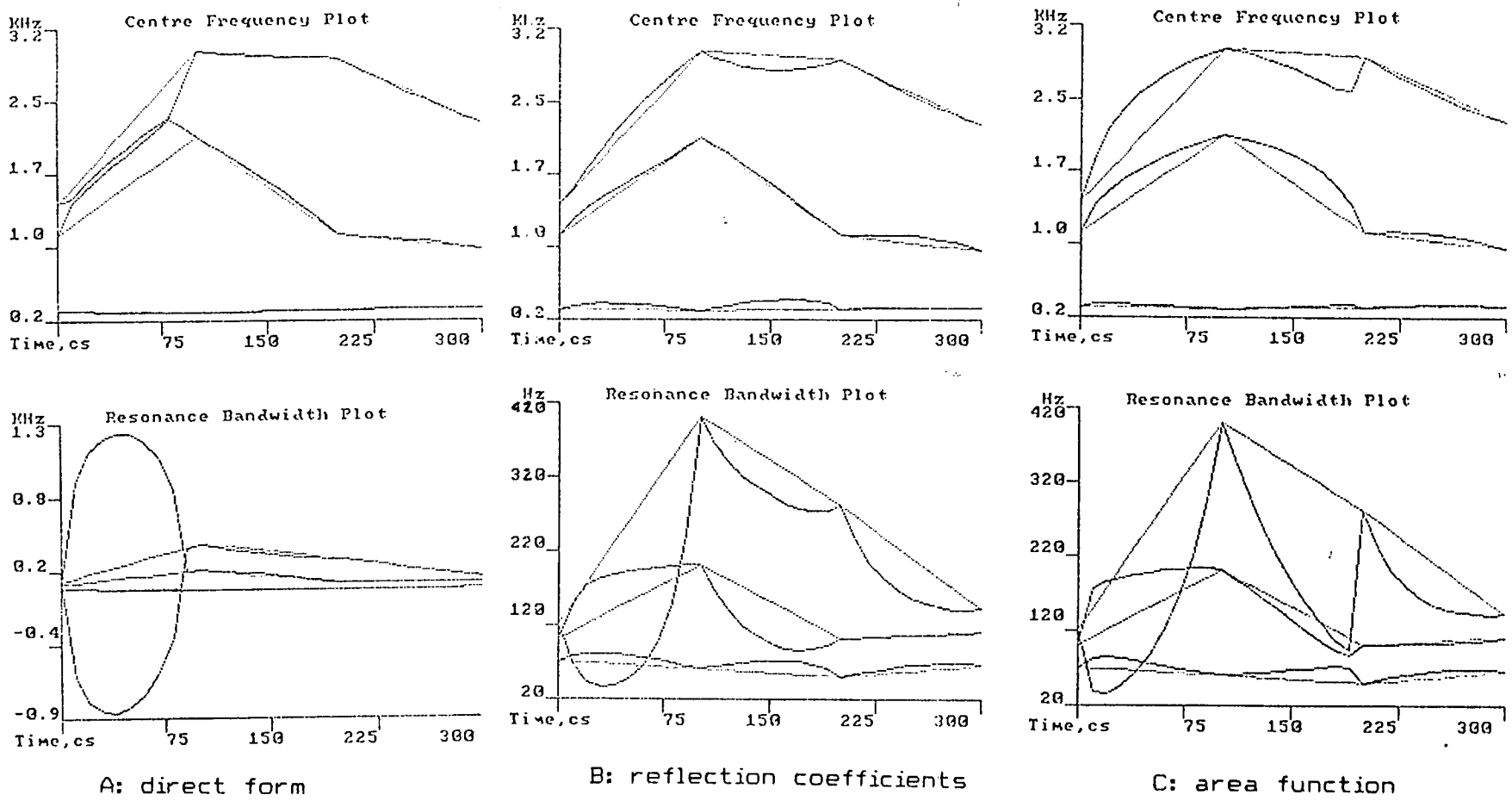


Figure A2.4: Parameters for /rilu/. Formant motion for linear interpolation of direct form, reflection coefficient, and area function synthesizer parameters. The straight lines represent comparison paths for series resonance synthesis.

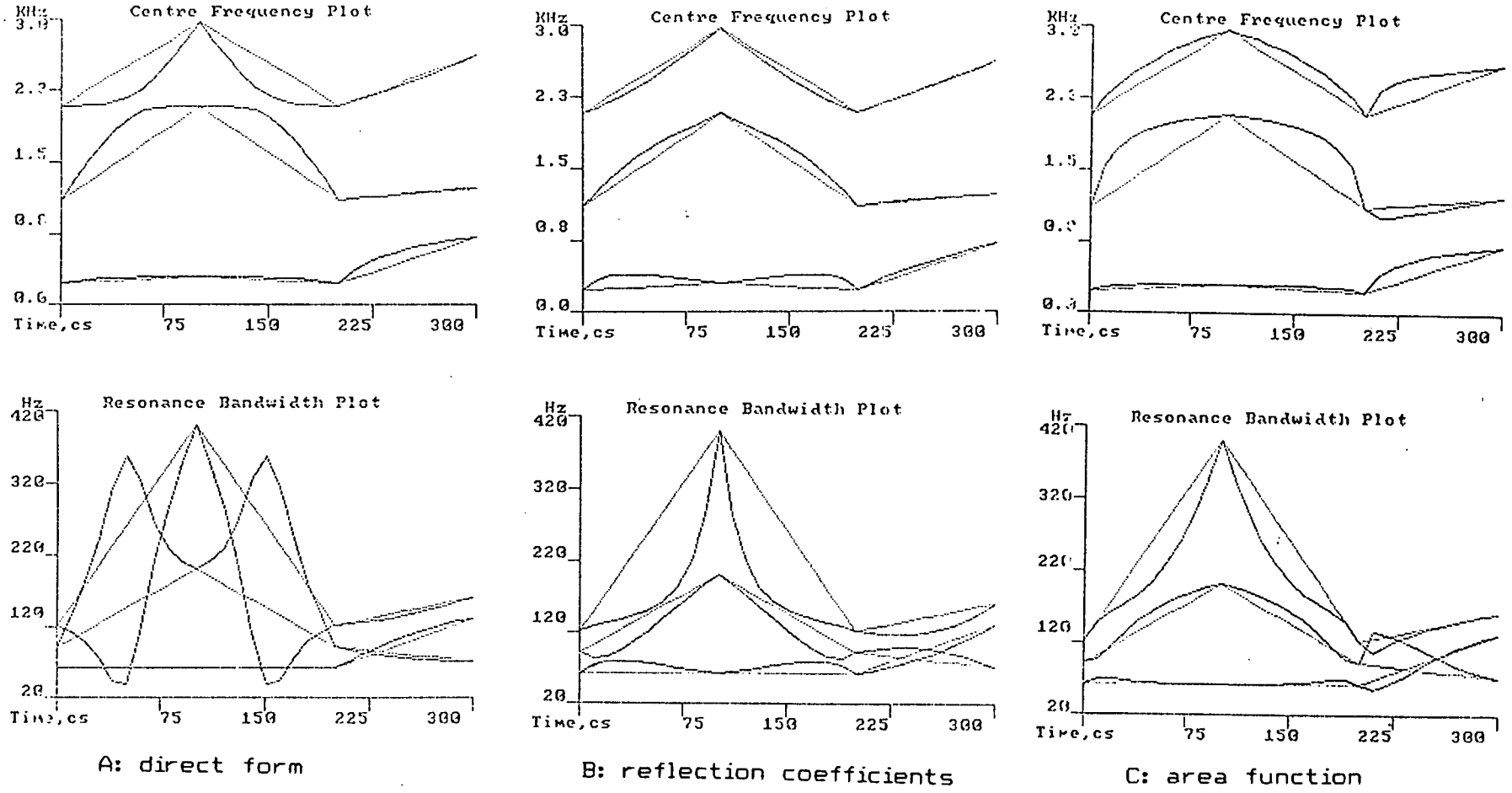


Figure A2.5: Parameters for /viva/. Formant motion for linear interpolation of direct form, reflection coefficient, and area function synthesizer parameters. The straight lines represent comparison paths for series resonance synthesis.

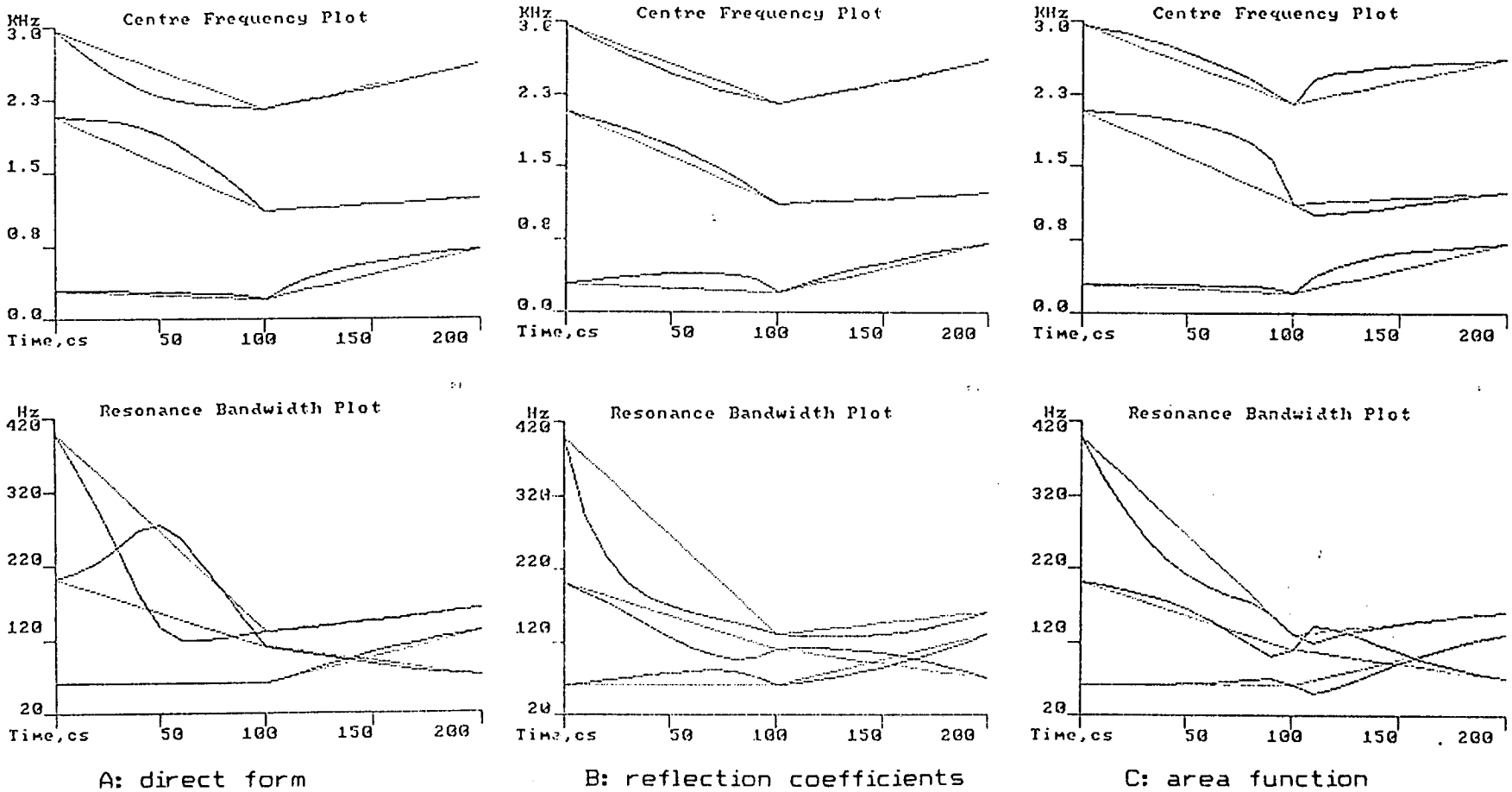
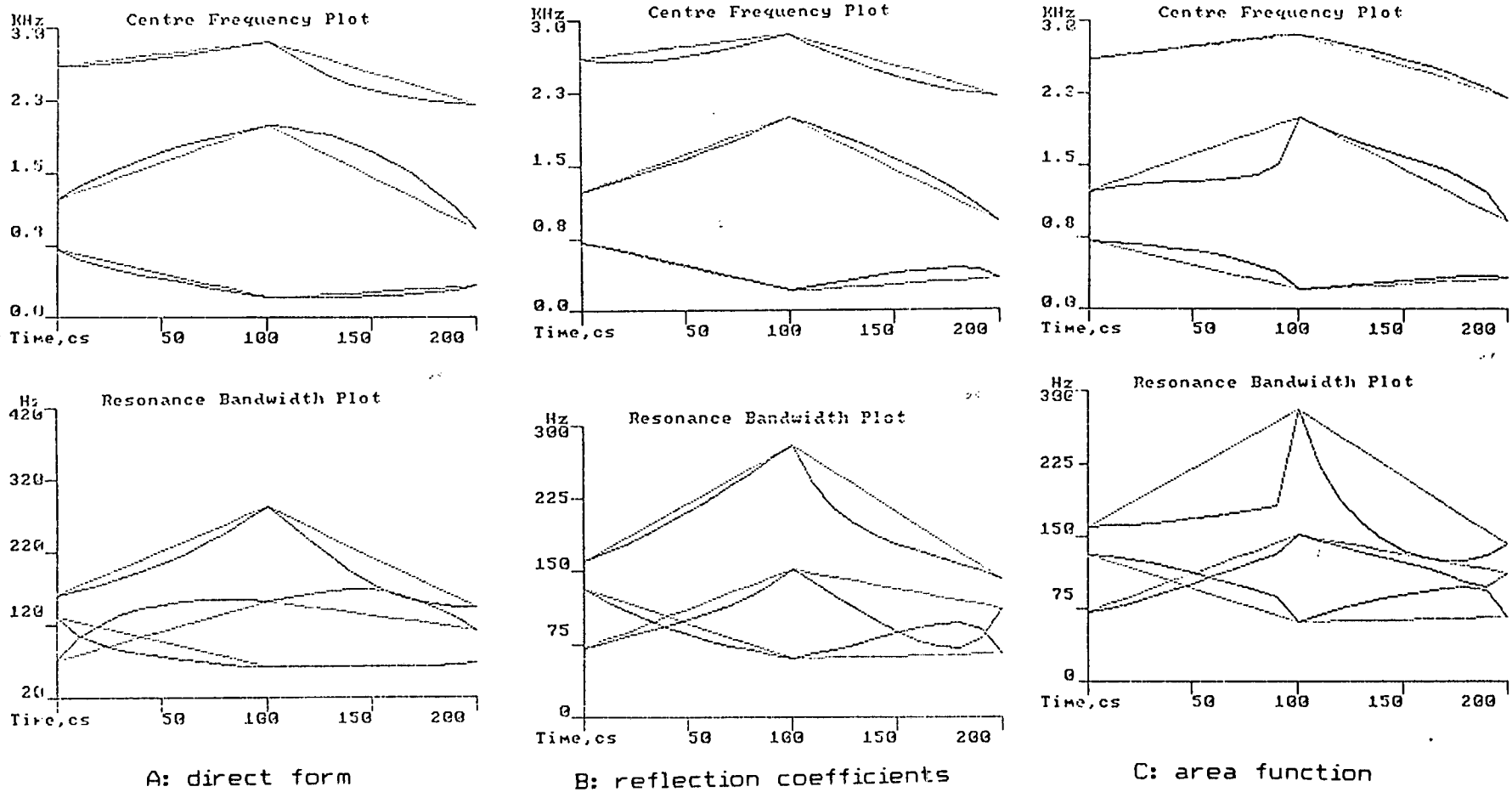


Figure A2.6: Parameters for /iba/. Formant motion for linear interpolation of direct form, reflection coefficient, and area function synthesizer parameters. The straight lines represent comparison paths for series resonance synthesis.



A: direct form

B: reflection coefficients

C: area function

Figure A2.7: Parameters for /agu/. Formant motion for linear interpolation of direct form, reflection coefficient, and area function synthesizer parameters. The straight lines represent comparison paths for series resonance synthesis.

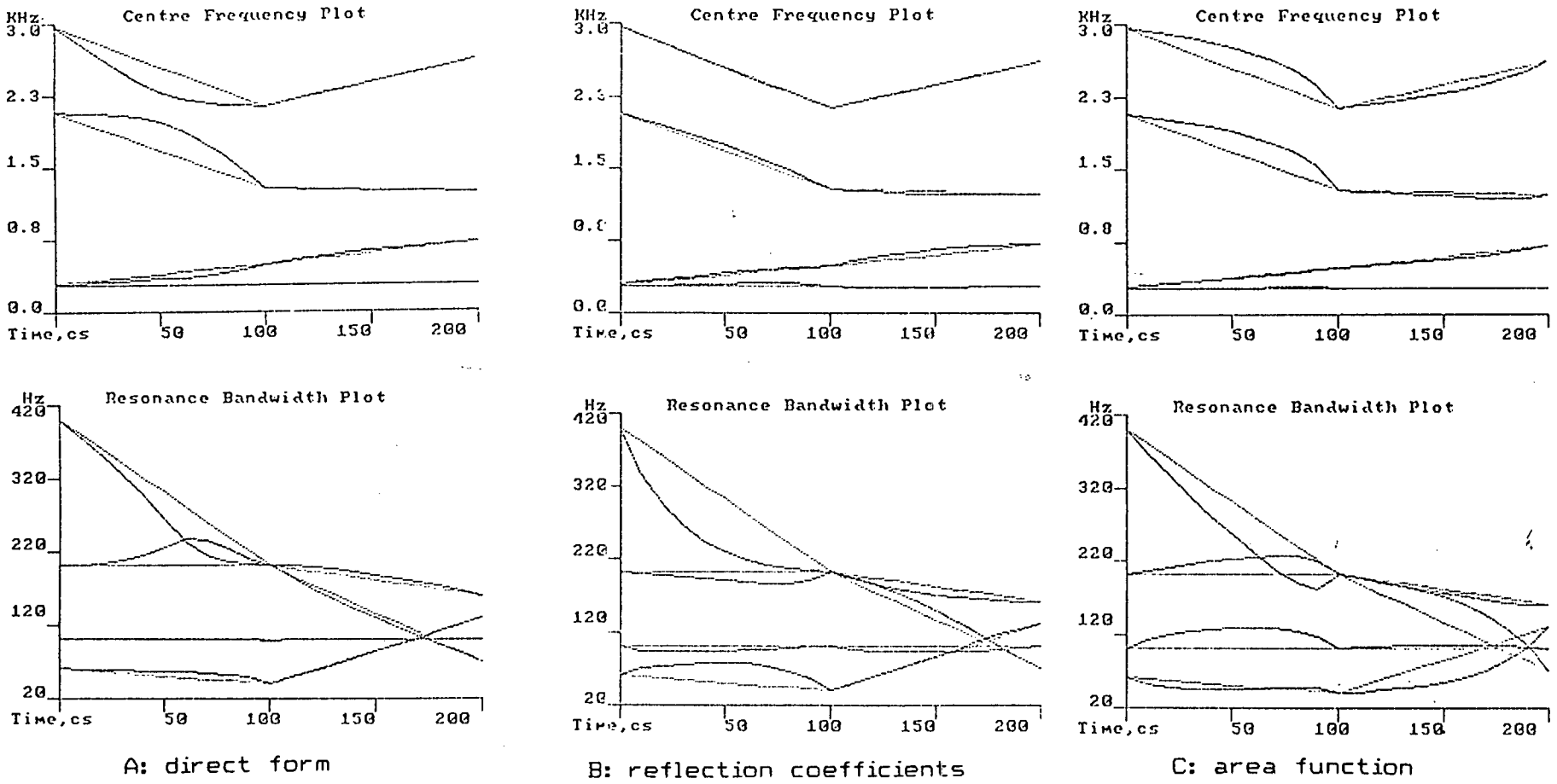


Figure A2.8: Parameters for /ima/. Formant motion for linear interpolation of direct form, reflection coefficient, and area function synthesizer parameters. The straight lines represent comparison paths for series resonance synthesis.

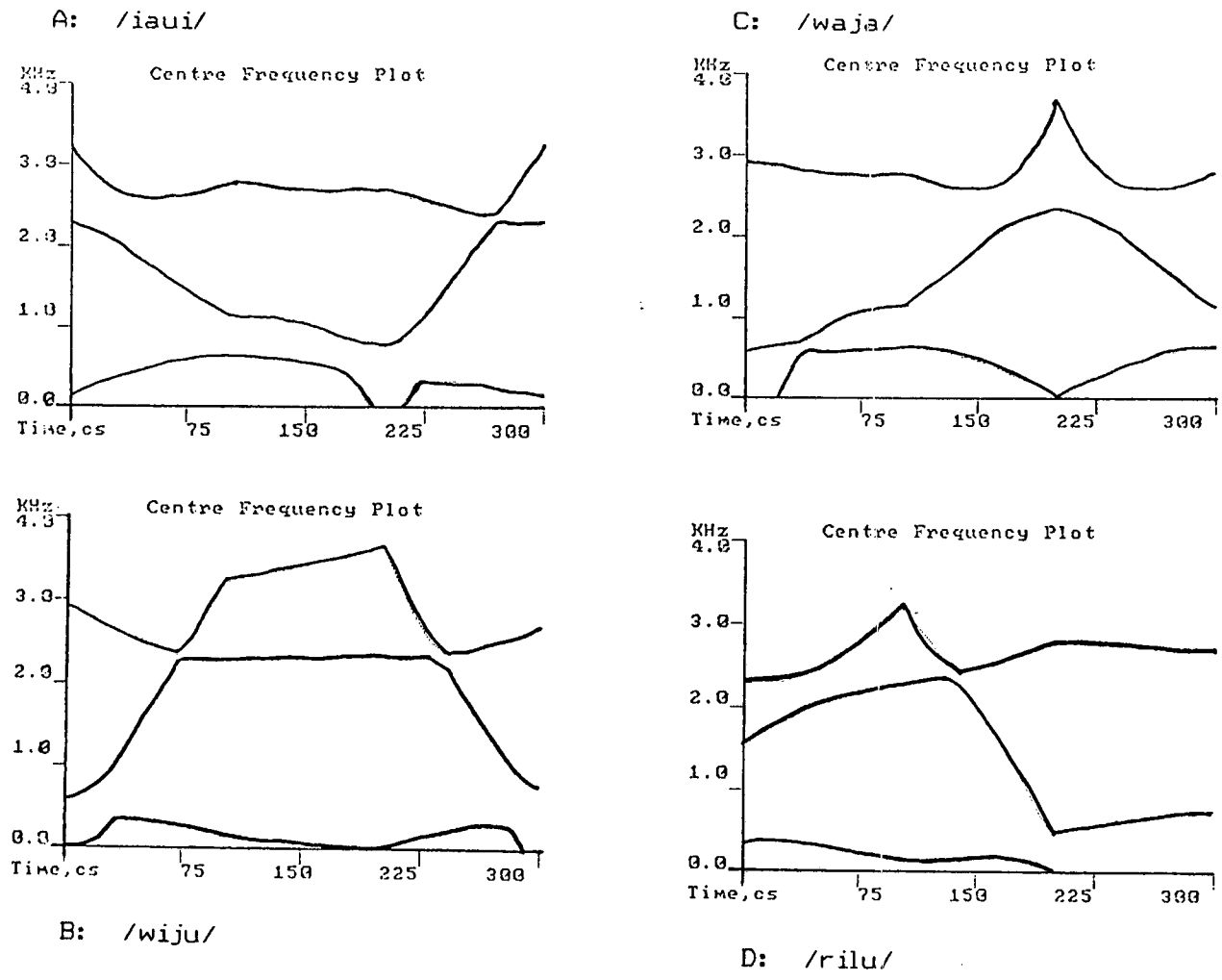
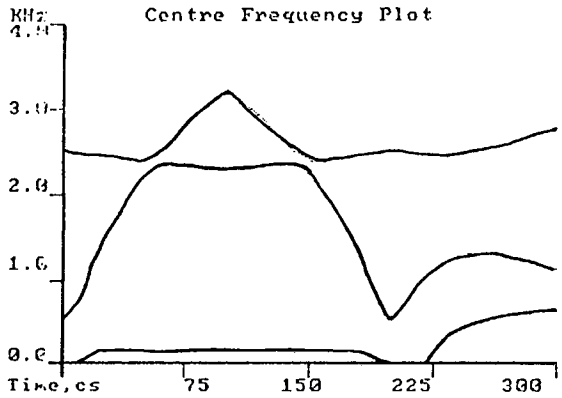
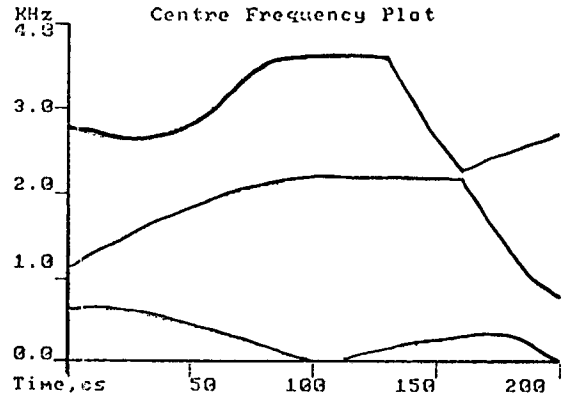


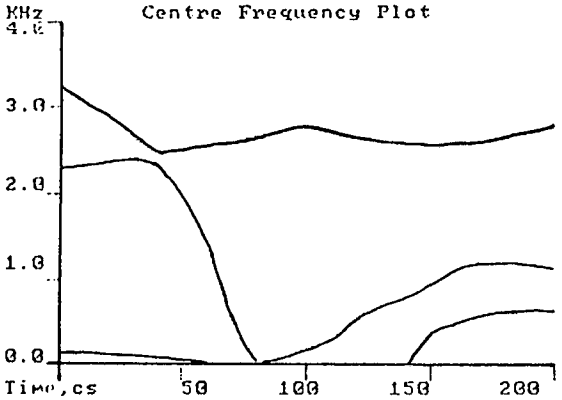
Figure A2.9: Formant motion for linear interpolation of articulatory parameters.

A: /viva/

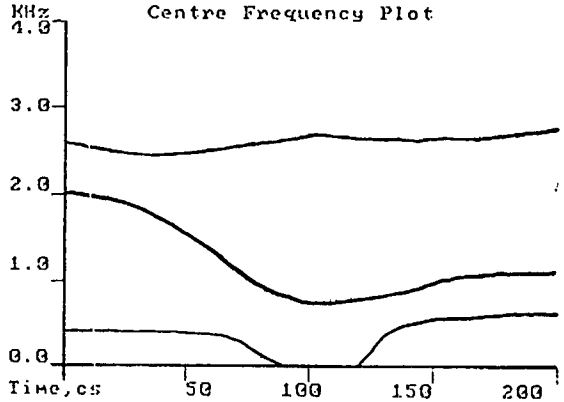


C: /agu/





B: /iba/



D: /ima/

Figure A2.10: Formant motion for linear interpolation of articulatory parameters

Appendix 2.2: Formant trajectories: Quantitative results

Table A2.2-1: Path Differences between linear series resonance transitions and transitions produced by linear interpolation of other synthesizer parameters. Averages of the data in this table are given in Table 6.3.

1. Word = /iaui/

A. Centre frequency differences

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Direct form	10.0	6.5	2.8	6.4	0.0	3.9
Reflections	15.9	4.0	2.5	7.5	0.2	4.6
Area func.	10.3	9.6	1.9	7.2	0.6	4.6
Articulatory	21.2	11.1	8.5	13.6	16.2	14.6

B. Differences in Q

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Parallel	35.0	88.0	68.0	63.6	33.3	51.5
Direct form	6.0	15.2	12.1	11.1	0.0	6.7
Reflections	4.7	26.3	14.4	15.1	3.8	10.6
Area func.	10.3	13.1	36.5	19.9	5.3	14.1
Articulatory	56.7	632.4	148.6	279.2	671.6	436.2

2. Word = /wiju/

A. Centre frequency differences

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Direct form	1.8	12.0	4.6	6.1	0.0	3.6
Reflections	23.6	8.4	2.7	11.5	0.3	7.0
Area func.	7.5	14.1	2.6	8.0	0.5	5.0
Articulatory	51.0	16.7	12.4	26.7	18.3	23.3

B. Differences in Q

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Parallel	9.0	72.0	89.2	56.7	31.9	46.8
Direct form	5.1	15.2	94.5	38.2	0.0	22.9
Reflections	5.9	77.1	27.3	36.7	5.3	24.2
Area func.	10.1	38.5	32.5	27.0	8.3	19.5
Articulatory	83.0	9238.6	286.0	3202.5	10765.3	6227.6

3. Word = /waja/

A. Centre frequency differences

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Direct form	19.1	4.6	1.7	8.4	0.0	5.0
Reflections	7.6	3.8	2.5	4.6	0.2	2.8
Area func.	18.2	13.4	0.9	10.8	0.4	6.6
Articulatory	24.5	9.3	10.5	14.7	16.2	15.3

B. Differences in Q

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Parallel	33.9	99.4	102.4	78.6	40.1	63.2
Direct form	25.0	57.5	13.2	31.9	0.0	19.1
Reflections	16.4	24.0	15.0	18.4	4.0	12.6
Area func.	2.9	11.5	28.1	14.1	4.1	10.1
Articulatory	54.6	881.4	119.0	684.9	3015.4	1617.1

4. Word = /rilu/

A. Centre frequency differences

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Direct form	1.3	6.9	3.0	3.7	0.0	2.2
Reflections	11.8	3.1	2.5	5.8	0.5	3.6
Area func.	4.6	9.5	7.3	7.1	0.8	4.6
Articulatory	56.2	29.7	14.8	33.5	17.8	27.3

B. Differences in Q

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Parallel	14.5	78.4	97.9	63.6	23.8	47.7
Direct form	3.4	23.2	37.3	21.3	0.0	12.7
Reflections	2.9	19.6	100.2	40.8	5.7	26.8
Area func.	7.8	16.0	106.4	43.4	8.9	29.6
Articulatory	70.4	621.2	3305.3	1332.3	2075.6	1629.6

5. Word = /viva/

A. Centre frequency differences

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Direct form	11.5	9.4	5.6	8.8	0.0	5.2
Reflections	25.8	3.8	1.6	10.4	0.2	6.3
Area func.	15.6	13.9	4.9	11.4	0.6	7.1
Articulatory	37.5	20.2	7.6	21.7	18.6	20.5

B. Differences in Q

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Parallel	27.0	86.2	95.2	69.4	31.8	54.4
Direct form	8.9	18.8	84.2	37.2	0.0	22.3
Reflections	15.1	18.1	32.0	21.7	4.7	14.9
Area func.	15.1	15.7	16.6	15.7	8.1	12.7
Articulatory	52.7	1075.6	2294.3	1140.9	4754.0	2586.1

6. Word = /iba/

A. Centre frequency differences

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Direct form	15.3	5.8	3.5	8.1	0.0	4.9
Reflections	26.2	2.3	1.4	9.9	0.1	6.0
Area func.	20.9	11.9	4.2	12.3	0.5	7.6
Articulatory	47.2	21.1	7.1	25.1	18.8	22.6

B. Differences in Q

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Parallel	32.3	71.4	83.9	62.5	35.6	51.7
Direct form	12.9	10.5	17.6	13.6	0.0	8.1
Reflections	17.9	15.7	20.0	17.8	3.4	12.1
Area func.	25.2	18.2	12.4	18.5	6.4	13.7
Articulatory	49.1	2760.5	6726.4	3178.7	20537.2	10122.1

7. Word = /ima/

A. Centre frequency differences

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Direct form	0.6	5.8	5.5	3.9	1.5	2.9
Reflections	4.0	4.9	1.7	3.5	0.1	2.1
Area func.	1.9	2.0	5.8	3.2	2.3	2.8
Articulatory	73.4	61.7	96.8	77.2	41.6	63.0

B. Differences in Q

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Parallel	40.1	245.2	594.4	293.2	2651.3	1236.5
Direct form	0.9	11.2	3.0	5.0	1.7	3.6
Reflections	7.6	14.2	6.2	9.3	7.9	8.7
Area func.	7.8	18.8	9.1	11.9	6.8	9.8
Articulatory	118.4	54.8	851.9	341.7	1454.0	786.6

8. Word = /agu/

A. Centre frequency differences

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Direct form	9.5	8.3	2.6	6.8	0.0	4.0
Reflections	15.6	3.5	2.6	7.2	0.1	4.4
Area func.	17.1	11.0	1.1	9.7	0.3	5.9
Articulatory	41.5	21.5	11.6	24.8	18.2	22.1

B. Differences in Q

Synthesizer type	F1	F2	F3	Ave F1-F3	Ave F4-F5	Ave F1-F5
Parallel	31.2	88.6	54.2	58.0	36.4	49.3
Direct form	8.7	9.2	5.8	7.9	0.0	4.7
Reflections	6.7	24.7	4.6	12.0	3.0	8.4
Area func.	13.4	12.2	32.1	19.2	3.3	12.8
Articulatory	84.4	4924.9	3549.9	2853.1	6746.6	4410.5

Table A2.2-2: Average Spectral Difference between linear series resonance transitions and transitions produced by linear interpolation of other synthesizer parameters. Averages of the data in this table are given in Table 6.4.

synthesiser type	word: iaui	wiju	waja	rilu
Direct form	2.57	2.06	3.12	1.26
Reflection coeff.	3.09	4.47	2.18	2.61
Area function	3.14	3.24	4.13	2.80
Parallel	3.77	3.10	4.75	3.59
Articulatory	8.18	16.28	9.83	16.46

	word: viva	iba	ima	agu
Direct	2.93	3.11	1.67	2.46
Reflection coeff.	4.27	4.08	1.43	2.99
Area function	4.19	4.31	1.82	3.42
Parallel	3.16	3.26	20.09	3.23
Articulatory	12.12	19.68	21.31	14.90

Appendix Three: Acoustic consequences of articulatory parameters.

This appendix contains contour plots to supplement Figure 5.6. Resonance frequencies and bandwidths are plotted as contours over the two-dimensional space of articulatory control parameters FR (front raising) and BR (back raising). Thick lines show the limit of the region of positive vocal tract areas. All values in Hz.

Figure 5.6 is for a lip opening of 1 cm^2. A total of ten lip openings were investigated, spaced two per octave from 0.25 cm^2 to 5.6 cm^2. Data for all five resonant frequencies were plotted, making fifty graphs in total. Of these, the data for the fourth and fifth formants showed very little variation with lip opening, and the graphs for lip openings of 4.0 cm^2 were nearly identical to those with a 5.6 cm^2 opening. The remaining 27 plots (nine lip openings, three formants) are given in this appendix.

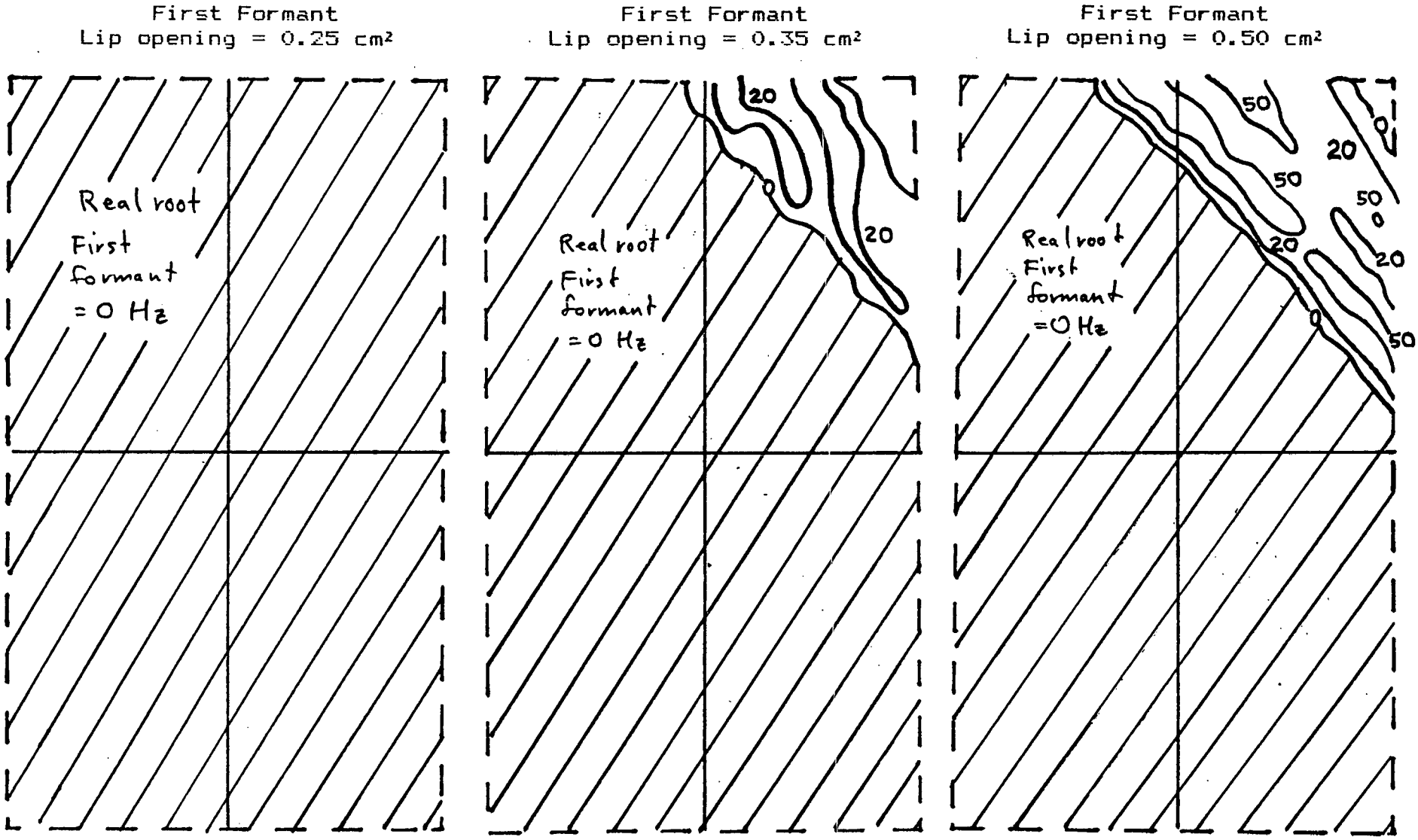


Figure A3.1: Contour plots of first formant frequency vs tongue parameters.

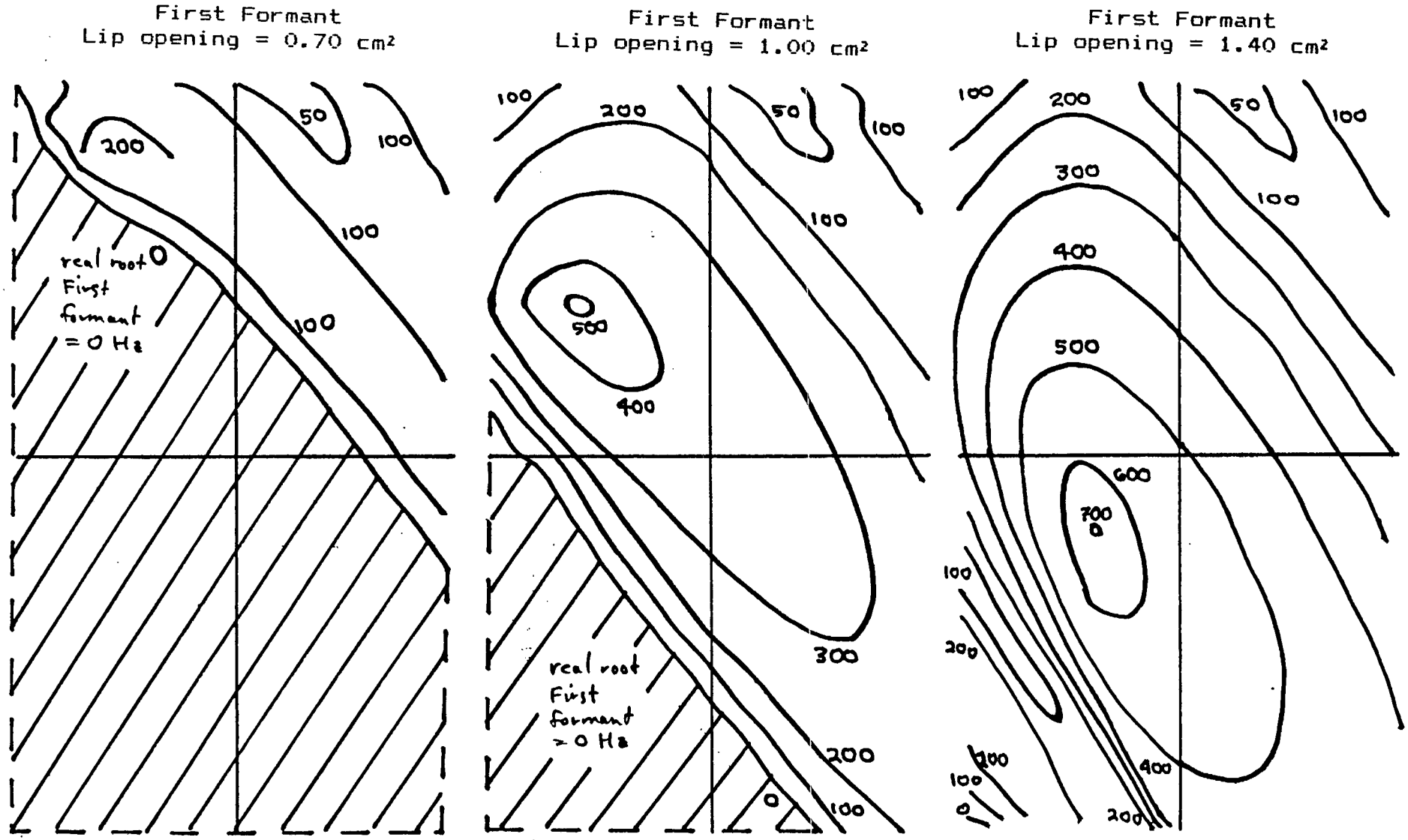


Figure A3.2: Contour plots of first formant frequency vs tongue parameters.

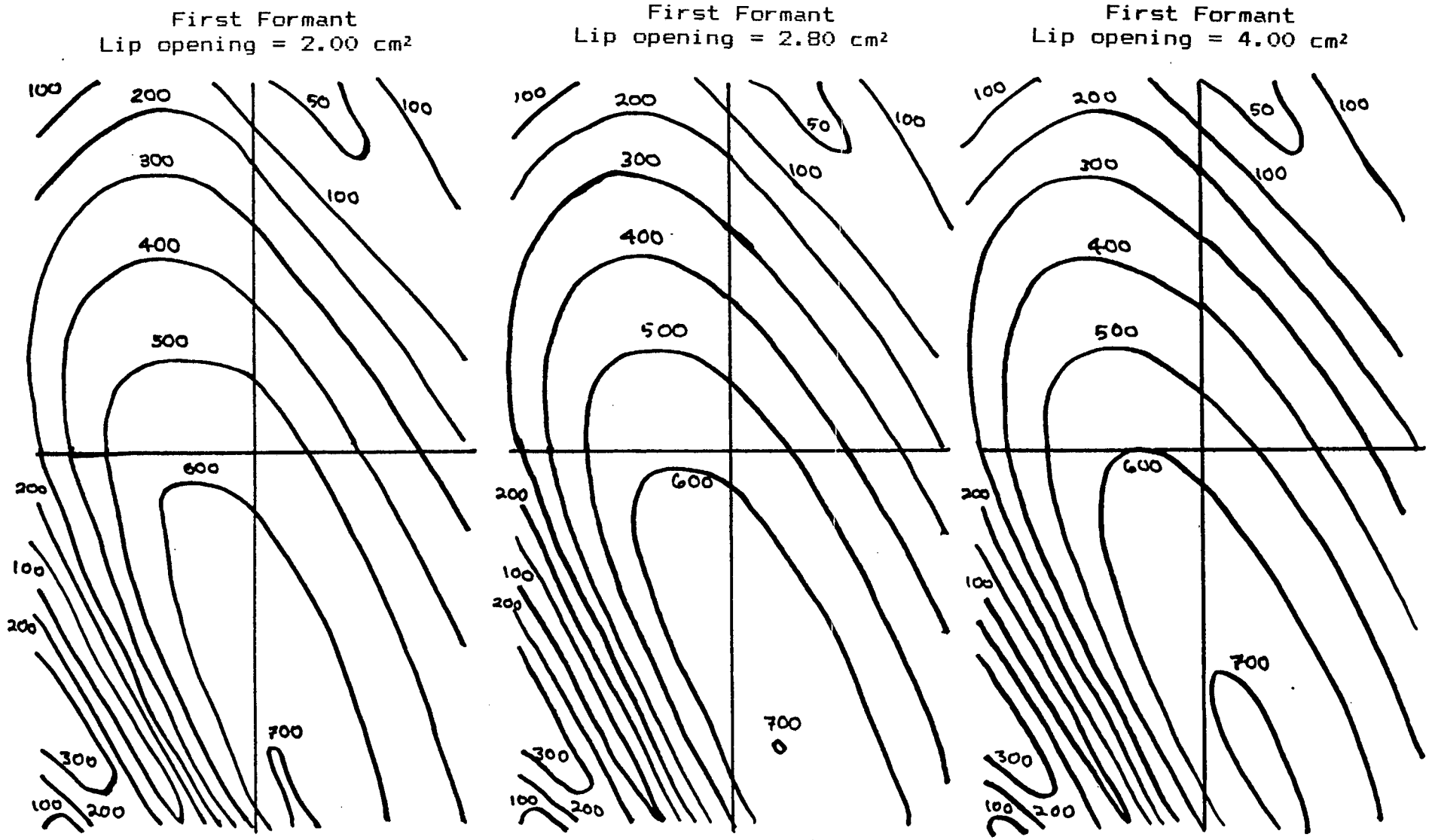


Figure A3.3: Contour plots of first formant frequency vs tongue parameters.

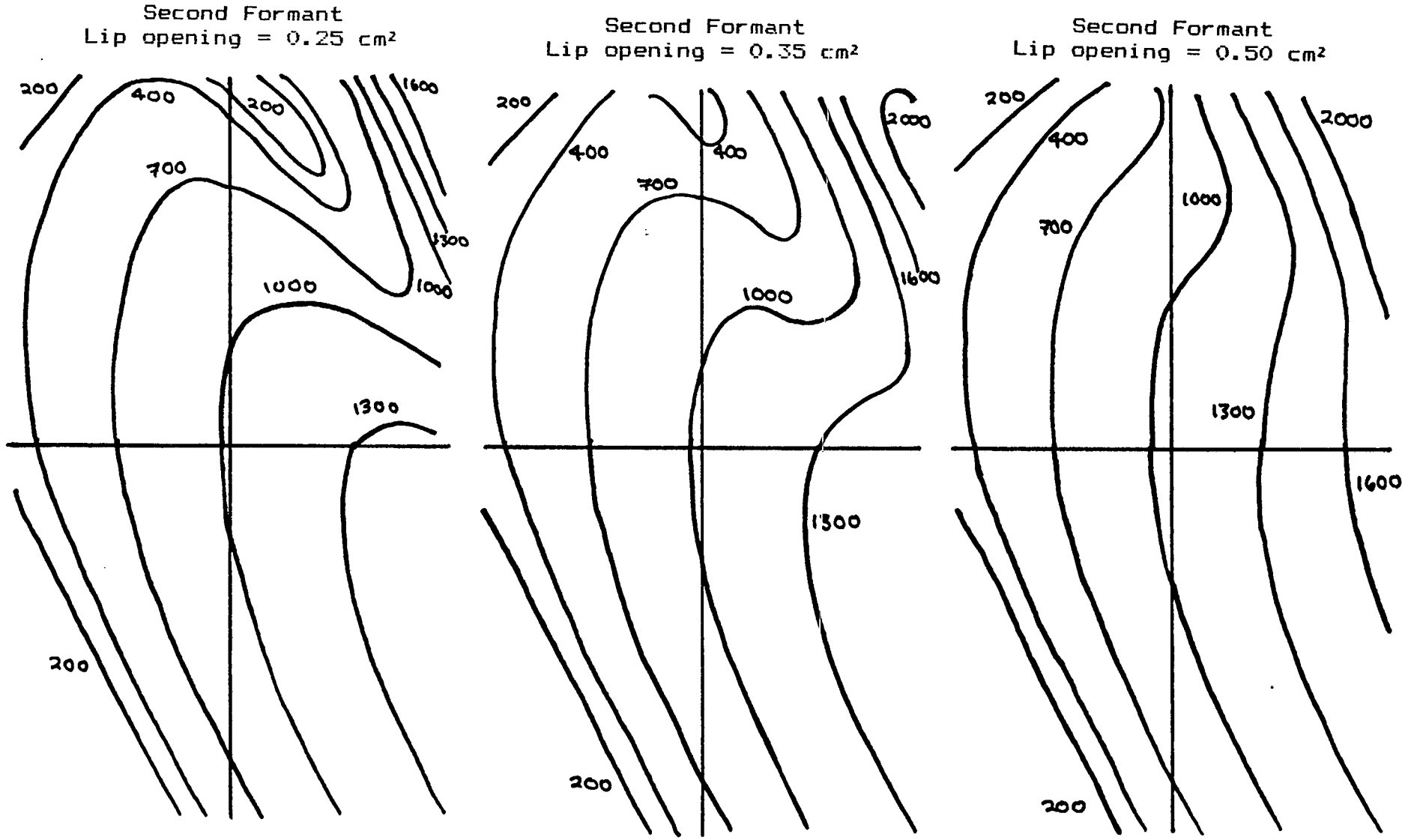


Figure A3.4: Contour plots of second formant frequency vs tongue parameters.

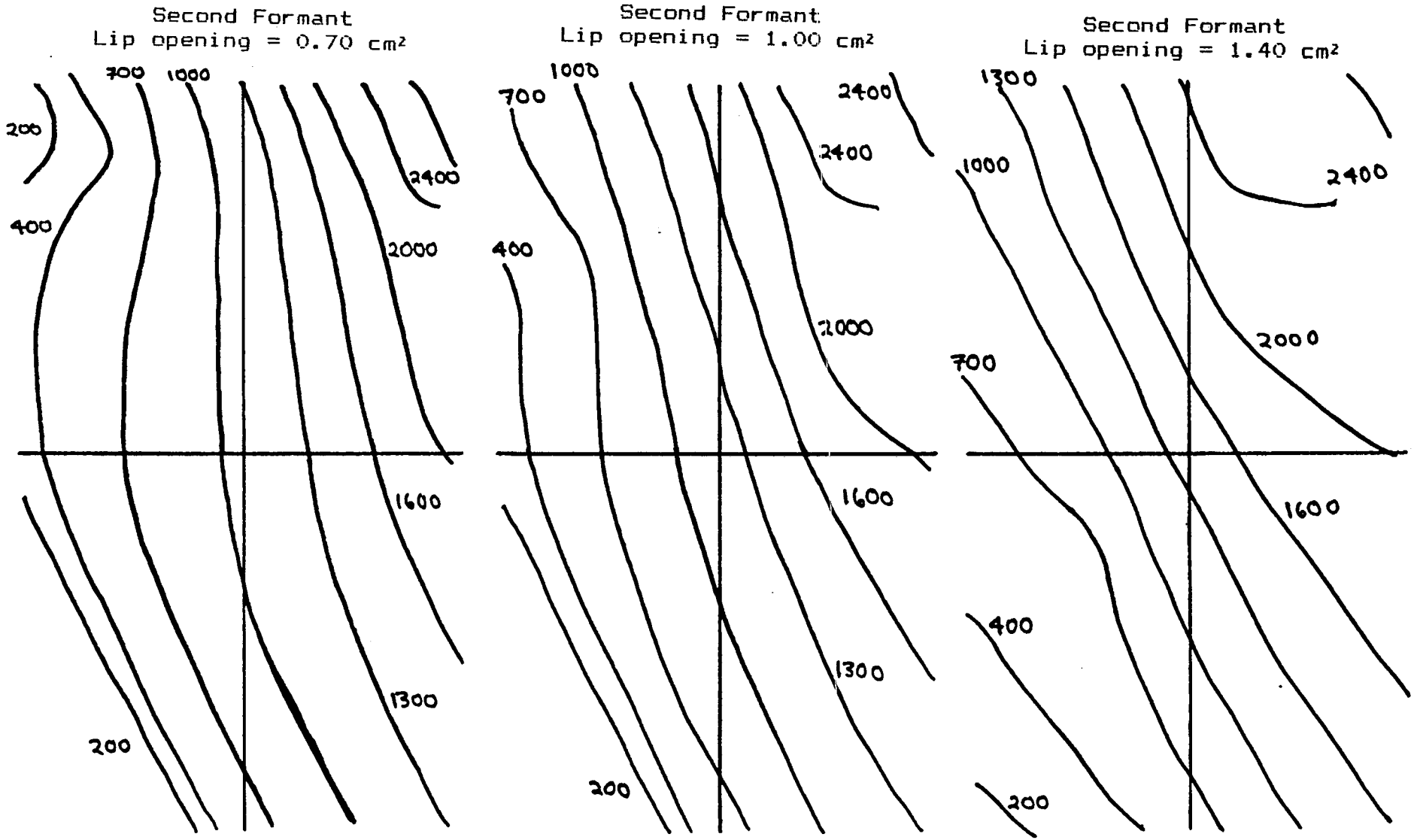


Figure A3.5: Contour plots of second formant frequency vs tongue parameters.

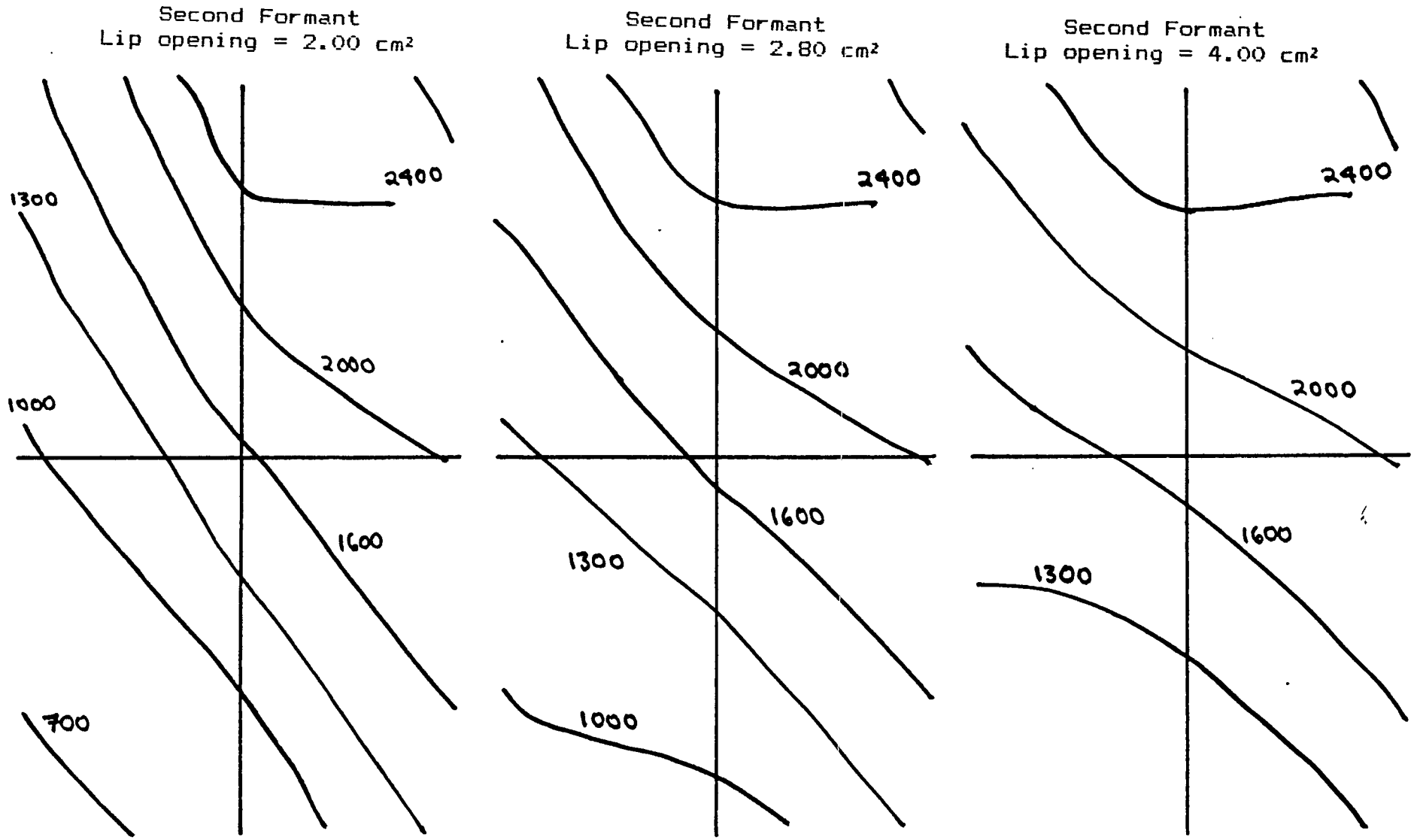


Figure A3.6: Contour plots of second formant frequency vs tongue parameters.

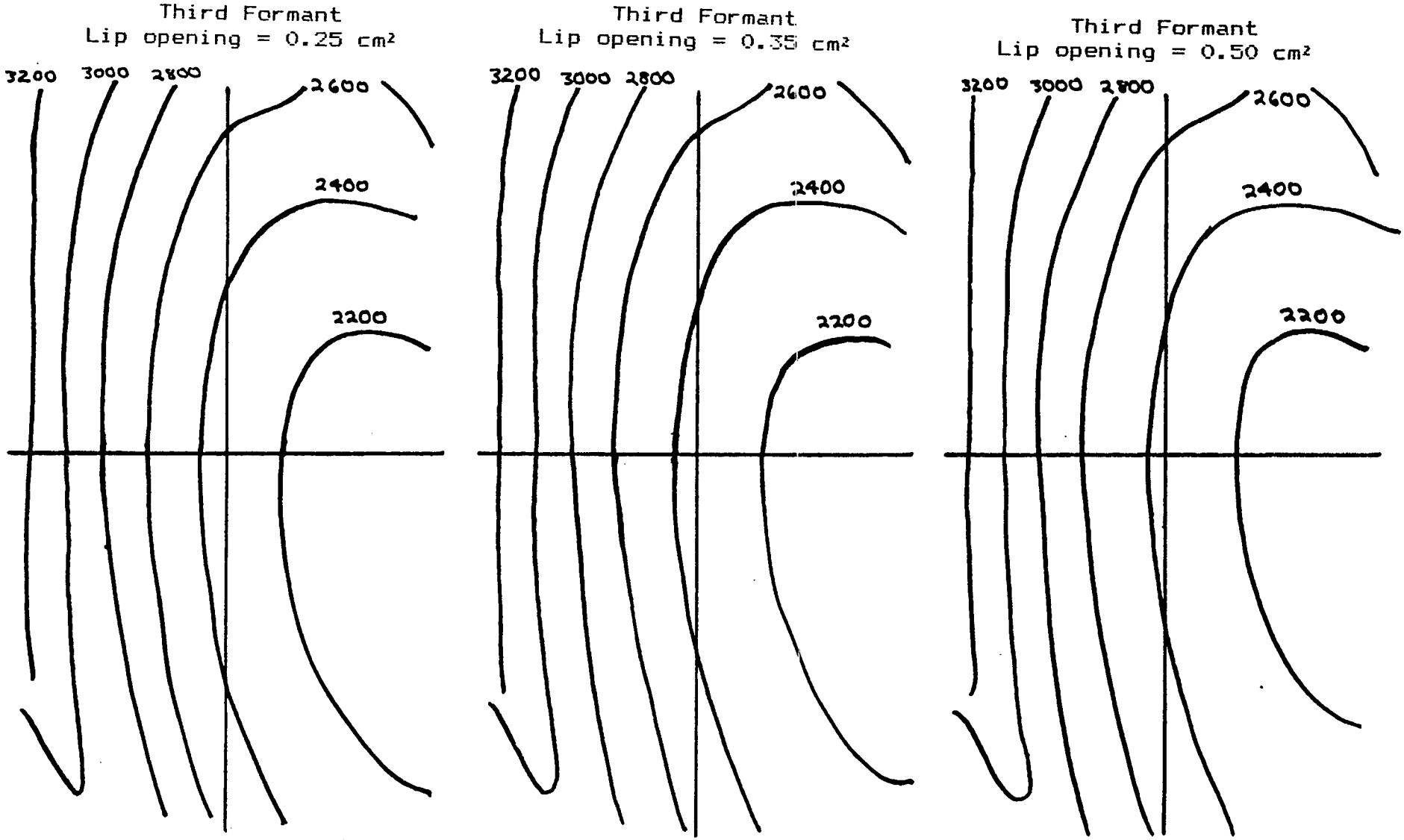


Figure A3.7: Contour plots of third formant frequency vs tongue parameters.

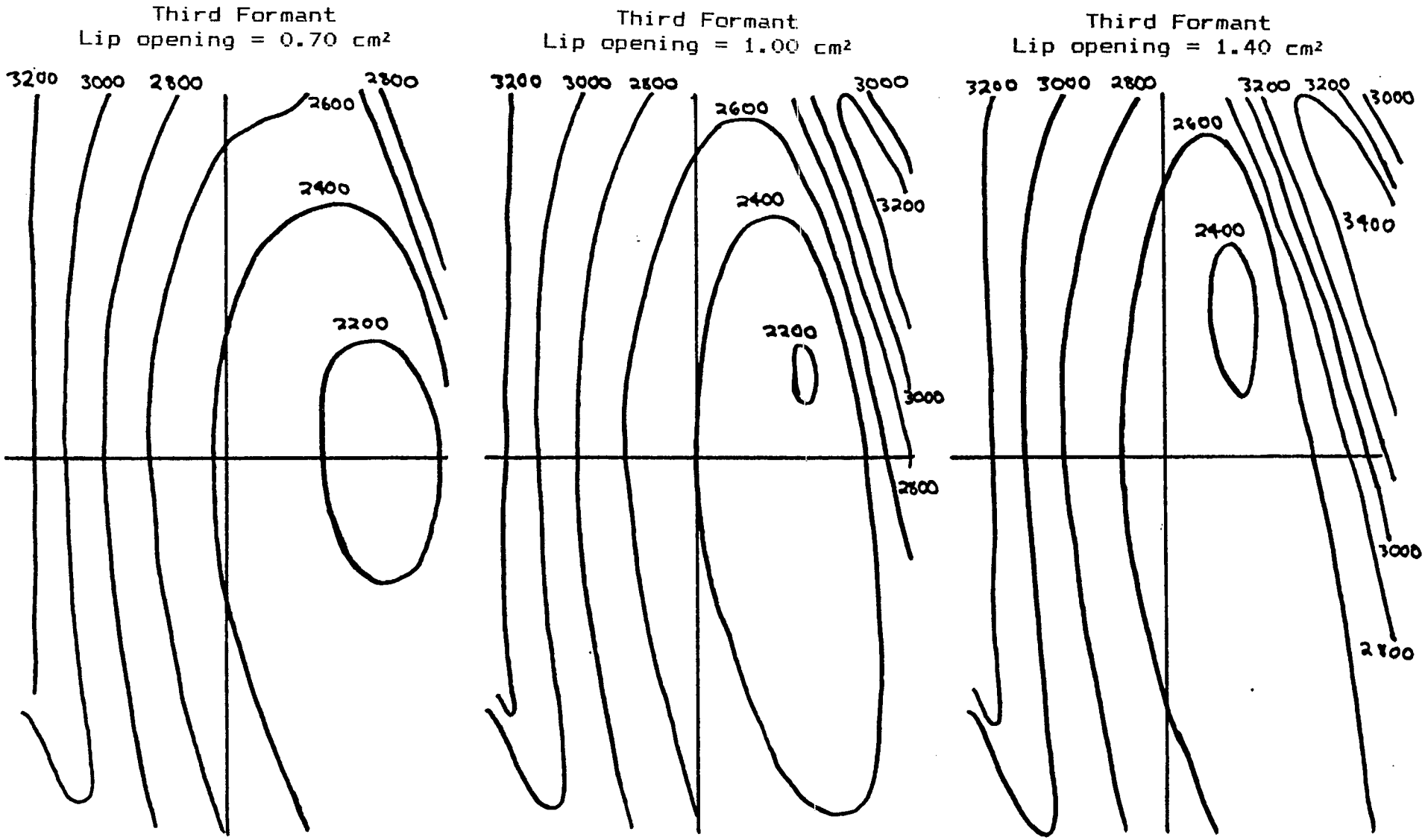


Figure A3.8: Contour plots of third formant frequency vs tongue parameters.

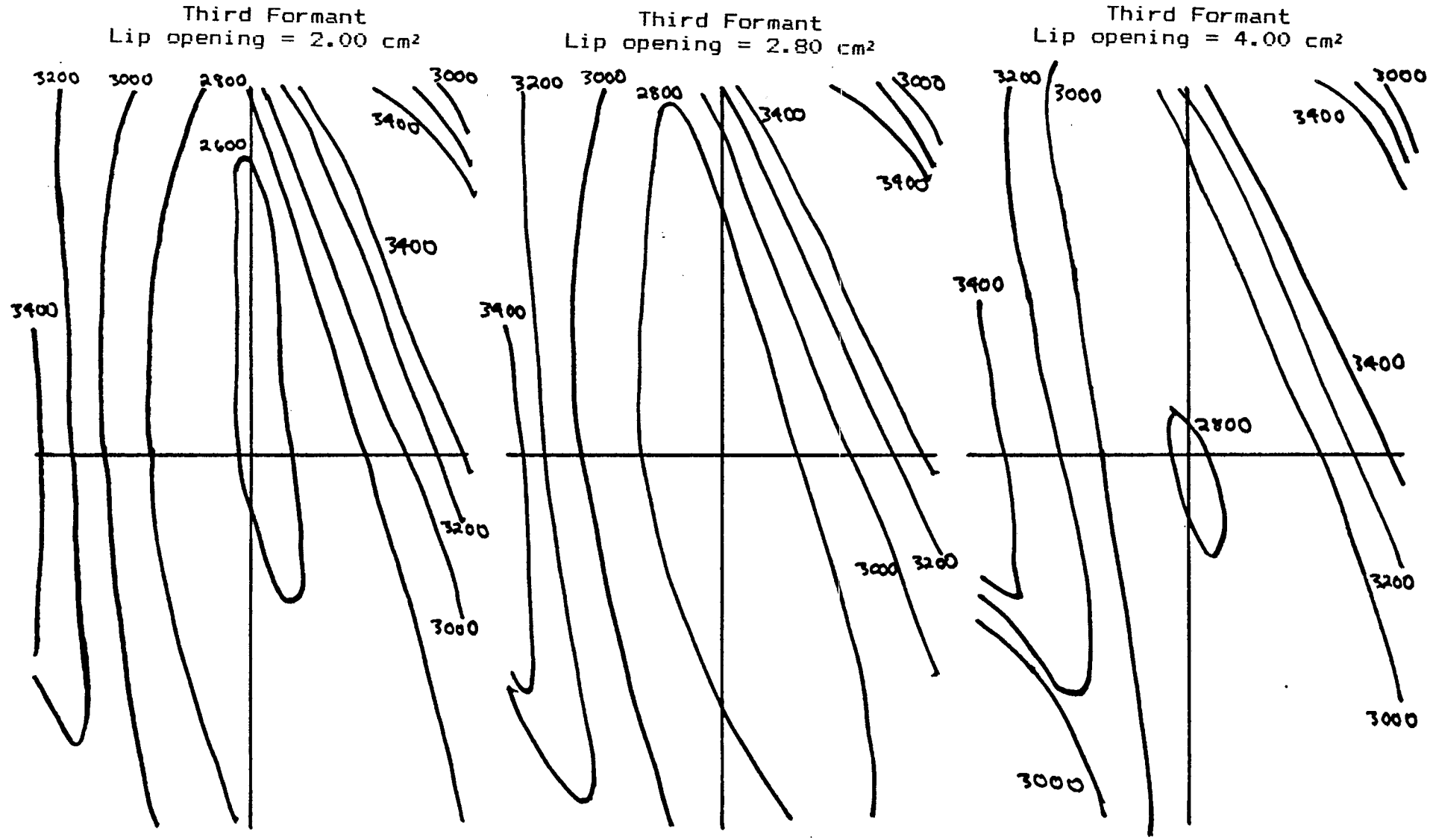


Figure A3.9: Contour plots of third formant frequency vs tongue parameters.

Appendix 4: Intelligibility test details

A4.1 Phoneme Parameter Values

Table A4.1: Klatt (1980b, 986-987) formant data.
[Symbols used for sounds defined in List of Symbols, page x.]

sound	cf1	bw1	cf2	bw2	cf3	bw3	
i	290	60	2070	200	2960	400	/i/
I	400	50	1800	100	2570	140	ih
e	530	60	1680	90	2500	200	eh
æ	620	70	1660	150	2430	320	/ae/
a	700	130	1220	70	2600	160	/a/
^	620	80	1220	50	2550	140	neutral
ɒ	540	80	1100	70	2300	70	US oh
o	600	90	990	100	2570	80	US aw
ʊ	450	80	1100	100	2350	80	omega
u	320	65	900	110	2200	140	/u/
w	290	50	610	80	2150	60	
j	260	40	2070	250	3020	500	
r	310	70	1060	100	1380	120	
l	310	50	1050	100	2880	280	

cf = centre frequency; bw = bandwidth;

Fourth and fifth formants have these fixed values for vowels:

cf4	bw4	cf5	bw5
3300	500	3850	700

a = amplitude in dB re nominal 'off' (silence) value of 0 dB.

							a2	a3	a4	a5	a6	ab
f	340	200	1100	120	2080	150	0	0	0	0	0	57
v	220	60	1100	90	2080	120	0	0	0	0	0	57
θ	320	200	1290	90	2540	200	0	0	0	0	28	48
ð	270	60	1290	80	2540	170	0	0	0	0	28	48
s	320	200	1390	80	2530	200	0	0	0	0	52	0
z	240	70	1390	60	2530	180	0	0	0	0	52	0
S	300	200	1840	100	2750	300	0	57	48	48	46	0
p	400	300	900	120	2150	220	0	0	0	0	0	63
b	200	60	900	110	2150	130	0	0	0	0	0	63
t	400	300	1600	120	2600	250	0	30	45	57	63	0
d	200	60	1600	100	2600	170	0	47	60	62	60	0
k	300	250	2400	120	2850	330	0	53	43	45	45	0
g	200	60	2400	120	2850	280	0	53	43	45	45	0

snd	cf1	bw1	cf2	bw2	cf3	bw3	fnp	fnz
m	480	40	1270	200	2130	200	270	450
n	480	40	1340	300	2470	300	270	450

(nasal pole and zero frequency)

Table A4.2: 10th order all-pole approximation to Klatt data.

snd	cf1	bw1	cf2	bw2	cf3	bw3	cf4	bw4	cf5	bw5	ampl
i	290	60	2070	200	2960	400	3300	500	3850	700	60
I	400	50	1800	100	2570	140	3300	500	3850	700	60
e	530	60	1680	90	2500	200	3300	500	3850	700	60
æ	620	70	1660	150	2430	320	3300	500	3850	700	60
a	700	130	1220	70	2600	160	3300	500	3850	700	60
^	620	80	1220	50	2550	140	3300	500	3850	700	60
ɒ	540	80	1100	70	2300	70	3300	500	3850	700	60
o	600	90	990	100	2570	80	3300	500	3850	700	60
ʊ	450	80	1100	100	2350	80	3300	500	3850	700	60
u	320	65	900	110	2200	140	3300	500	3850	700	60
w	290	50	610	80	2150	60	3300	500	3850	700	50
j	260	40	2070	250	3020	500	3300	500	3850	700	50
r	310	70	1060	100	1380	120	3300	500	3850	700	50
l	310	50	1050	100	2880	280	3300	500	3850	700	50
f	340	200	1100	120	2080	150	3300	1000	3850	1000	45
v	220	60	1100	90	2080	120	3300	1000	3850	1000	50
					Simulate AB parameter with flat F4, F5						
θ	320	200	1290	90	2540	200	3300	1000	4900	600	40
ð	270	60	1290	80	2540	170	3300	1000	4900	600	40
									Lift high end of spectrum		
s	320	300	1390	200	2540	300	4000	2000	5000	300	50
z	240	70	1390	60	2540	180	4000	2000	5000	300	55
S	300	200	1840	100	2500	300	2750	500	4900	1000	50
Z	240	70	1840	100	2500	300	2750	500	4900	1000	55
			Sibilants; hand-trimmed; main energy below 3 kHz								
p	400	300	900	120	2150	220	3300	2000	3850	2000	45
b	200	60	900	110	2150	130	3300	2000	3850	2000	45
					Simulate AB parameter with flat F4, F5						
t	400	300	1600	120	2600	250	3300	500	4900	500	45
d	200	60	1600	100	2600	170	3300	500	4900	500	45
										Lift F4, F5	
k	300	250	2400	120	2850	330	3300	1000	4900	1000	45
g	200	60	2400	120	2850	280	3300	1000	4900	1000	45
							Intermediate case of F4, F5				
m	300	100	900	200	2150	300	3300	500	3850	700	45
n	300	100	1600	200	2600	400	3300	500	3850	700	45
N	300	100	2400	200	2850	400	3300	500	3850	700	45
	F1 for nasal resonance; nasals should have a zero										

A4.2 FAAF Wordlist

mail	bin	boast	dab	man	taught	seal	sheen	mash	ritz
bail	din	ghost	gab	nan	port	feel	seen	mass	rich
nail	pin	coast	tab	van	thought	veal	keen	match	ridge
dale	tin	post	cab	than	fought	zeal	teen	mats	rids
some	rib	cob	bag	get	mix	rose	lands	bang	ham
sun	rig	cod	bad	bet	milks	rove	lads	ban	high
sub	rip	cop	bat	wet	mick	robe	lad	bad	hang
sud	rick	cot	back	yet	milk	rode	land	bag	how

Practice items:	what	hold	lest	stone	stream
	rot	cold	nest	own	scream
	lot	old	rest	tone	scheme
	yacht	gold	messed	sown	steam

A4.3 Synthesizer Control Data

A4.3.1 Synthesizer control 'spellings' examples

Lines beginning with * are just comments, and are used here as reminders of the requisite excitation data and gain data.

The format for the actual command lines is: phoneme code, segment duration, transition type, and interpolation interval.

The following 'spellings' will synthesize the four tokens mix, milks, milk and Mick. All use the excitation file VVR, which has 400 msec of pulses (fully voiced), and then 100 msec of random noise (fully unvoiced). The two tokens with stop+fricative clusters use the gain contour CVSC2, for a stop followed by a voiceless consonant in syllable-final position. The other two words end in a voiceless stop, and require the gain contour CVS2. Note that the gain (and excitation) are identical irrespective of presence of the approximate /l/ before the /k/. This is because the /l/ can be treated as part of the vowel, as far as gain, excitation and duration are concerned. (Indeed in many dialects the /l/ in this position would be vocalic, eg [mIuk].)


```
* mix
* exc: vvr
*gain: cvsc2
*
m 0    B
m 100  L 100
I 50   L 10
I 170  L 170
k 50   L 10
k 30   L 30
s 20   L 10
s 80   L 80
*
*
*
* milk
* exc: vvr
*gain: cvs2
*
m 0    B
m 100  L 100
I 50   L 10
I 100  L 100
l 40   L 10
l 30   L 30
k 50   L 10
k 130  L 130
*
```

```
* milks
* exc: vvr
*gain: cvsc2
*
m 0    B
m 100  L 100
I 50   L 10
I 100  L 100
l 40   L 10
l 30   L 30
k 50   L 10
k 30   L 30
s 20   L 10
s 80   L 80
*
* mick
* exc: vvr
*gain: cvs2
*
m 0    B
m 100  L 100
I 50   L 10
I 170  L 170
k 50   L 10
k 130  L 130
*
```

A4.3.2 Amplitude contour specification

```
*A control file for GAIN.PAS 22.1.87 RDW **** svc ****
*stop + vowel + continuant
*fast onset at release, slower offset ramp
*begin:
*
*format: amplitude (dB), duration, transition type, interval
-50    0   B            /Begin at -50 dB
-50   40   L 40         /Stay at -50 dB for 40 msec
0     10   J 10         /Jump to 0dB after 10 more msec
0    420   L 420        /Stay at 0dB another 420 msec
-50   30   L 10         /Linear transition back to -50dB
*                  over the next 30 msec, in 10 msec steps
*Total duration of 40+10+420+30 = 500 msec
.....
end of file
.....
```

A4.3.3 Excitation specifications

```
*an EXCITATION control file 18.1.87; ****  VVV  ****
*all voiced
*samplerate:
 10000
*control data:
* fund(Hz) exc_code time(msec) interp_code (interp_param)
* fund & time are integers
* the codes are char: { S ! V ! R ! M } and { B ! J ! L }
* meaning Silence Voiced Random Mixed and Begin Jump Linear
* the interpolation param is the interpolation interval
* (which is not required for interp_code = "B")
 150 V   0 B
 150 V 200 L 100    /stay at 150 Hz for first 200 msec
  75 V 300 L 10     /ramp to 75, linear on a log scale
.............
eof
.............
```

```
*an EXCITATION control file 18.1.87; ****  VVR  ****
*final unvoiced consonant
*samplerate:
 10000
*control data:
* fund(Hz) exc_code time(msec) interp_code (interp_param)
* fund & time are integers
* the codes are char: { S ! V ! R ! M } and { B ! J ! L }
* meaning Silence Voiced Random Mixed and Begin Jump Linear
* the interpolation param is the interpolation interval
* (which is not required for interp_code = "B")
 150 V   0 B
 150 V 200 L 100
  95 V 200 L 10     /gets to 95Hz after 2/3 of log trans to 75
  75 R 100 L 100    /no periodic component, no need to interp.
.............
eof
.............
```


A4.4 Batch Commands for Synthesis

A4.4.1 The BIGFAAF command

Synthesize an entire wordlist.
Runtime parameters: synthesizer type, interpolation type.

```
REM BIGFAAF.BAT, a file to make the first 85 FAAF words
REM uses batch command synth (which doesn't return here) so:
REM use secondary command processor!
REM
REM %1 is synth type, %2 is interpolation class
REM synth types are S,D,R,A,H for Hump and P
REM interpolation classes pick up a suffix letter on the
REM files of spellings: J=jump, C=cosine, 2=JSRU 2-part linear
REM group 1
command /c synth mail vvv cvc %1 %2
command /c synth bail vvv svc %1 %2
command /c synth nail vvv cvc %1 %2
command /c synth dale vvv svc %1 %2
REM
...
REM group 16
command /c synth mix   vvr cvsc2 %1 %2
command /c synth milks vvr cvsc2 %1 %2
command /c synth mick  vvr cvs2  %1 %2
command /c synth milk  vvr cvs2  %1 %2
REM
etc.
```

A4.4.2 The SYNTH command

Selection of spelling, excitation, and amplitude contour.
Runtime parameters: syntheisizer type, interpolation type.

```
REM batch file SYNTH.BAT name excitation gain 22.01.87
REM name is the MAKE2 inputfile
REM execute MAKE2, TRANS2, GAIN and RESULT2 for FAAF words
REM get segmental data and durations from subdirectory FAAF
copy faaf\%1%5 %1
REM %5 is the interpolation class: J, C, 2 or blank for linear
REM run MAKE3 to look up cf,bw and ampl data
REM and convert to parameters of chosen synthesiser type
REM param %4 is synth type; if blank, use type from input file
make3 %4%1
REM run TRANS3: interpolation of parameters and amplitude
REM and convert back (if necessary) to cf, bw data
trans3 %1<false.txt
```

```
REM GAIN need gains and TRANS3 output from files with
REM the same name, differing only in extension; so copy over:
copy %3.gn %1.gn
REM and apply gains, which add (in dB) to amplitudes.
gain %1
REM RESULT2 command file uses file named 'word'; copy over:
REM first copy the required excitation
copy %2.exb word.exb
REM and then the gain-modified .GNT interpolated parameters
copy %1.gnt word.int
REM now synthesize using S option (in COMMAND) to RESULT2
result2 command
REM put results under proper name back in subdirectory FAAF
copy word.syn faaf\%1.w%4%5
REM %4 keeps various kinds of synth separate.
REM Now clean up debris:
erase %1
erase %1.*
REM that's it
exit
REM exit returns to calling batch, for nesting
```

A4.5 Control File for Making a Wordlist Tape.

A4.5.1 Control file: standard FAAF test

Control file for practice items and start of page one of a standard FAAF test (order A).

```
*This is a control file for SPCH, Faaf list A
*For 1st exp, list A is series synth
*Change the extensions for the various synthesizers.
*
Slevel.wavD=Slevel.wavD     *level setting tones
Sbleep.wavD                 *one bleep = order A
====================        *pause before practice items
Sbleep.wavD=Srot.wvsD       *warning tone, stimulus item
Sbleep.wavD=Sold.wvsD             "            "
Sbleep.wavD=Smessed.wvsD          "            "
Sbleep.wavD=Sstone.wvsD           "            "
Sbleep.wavD=Sscheme.wvsD          "            "
=====                       *pause to turn page
Sbleep.wavD=Sbail.wvsD      *warning tone, stimulus item
Sbleep.wavD=Sbin.wvsD             "            "
Sbleep.wavD=Spost.wvsD            "            "
Sbleep.wavD=Sgab.wvsD             "            "
Sbleep.wavD=Svan.wvsD             "            "
etc.
```


A4.5.2 Control file for naturalness test

Control file for practice items and page one of the naturalness test.

```
*This is a control file for SPCH, Naturalness test
*
Slevel.wavDrrrr          *level setting tones  (r=repeat)

Sbleep.wavDrrrrr         *six bleeps = naturalness test
========                 *pause before practice items
Sb.wavD=Sthan.wvsD=Sb.wavD=Smilk.wvsD=Sb.wavD=Snan.wvsD
                                    =Sb.wavD=Smilks.wvsD

*four practice items, each preceeded by a warning tone
====                                 *pause to turn page

*then 20 items on page one, each preceeded by a warning tone
*in groups of five, mainly for legibility

Sb.wavD=Sthan.wvnD=Sb.wavD=Smilks.wvdD=Sb.wavD=Sthan.wvlD
                    =Sb.wavD=Smix.wvrD=Sb.wavD=Sthan.wvaD

Sb.wavD=Smix.wvpD=Sb.wavD=Sman.wvlD=Sb.wavD=Snan.wvdD=Sb.wavD
                                   =Svan.wvlD=Sb.wavD=Smilk.wvcD

Sb.wavD=Sman.wviD=Sb.wavD=Sman.wvnD=Sb.wavD=Sthan.wvpD
                  =Sb.wavD =Smix.wvnD=Sb.wavD=Smick.wvcD

Sb.wavD=Snan.wvlD=Sb.wavD=Smix.wvcD=Sb.wavD=Sthan.wvcD
                  =Sb.wavD =Snan.wvhD=Sb.wavD=Sman.wvdD

====                                 *pause to turn page

etc.
```

A4.6 FAAF Test Questionairre and Instructions.

A4.6.1 FAAF Tests Questionaire

All supplied information will be kept completely confidential and will NOT be stored on a computer. Age _____

Are you currently in good health? _____

Do you consider your hearing to be normal? _____
(If not, supply details overleaf)

Do you have a cold? ___ Blocked nose? ___ Ear infection? ___
(If YES, supply details overleaf)

Have you ever had a hearing test? _____ Result? _____

Are you subject to dizziness? _____ Tinnitus? _____
(If YES, supply details overleaf)

Is there any reason you know of for not participating in a one-hour listening test? ____ (If YES, give details overleaf)

A4.6.2 INSTRUCTIONS FOR FAAF TESTS

When you are ready, I will start the tape which you will hear through headphones. On the tape is a voice which has been distorted. The aim of the test is to find out how easy or difficult it is to identify what the voice is saying.

The voice will say a word, and you have to identify the word by choosing from the four possibilities given on the response sheet. Each word will be preceeded by a warning tone. When you hear the word, decide which of the alternatives on the response sheet is closest to what you heard, and circle that word. If you are unsure, still indicate what you consider to be the best guess.

There will be five practice words, and then 80 test words divided into groups of 20. There is a separate response page for each group of 20, and extra time is allowed for turning the page between groups.

The whole test will take about 10 minutes. But there will be five tests in all, with short breaks between tests, and a longer break after the first three tests.

Please keep alert. You will be paid a minimum of £3 for participating, but there is a reward of 2p for each correct answer. You can earn a maximum of £8 for a perfect score.

A4.7 Analysis of Variance

A4.7.1 Two-way analysis of variance:

Columns i=1...nj; Rows j=1...k;

Total sum of squared deviations:

$A = \Sigma(x-\mu)^2 = \Sigma x^2 - IJ\mu^2 = 85031 - 25*57.72^2 = 85031 - 83290$
$= 1741$

correction factor $= K = IJ\mu^2 = (\Sigma x)^2/IJ$
(Thus $A = \Sigma x^2 - K = \Sigma x^2 - (\Sigma x)^2/IJ = 85031 - 1443^2/25 = 1741$ also)

among rows:
$B = J[\Sigma\mu i^2 - I\mu^2] = \Sigma(Xi.)^2/I - K = 417455/5 - K$
$= 83491 - 83290 = 201$

among columns:
$C = I[\Sigma\mu j^2 - J\mu^2] = \Sigma(X.j)^2/J - K = 423247/5 - K$
$= 84649.4 - 83290 = 1359.4$

$D = A - B - C = 1741 - 201 - 1359.4 = 180.6$

D/σ^2 is chi-square with (I-1)(J-1) dof

H1: row mean = bulk mean; B/σ^2 is chi-square with (I-1) dof
[F dist; (I-1) and (I-1)(J-1) dof]
$F1 = (B/(I-1)\sigma^2) / (D/(I-1)(J-1)\sigma^2) = (J-1)B/D$
$= 4*201/180.6 = 4.45$
$0.01 < p(H1) < 0.025$ *

H2: column mean = bulk mean; C/σ^2 is chi-square with (J-1) dof
[F dist; (J-1) and (I-1)(J-1) dof]
$F2 = (I-1)C/D = 4*1359.4/180.6 = 30.1$ $p(H2) < 0.005$ **

A4.7.2 Latin Squares Analysis of Variance: order effect.

I=J=r

$A = \Sigma x^2 - r^2\mu^2 = \Sigma x^2 - K = 1741$

$B = r[\Sigma \mu i^2 - r\mu^2] = \Sigma(Xi.)^2/r - K = 201$

$C = r[\Sigma \mu j^2 - r\mu^2] = \Sigma(X.j)^2/r - K = 1359.4$

$D = r[\Sigma \mu t^2 - r\mu^2] = \Sigma(Xt)^2/r - K = 416931/5 - K$
$= 83386.2 - 83290 = 96.2$

$E = A - B - C - D = \Sigma(x - \mu i - \mu j - \mu t + 2\mu)^2$
$= 1741 - 201 - 1359.4 - 96.2 = 84.4$

E/σ^2 is chi-square with $(r-1)(r-2)$ dof.

$F1 = (B/(r-1)\sigma i^2) / (E/(r-1)(r-1)\sigma^2) = (r-2)B/E$
$= 3*201/84.4 = 7.15;$ $p<0.005$ **

$F2 = (C/(r-1)\sigma j^2) / (E/(r-1)(r-1)\sigma^2) = (r-2)C/E$
$= 3*1359.4/84.4 = 48.3;$ $p<<0.005$ ***

$F3 = (D/(r-1)\sigma t^2) / (E/(r-1)(r-1)\sigma^2) = (r-2)D/E$
$= 3*96.2/84.4 = 3.41;$ $0.025 < p < 0.05$ *

all with $(r-1)$ and $(r-1)(r-2)$ dof.

Appendix 5: Naturalness test details

A5.1 Subject instructions for naturalness test

INSTRUCTIONS

When you are ready, I will start the tape which you will hear through headphones. On the tape is a voice which has been distorted. The aim of the test is to find out how easy or difficult it is to identify what the voice is saying. We also want to ask you how natural the voice sounds.

The voice will say a word, and you have to identify the word by choosing from the four possibilities given on the response sheet. When you hear the word, decide which of the alternatives on the response sheet is closest to what you heard, and circle that word. If you are unsure, still indicate what you consider to be the best guess. Then also circle a number to show how natural the voice was.

The naturalness is scored on a scale from one to ten. A score of one is for a voice which is completely natural. A score of ten is for a voice which is totally unnatural. Please try to use the whole range of the scales.

There will be four practice words, and then 80 test words divided into groups of 20. There is a separate response page for each group of 20, and extra time is allowed for turning the page between groups.

The whole test will take about 10 minutes. Please keep alert. There is a reward of 2p for each correct answer.

A5.2 FAAF words: rank order of recognition difficulty

Synthesis Difficulty Rank Order
(and rankings on natural speech in noise as measured by Institute of Hearing Research, Nottingham)

0 = most difficult item

	word ranking 1-80		group ranking 1-20	
word	synthetic	IHR	synthetic	IHR
mail	69	34	15	14
bail	36	58		
nail	57	48		
dale	25	45		
bin	43	59	5	18
din	39	37		
pin	29	39		
tin	9	78		
boast	63	35	2	6
ghost	13	63		
coast	15	24		
post	2	9		
dab	29	53	10	15
gab	29	47		
tab	50	57		
cab	22	36		
man	57	28	1	1
nan	2	8		
van	1	18		
than	11	5		
taught	67	65	14	2
port	25	25		
thought	29	6		
fought	43	3		
seal	8	77	6	8
feel	55	4		
veal	43	62		
zeal	29	43		
sheen	21	69	13	20
seen	64	79		
keen	50	52		
teen	16	80		
mash	69	19	19	16
mass	67	67		
match	50	61		
mats	57	56		
ritz	43	75	9	10
rich	21	55		
ridge	25	17		
rids	36	16		
some	10	2	7	4
sun	76	12		
sub	39	27		
sud	12	70		

rib	57	14	4	5
rig	43	11		
rip	5	23		
rick	18	32		
cob	69	31	3	9
cod	50	13		
cop	4	38		
cot	6	64		
bag	57	73	8	11
bad	50	10		
bat	6	72		
back	39	29		
get	57	26	17	13
bet	25	40		
wet	69	46		
yet	69	68		
mix	77	71	20	19
milks	74	49		
mick	18	51		
milk	77	50		
rose	77	76	11	3
rove	16	7		
robe	29	1		
rode	13	21		
lands	22	44	18	17
lads	75	74		
lad	65	42		
land	65	41		
bang	43	33	12	12
ban	56	20		
bag	18	60		
bad	29	66		
ham	43	22	16	7
high	77	15		
hang	36	54		
how	39	30		

A5.3 Contingency tables for hard/easy and correct/false

natural speech		hard	easy
	correct	2.08	2.29
	false	2.88	5.00
LPC		hard	easy
	correct	4.76	2.19
	false	4.19	4.50
Series cosine		hard	easy
	correct	5.36	6.17
	false	6.04	6.12
Series linear		hard	easy
	correct	7.00	5.95
	false	5.50	6.84
Series lin, deemph		hard	easy
	correct	5.38	6.23
	false	4.92	6.56
Direct form		hard	easy
	correct	5.38	6.25
	false	5.38	6.80
Reflections		hard	easy
	correct	5.81	7.09
	false	5.71	6.72
Area function		hard	easy
	correct	5.12	6.87
	false	5.71	6.59
Articulatory		hard	easy
	correct	5.00	6.32
	false	5.92	7.78
Parallel		hard	easy
	correct	6.38	6.22
	false	6.25	7.00

A5.4 Word pairs for significance tests on recognition scores (Table 5.4)

pair	recognition rate
1. van	29
than	31
2. nan	39
mick	42
3. mix	65
milks	67
4. man	73
milk	80

A5.5 Significance test for recognition scores

wordpair	speech material natural speech Y1	reference series linear Y2	difference d = Y2 − Y1
1	12	7	5
2	10	5	5
3	20	11	9
4	20	12	8

$\Sigma d = 27$

$\Sigma d^2 = 195$

$D = \text{ave } d = 6.75$

$s^2 = (\Sigma d^2 - [\Sigma d]^2/n)/(n-1)$

$= (195 - (729/4)) / 3 = (195 - 182.25) / 3 = 4.25$

$s^2/n = 4.25/4 = 1.06$

$s = \sqrt{(s^2/n)} = 1.03$

$t = D/s = 6.75/1.03 = 6.55$, dof=3.

two-tailed test for difference from mean of zero: $p<0.005$

A5.6 Significance estimates for naturalness scores (Tables 5.6, 5.7 and 5.8)

For the recognition scores in Table 5.4 the mean (of the wordpair scores) was compared with the mean for series linear synthesis. Recognition performance on series linear was used as a benchmark. But there are no other naturalness results, so no reason to select one particular type of synthesis as a reference. The t-test is for a change or difference in mean value, and so requires a reference.

A simple approximate procedure is to treat the observed mean and standard deviation of results on the eight types of synthesis as though they were in fact the true parameters of an underlying distribution, rather than estimates based on a certain number of observations. Then one can simply treat each score as a certain distance from the true mean, normalised by the standard deviation. A variable with a normal distribution of zero mean and unit standard deviation is usually denoted Z. Conversion of samples from any other normal distribution to Z-scores is a matter of subtracting the mean and dividing by the standard deviation. The probability of a sample larger than a particular value of Z is the information given in a table of the normal distribution.

The observed naturalness scores in Tables 5.6, 5.7 and 5.8 were converted to Z-scores using the mean and standard deviation computed for the each of the eight types of synthesis. The results are the values labelled 'synthesis mean, sd' in these tables. Conversion to Z-scores allows the probability of the observed disparity from the 'synthesis mean' to be taken from a table of the normal distribution, and this probability is the value given as a significance score in Tables 5.6, 5.7 and 5.8. It would be a true significance if the scores arose from a normal distribution, and the true mean and standard deviation were known. In the present case normality is assumed, and no allowance is made for the fact that the parameters of the distribution are estimates.

Not treating the parameters as estimates reduces the probability scores. A proper t-test with eight degrees of freedom needs a Z of about 2.3 to reach the 5% significance level. Simply using the normal distribution as just described requires a Z of only about 2.0. Put another way, the method used reaches 5% significance with about a 15% smaller disparity from the mean. Thus the resultant significance scores are slightly generous.

Speech Audiometry

Edited by

Michael Martin, OBE

The Royal National Institute for the Deaf, London

Taylor & Francis
London New York Philadelphia
1987

Chapter 1

Basic Properties of Speech

R. Wright

Many areas of research and clinical practice involve some aspect of speech, and many people are thus expected to have some knowledge of the subject. But what kind of knowledge? Linguistics, phonetics, psychology, acoustics, signal processing, physiology, communication engineering: there is no end to it. It is tempting to assume that one's own innate knowledge will suffice, at least for clinical subjects like audiology. After all, most of us are (given normal hearing and a few other favours) expert in the use of speech; is that not sufficient?

It is not enough when dealing clinically with hearing loss, because we are then intervening in a process which has gone wrong. No mechanical knowledge may be required to drive a car, but we should like rather more when faced with a breakdown. One useful aspect of the multidisciplinary nature of speech studies is that quite a lot of information is available at the introductory level. There are excellent texts to introduce linguistics, phonetics, the physics of sound and of speech, and the basics of psychoacoustics and speech perception.

This chapter will refer to all the above areas at an introductory level, but the reader is referred to Denes and Pinson (1963), Fry (1979), Gimson (1980), Ladefoged (1962, 1982) and Moore (1982) for a more complete introduction. None requires a degree in maths or other prerequisites. They all take the time and space to provide a proper foundation, whereas this chapter is only a summary, a list of findings.

Given a basic understanding at the introductory level, this chapter will try to emphasize two particular aspects of speech:

1. the importance of viewing speech as something very different from a sequence of sounds. It is usual to introduce phonetics with a description of 'the sounds'; this approach leads to many problems. It encourages a view that speech can be adequately described at a single level of analysis. This in turn makes it awkward to introduce linguistics and a heirarchy of descriptive levels. It is then equally awkward to describe prosodic aspects

such as stress and intonation. They tend to be introduced as something added to speech more or less as an optional extra or afterthought. This chapter will go very much to the opposite extreme, and will try to cover almost all of speech perception without recourse to a segmental description.

2. the importance of decision making. Speaking and hearing are often treated as opposite ends of a 'speech chain'; articulatory phonetics is often viewed as opposed to acoustic phonetics; but the speaker and hearer are joined in the task of making a set of decisions concerning units at various linguistic levels. Speech perception will therefore be presented in terms of a set of yes/no (or at worst three-way) decisions, and the acoustic characteristics which encode the speaker's decisions and provide the cues to the listener's decisions.

This approach is not meant to be idiosyncratic. Hearing impairment only becomes a problem for speech when decisions about the speech signal begin to be made incorrectly. By and large, many segmental decisions can be incomplete before communication begins to be affected (because they were not necessary for correct decisions at higher levels). This process can only begin to be adequately explained by consideration of a heirarchy of suprasegmental aspects of the decision making involved in speech perception. The point of speech audiometry is to provide a methodology which tests a person's ability to make these decisions.

The Speech Signal

Periodic and Aperiodic Waveforms

The physical description of speech usually begins with a consideration of waveforms (Figure 1). Speech at this level can be considered as simply a disturbance of the air pressure, a sound wave.

The first distinction to be made concerning waves is whether or not they repeat. A repetitive wave is termed *periodic*, as shown in Figure 1(a). A non-repetitive or *aperiodic* wave is shown in Figure 1(b).

Many interesting physical phenomena have a repetitive aspect, including the idealized vibration of the larynx when used for speech. The resultant periodic waves are essentially different from aperiodic waves, because they are completely described by one repetition (one period). Thus Figure 1(a) is a complete description whereas for the aperiodic wave, 1(b) the figure is only a portion, an incomplete description.

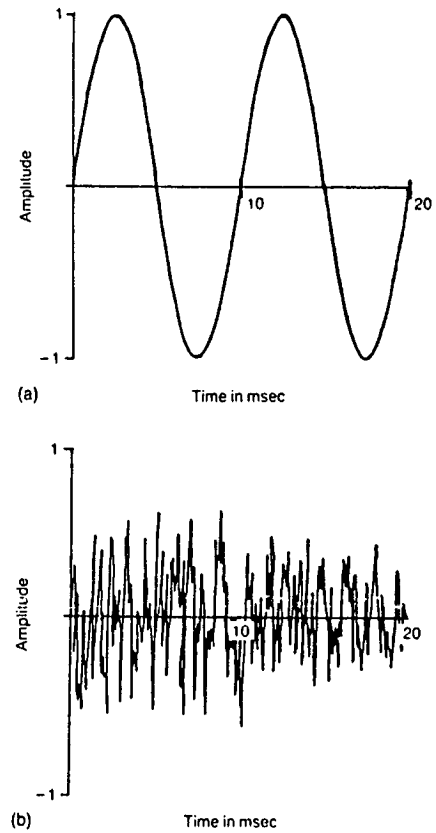


Figure 1. Periodic and Aperiodic Waveforms: (a) An example of a periodic wave, a 100 Hz pure tone; (b) An aperiodic waveform, random noise.

Properties of a Periodic Wave

A periodic wave is characterized by three properties:

(a) the period, which is the repetition interval.
(b) the amplitude, which is simply the height for the pure tone in Figure 1(a). (An exact definition of amplitude can be made for any periodic wave, however complicated the shape.)
(c) the wave shape (wave form).

The pure tone in Figure 2 has a period of 1 msec (0.001 second). Thus it repeats 1000 times per second and has a *fundamental frequency* of 1 kHz. The amplitude shown represents the quietest sound that the normal ear can hear at this frequency. In this case the amplitude has units of length, representing the actual amount of motion of notional small volumes of air. It should be noted that this motion is very small: it is of the order of the diameter of an air molecule!

It is more usual to give pressure rather than amplitude, because pressure can be directly measured. The equivalent pressure is also shown in Figure 2.

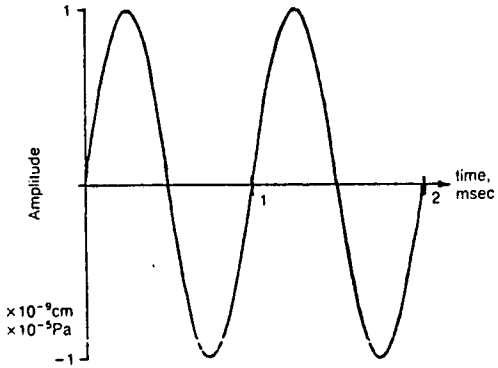


Figure 2. A pure tone with a frequency of 1000 Hz, at a level corresponding to the detection threshold for normal hearing.

Wave Shape and Spectrum

The key to formally characterizing a wave form is presented in Figure 3, showing how a complicated shape is equivalent to a combination of pure tones, varying in period and amplitude. This principle was developed by the French mathematician (and Utopian socialist) Fourier, ca 1800, and is called *Fourier analysis*.

Fourier analysis shows that ANY periodic waveform can be represented as a combination of pure tones. Further, the frequencies of the pure tones to be used are specified: it is sufficient merely to use those tones whose frequencies are integer multiples of the fundamental frequency. These tones constitute a *Fourier series*. The wave shape can be exactly specified in terms of the amplitudes and phases of a set of pure tones. The lowest frequency tone (with period equal to that of the complicated waveform) is the *fundamental* of the Fourier series. The remaining tones are the *harmonics*, beginning with the second harmonic. This method of representation of waveforms is also called harmonic analysis.

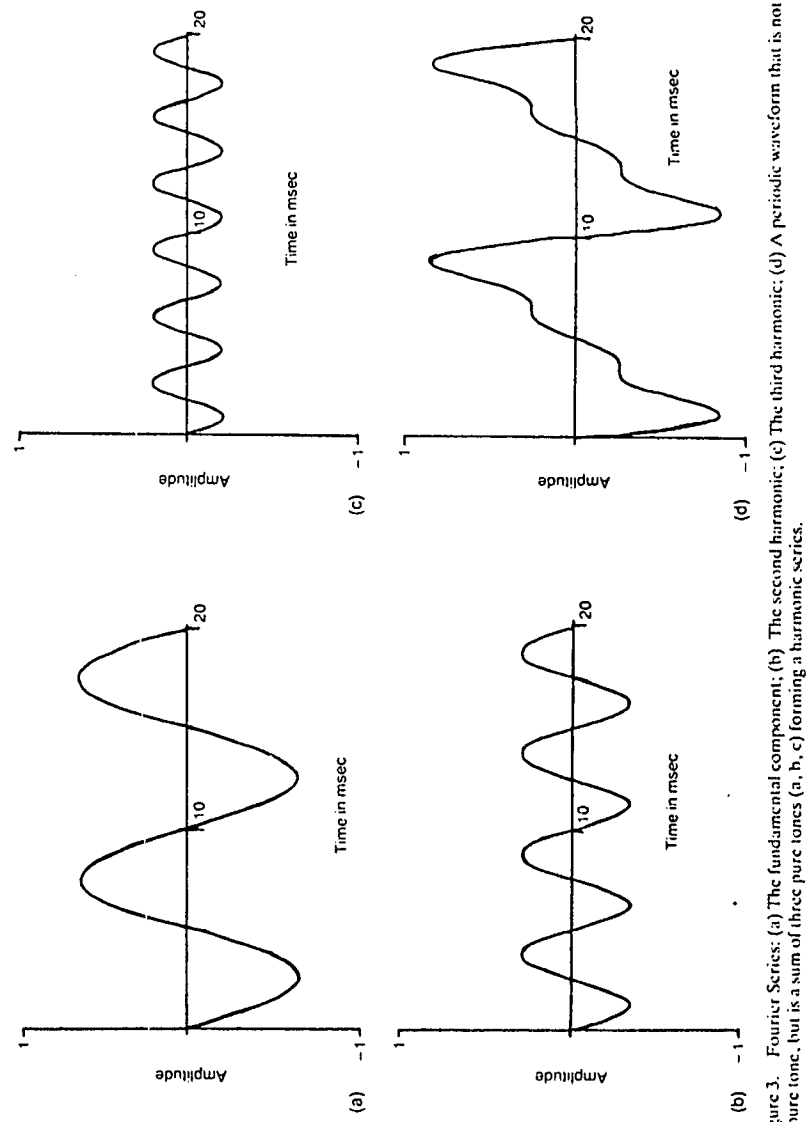


Figure 3. Fourier Series: (a) The fundamental component; (b) The second harmonic; (c) The third harmonic; (d) A periodic waveform that is not a pure tone, but is a sum of three pure tones (a, b, c) forming a harmonic series.

A bar graph can be constructed showing the amplitude of each harmonic of the series. Figure 4. This shows the same information as in Figure 3, but in a compact way (especially if many more harmonics were to be used). This graph is a *line spectrum* or *discrete spectrum*.

Knowledge of the period, the overall amplitude, and the spectrum completely characterize a periodic signal. Further, the period is evident in the spectrum, and the overall amplitude is defined in terms of the individual amplitudes which constitute the spectrum. Thus the spectrum is a complete description of a periodic waveform.

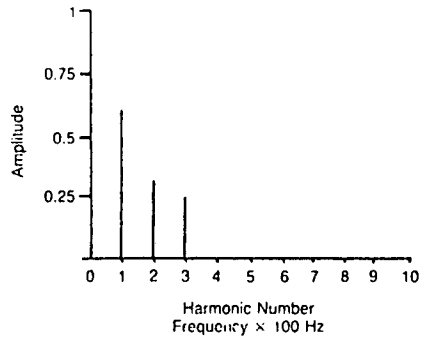


Figure 4. The line spectrum for the waveform in Figure 3(d).

Aperiodic Waveforms

Fourier analysis can be extended to non-repetitive signals. A non-mathematical interpretation would be to consider such signals as actually having a period, but one of duration approaching infinity. Thus the fundamental frequency becomes very low (approaching zero), and the lines in the spectrum get very close together. The result is a *continuous spectrum*, a spectrum which has energy (or could have energy; some frequency regions may make no contribution) at all frequencies. Because it is impossible to draw an infinite number of vertical bars, it is conventional to just draw the tops of the bars. Figure 5 shows an aperiodic sound and its continuous spectrum.

It is worth emphasizing that sounds which occur at very low repetition rates can have very high frequency content. Thus although the normal ear is said to reach a lower frequency limit at about 20 Hz, this does not mean that one must slam a door at a frequency greater than 20 Hz in order to be heard. A person may slam a particular door once per day. This is not a very low frequency (one cycle/day) sound, because it is not a pure tone. It can be analysed as an aperiodic signal with a continuous spectrum, and with significant energy at audible frequencies.

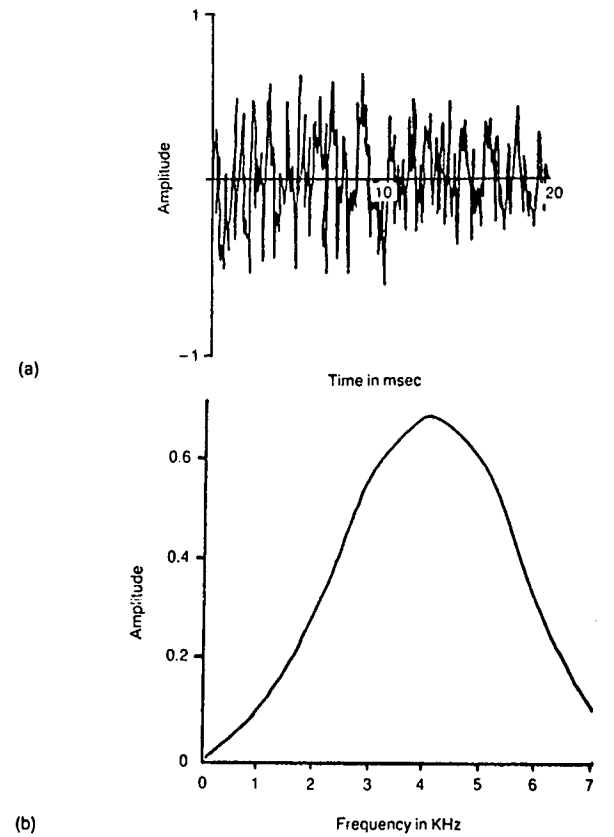


Figure 5. Aperiodic signals: (a) A noise waveform; (b) The continuous spectrum of the noise waveform.

Duration

The fact that a waveform can be described in terms of amplitude, period, and spectrum has already been discussed. One more dimension is necessary: time. We wish to deal with speech signals, and so must describe signals which are of finite duration: they start and stop.

This involves a compromise with the definition of a periodic signal, and hence with the requirements of Fourier analysis. Speech signals will have a repetitive aspect, but will not be perfectly repetitive because of their finite

duration. Thus making a line spectrum from one period of a speech signal is an approximation: a very good approximation for a sustained sound with many similar periods: a poor approximation for a rapidly changing sound which is really not at all repetitive.

Given the above reservation concerning periodic signals, we can physically describe speech signals in terms of their amplitude, period, spectrum and duration. The duration is simply the time from the beginning to the end of the signal involved. If the signal changes over the course of its duration, then properly we need multiple values of amplitude, period and spectrum. A three-dimensional graph of spectrum vs time could be produced (amplitude vs frequency vs time) and this is exactly the information in a speech spectrogram, the basic tool of acoustic phonetics research (Figure 6).

Figure 6. A speech spectrogram, a three-dimensional representation of speech. Frequency is the vertical dimension, time is horizontal, and the signal level is represented by the amount of darkness.

Spectrogram courtesy IBM (UK) Science Centre, Winchester.

Psychoacoustics

The human observer is not just like a microphone and a Fourier analysis system. Human response on any physical dimension does not in general uniformly equate to the physical units which describe that dimension. This is the whole subject of psychophysical scaling, and deserves separate study in its own right. We can only summarize the most relevant results.

For each of the physical descriptions so far discussed (duration, amplitude, period and spectrum) we can present an equivalent perceptual description, Table I.

Perceptually speech sounds can thus be described in the following terms:

LOUDNESS. There is a proportionality effect to the perception of loudness (and pitch; and many other sensory phenomena); small changes to small signals are as significant as much larger changes to large signals. Thus a

Physical	*Perceptual*
Amplitude	Loudness
Period	Pitch
Spectrum	Quality
Duration	Length

Table I: Physical (Acoustic) and Perceptual (Auditory) Descriptions

change in sensation depends upon the *ratio* of the stimulus change to the size of the original stimulus. This is not at all the same as a *linear* response. In a linear system the sensation change (or measurement) depends only upon the stimulus change, and no ratio is involved. Most physical devices (like microphones) are built to have a linear response, and hence differ in a basic way from human auditory perception.

The decibel scale is a way of numerically coping with the wide range of sound levels. Human auditory processing is faced with the same problem, and also solves it by the use of a logarithmic relationship. Thus the decibel scale is an approximation to loudness, the auditory scaling of acoustic intensity.

So decibel steps should represent loudness steps. Indeed, one decibel is approximately the minimum detectable loudness change over a wide range of intensities and frequencies. Although decibels are still a physical measurement, a logarithmic conversion from linear physical measurements, the decibel uses a mathematical relation (the logarithm) which is a good approximation to the perceptual scaling for loudness.

A common problem in hearing impairment is the phenomenon of recruitment, in which a person is abnormally sensitive to changes in sound level. This can be viewed as an alteration to the proportionality constant in the logarithmic scaling such that loudness increases faster than for the normal ear.

PITCH. A periodic sound will produce a sensation of pitch. (So will other sounds; this is a complex subject and the reader is referred to Moore, 1982). Pitch is determined mainly by the repetition period.

It is a commonplace error to confuse fundamental frequency with the *amplitude* of the fundamental component in the line spectrum. This leads to the conclusion that if the amplitude of this component is reduced to zero, then the fundamental frequency is eliminated. People then marvel at the power of human perception in 'recovering' the 'missing' fundamental frequency (as in telephone bandwidth speech).

Simple inspection of the time waveforms can help clarify the issue. Figure 7(a) shows a signal and its spectrum; Figure 7(b) shows the same signal except the amplitude of the fundamental component is zero. As is evident from the figure, the obvious period is unaltered. Changing the amplitudes in the spectrum will change the quality, not the period. There is no missing

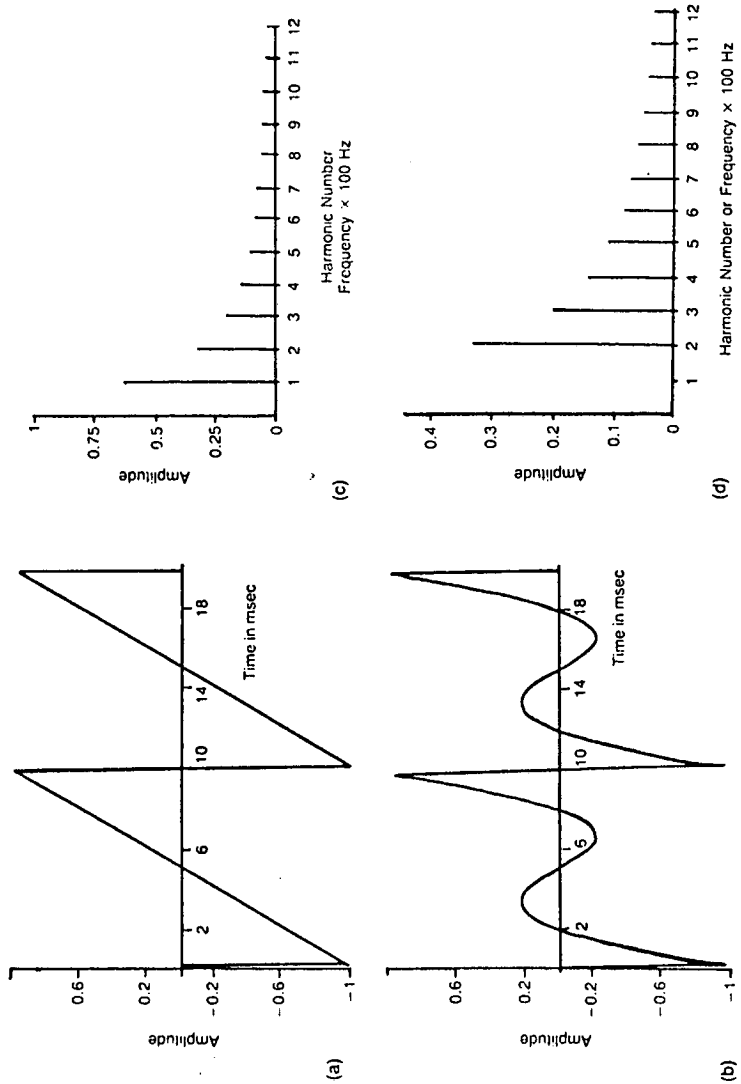


Figure 7. The 'Missing' Fundamental: (a) A sawtooth wave of 100 Hz; (b) The result of deleting the fundamental component. The period is unaffected; (c) The spectrum of (a), consisting of the fundamental and harmonics; (d) The spectrum of (b). The greatest common divisor of the harmonics is still the 'missing' fundamental.

fundamental as far as the eye (or the ear) is concerned: the period of Figure 7(b) is as evident as that of Figure 7(a). Further, even in the spectrum the fundamental frequency is still evident, as the harmonic spacing. In practice, as few as three higher harmonics (such as 17, 18 and 19 in Figure 8) are sufficient to produce a clearly evident periodicity (again, clear to the eye in the figure, and clear to the ear as an actual signal).

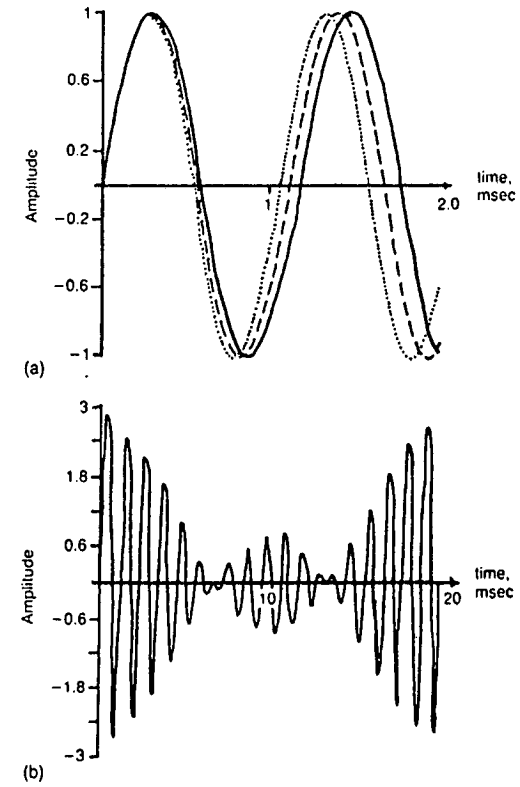


Figure 8. An extreme case of periodicity determined solely by higher harmonics: (a) The first cycles of pure tones at 850, 900 and 950 Hz, which are harmonics 17, 18 and 19 of a series beginning at 50 Hz; (b) The sum of the three very high harmonics, with obvious periodicity at 20 msec corresponding to 50 Hz.

The basis of pitch in periodicity (NOT in strength of the fundamental of the spectrum) should be borne in mind in any consideration of pitch perception through either an impaired ear or a band-limiting device.

As with loudness, pitch relations are on approximately a log scale. Thus uniform intervals in pitch relate to uniform ratios in frequency. In music this leads to units such as semitones and octaves, which do not represent a fixed step size in physical terms, but rather represent a fixed ratio. Thus an octave is a doubling in frequency. And an octave rise in pitch can occupy a small (from 50 Hz to 100 Hz) or large (300 Hz to 600 Hz) frequency range. Both of these one-octave pitch rises would be perceived as being of approximately the same 'size'.

QUALITY. There is a difference between the sustained sounds produced by two different musical instruments playing the same note (i.e., same pitch, loudness and duration). In music this difference is called timbre but in speech it is called quality. The quality difference depends upon the spectrum, or equivalently the waveform.

For speech sounds, two vowels can be matched for pitch, loudness and duration, but their phonemic category (identity) will be determined by their quality.

LENGTH. Perception of length is not so regular as for pitch and loudness. There is not an accepted psychophysical scale in general use in audiology or phonetics, primarily because it is often unclear in speech just where any particular unit begin and ends (either perceptually or physically). Phoneticians use 'short' and 'long' to represent phonological contrasts (especially for vowels) which may or may not relate to perception of length, and may not be reducable to physical dimensions.

Linguistics

Speech is used for communication between persons. It is this role which is considered at the linguistic level.

Communication proceeds by the encoding (at the source) and the decoding (at the receiver) of information. The encoding and decoding consist of specific yes/no decisions operating upon various units which constitute a *linguistic hierarchy*. As with psychophysics and phonetics, linguistics is a major field of study in its own right.

Speech can be analysed for decision-making purposes according to the following (simplified!) linguistic hierarchy, summarized in Table II:

The UTTERANCE. An utterance constitutes the largest unit of *syntax*, which is the system of word-arrangement constraints.

The PHRASE (tone group). Within an utterance, divisions can be made into the major syntactic constituents. Acoustically the divisions may be marked with pauses or changes in fundamental frequency, or may be determined purely by syntactic structure. A defining characteristic of a tone group is that it has within it one main pitch change, called the *nucleus* of the *intonation*

Linguistic Unit	*Decision System*
Utterance	Syntax
Phrase	Intonation
Foot	Stress
Syllable	Vocalic vs Consonantal
Syllable-Part	Manner
Segment	Place

Table II: Hierarchy of linguistic units and their associated speech contrasts.

pattern. As soon as we attempt to divide an utterance into any smaller information-bearing units we must have pitch information, operating within an intonation system. The first decisions made about an utterance (first in terms of the size of the unit which the decision governs) are decisions based upon pitch perception.

The FOOT (stress group). The next level down requires the identification of stressed syllables, or at the very least the unit of 'stressed and following unstressed syllables', called the foot. Only a stressed syllable can carry an intonation marker (a pitch change), and thus only a stressed syllable can be the nucleus of a tone group. In English, an unstressed syllable MAY (not must; some do not) reduce, and a reduced syllable is shorter in length than an unreduced syllable. There is often also a difference in quality: the vowel in a reduced syllable 'reduces' to /I/ or /ɔ/.

Thus we see that just to divide an utterance into stress groups requires all of the perceptual dimensions except loudness, although the use of quality may not be necessary for the determination of stress. Rather it should be viewed as a correlate, with duration as the determiner. Similarly, loudness will also be affected (modulated) by stress and intonation patterns, as a secondary effect. Interestingly, when periodicity cues are removed (in whispered speech) intonation patterns can still be conveyed through the use of the secondary cue (correlate) of loudness; duration as a cue to stress is unaffected by presence or absence of periodicity.

The SYLLABLE. The foot is defined as a stressed syllable followed by any number of unstressed syllables. We have indicated that stress on a syllable is determined principally by duration. How is the syllable defined?

A syllable must have one and only one *syllable centre*, a sound with a vocalic role (a vocoid). Thus a foot divides into syllables, each of which has a centre (a vowel sound or another sound with a vowel role); then anything left over must attach to one or the other (preceeding or following in time) adjacent centres as a *syllable margin*. A part of the speech signal 'attached' to a following centre is thus *syllable initial*; otherwise it becomes *syllable final*.

The SYLLABLE-PART. Conventionally in English a syllable centre is a vowel, and an initial or final syllable margin is a consonant or consonant cluster. Additionally certain consonants may extend their roles to also serve as syllable centres. Thus 'syllabic' nasals or /r/ or /l/ are consonantal segments with a role (one level up) as syllable centres.

In English a syllable centre is either a single vowel, diphthong or syllabic consonant. A syllable margin, however, can consist of from zero to three consonants (and very rarely four as in 'twelfths').

It is possible to describe syllable parts without reference to vowels and consonants. From the decision making point of view a syllable centre requires a certain set of decisions about quality, and the margins require different decisions about 'manner'. Only one yes/no decision about each manner type may be 'loaded' onto a given syllable margin. This fact can be used to divide margins and thus divide syllables. This approach will be discussed in detail in the acoustic phonetics section.

The SEGMENT. The lowest-level unit is the individual speech sound. Decisions about speech at this level are *phonemic*; thus a phoneme is the minimum information bearing unit of the speech signal. It is only at this very lowest level that we encounter the units naïvely thought of as constituting speech. A wealth of interpretation must be accomplished before decisions at the segmental level are reached.

Of course, if only the stylized enunciation of isolated monosyllabic words is considered (as in some types of speech audiometry) then most of the higher levels are completely eliminated. We should at least be aware of what is being thrown away.

Acoustic Phonetics

This section will cover the description of speech (in particular, spoken English) in terms of the acoustic consequences of speech production, and the acoustic cues to speech perception. A complete description of speech would cover the many overall characteristics (voice quality, pitch range, rate; speaker sex, class, dialect, nationality) which must be present but do not encode a message. Thus there are overall features and contrastive features. The contrastive features carry the information, and require the decisions which form production and perception. The contrastive features will be the main concern of this discussion. This section will attempt to cover every speech decision (from top to bottom) and the associated acoustic cues.

As mentioned earlier, it is usual in introductory phonetics to concentrate on speech sound contrasts. Such an approach undervalues the linguistic complexity of speech, and leaves out decisions about syllable structure, stress and intonation. These higher linguistic levels also have contrasts, and are

referred to collectively as prosodics or suprasegmentals. To give these aspects their due, this discussion will begin with prosodics and only arrive at segmentals after all the higher levels of decision making have been described.

Prosodic Aspects: Pitch, Length, Loudness

The last section endeavored to show how (at least for English) the whole shape of an utterance and all the perceptual decisions down to the level of the syllable are prosodic, meaning based on pitch and length and perhaps loudness, and not based on spectral quality or anything to do with individual speech sounds.

The first decision about an utterance is the division into tone groups, each with its major pitch change on the nucleus. For simple utterances there will be just one tone group, and one nucleus. The role of the pitch change to mark the nucleus is especially significant, as this will be the most important word in the utterance.

Pitch is primarily determined by the periodicity of the speech signal. In normal speech, median values for fundamental frequency range from about 120 Hz for adult male voices to 180 Hz for adult females and roughly up to 250 Hz for children. Variation about this median in ordinary speech does not usually span much more than an octave, and variation tends to be toward higher pitches. Thus the median is usually near the low end of a person's range. Most people are actually capable of producing nearly a two octave range of fundamental frequencies, but for speech they tend to use about an octave located (again) at the low end of their range of possibilities. Within a single utterance a range of half an octave is quite usual, though the variation can be anything from nearly zero (a dull monotone) to an octave or more (extreme emphasis).

The next decision concerns stressed and unstressed syllables. This is an area which has considerable variation from language to language. In English the stress system is rather complex, because two factors are principally involved: pitch and length.

The nucleus is not the only syllable in an utterance which may have a pitch change; other stressed syllables may have a pitch marker (pitch motion). The important distinction is that unstressed syllables can never have a pitch marker.

Furthermore, in English an unstressed syllable may reduce. The vowel duration may be shortened by 50% or more, and the vowel quality may also reduce to /I/ or /ə/.

Thus there are four sorts of syllable:

1 -- stressed plus pitch motion
2 -- stressed
3 -- unstressed but not reduced
4 -- reduced

The clearest distinction is the three-way division between 1, 2 and 3 together, and 4. Type 1 has a pitch motion, type 4 has shortened length, and types 2 and 3 are both of full length but without pitch motion.

Many discussions of this subject founder on treating stress as a question of separating type 2 vs 3. It is much clearer (and more important for speech perception) to begin with the very much easier separation of 1 vs 2 + 3 vs 4. Then type 2 vs type 3 can be put in proper perspective. It is not the first decision to be made regarding stress; rather it is the last. The physical basis of this decision is not well established. According to Ladefoged (1982) it relates to 'effort', but effort can manifest itself in any (or any combination) of the prosodic dimensions.

Typical conversational speech proceeds at 100 to 150 words per minute, or two to three words per second. Thus the syllabic rate is approximately five syllables per second. A stressed syllable marked with a pitch change can typically have a duration of 100 to 200 msec, and could extend to 500 msec or more. A reduced syllable can easily have a duration of less than 50 msec, and in the extreme can reduce to nothing.

Segmental Aspects: Vowel and Consonant Contrasts

At the segmental level information in the speech spectrum (and hence the sound quality) becomes important for the first time. Of particular interest is the change of quality with time, as quality on its own provides limited information.

Vowel Contrasts

A uniform tube (such as an organ pipe) has the property of resonance. Certain frequencies are favoured, others are subject to cancellation. For speech the resonant frequencies (or modes) of the vocal tract are called *formants*. Vowel perception is based mainly on the position in frequency of the first two formants, F1 and F2.

These two formants determine a two-dimensional space, as shown in Figure 9(a), which shows the F1 and F2 values for the vowels of American English. Perceptual studies have shown a close agreement between physical formant measurements and perceptual scaling of vowel similarities, as shown in Figure 9(b). Furthermore, the physical and perceptual data bear a striking similarity to the 'vowel quadrilateral' used in the traditional teaching of phonetics, Figure 9(c). Thus the vowel quadrilateral closely represents speech perception, although it is conventionally labelled in terms of speech production. Further, the perception is closely linked to the acoustic description of the vowels, not the articulatory description. Thus the real success of the vowel quadrilateral (and the cardinal vowels) in phonetics can be explained: it is a good description of perception, and phonetics teaching is mainly a matter of the training of perception.

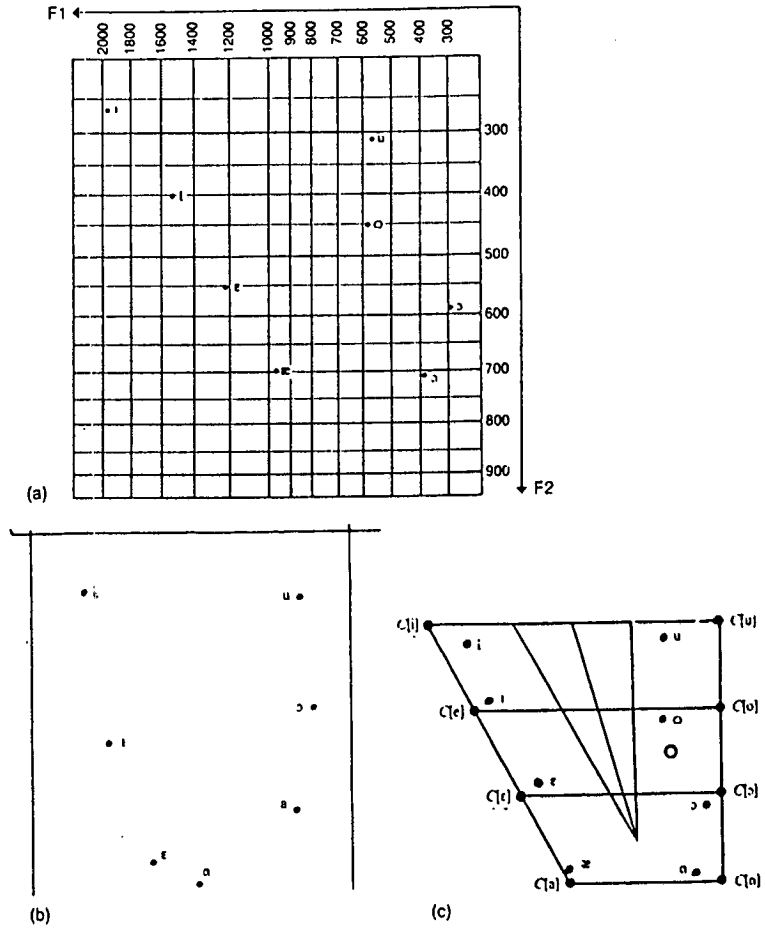


Figure 9. Vowel Space and the Significance of Formants: (a) Acoustic – The representation of the English steady vowels plotted according to the two lowest resonant frequencies of the vocal tract, F1 and F2; (b) Perceptual – The result of judgements of perceptual distances between vowels; (c) Phonetic – The vowel quadrilateral used in descriptive phonetics.

Figures taken from:

(a) Ladefoged (1982), Fig. 8.7; The vowels are for American English.

(b) Klein, Plomp and Pols (1970), Fig. 1.5.1; from a study of Dutch vowels. Their figure has been rotated and reflected about the central axis.

(c) Gimson (1980), Fig. 6; data on American vowels from Ladefoged (1982, Fig. 4.2) have been added.

Both F1 and F2 can range over about two octaves (adult male F1: 200–800 Hz; F2: 700 Hz–2.8 kHz). Formants can be about 40% higher for adult females, and still higher for children. Unfortunately, each formant does not scale uniformly with vocal tract length, so a child's formants will not be a fixed percentage increase on the adult male values quoted. But one would not expect a child's formants to be more than double those given above.

The basic decisions about vowels are:

1. F1 and F2 far apart vs close together ('front' vs 'back' vowels).
2. F1 high vs F1 low ('open' vs 'close' vowels).

Many of the world's languages can use the above decisions to produce a system using from three to five vowels. In English the perception can be modelled as based on five vowels: a 3-way decision on F1 and a 2-way decision on F2-F1, which yields six categories. This collapses to five because F2 is constrained when F1 is high.

But English has many more than five vowels! A second decision is added, based upon length, yielding 5 more vowels (there may well also be quality differences between the 'long' and 'short' members of a pair). Finally, if the formants move slowly diphthongs are produced, of which Southern British English has a particularly large number. Interested readers are referred to Gimson (1980) and Ladefoged (1982) for further details.

In Southern British English, perception is aided by constraints imposed upon vowels. Syllables can be divided into open (no syllable-final consonants) vs closed (ending in a consonant or cluster of consonants). ONLY long vowels can occur in open stressed syllables. In unstressed open syllables the reduced forms /ə/ and /ɪ/ also occur. Thus whether a vowel is heard as long is governed partly by the syllable structure. If it is an open syllable it will be a long vowel, regardless of the actual acoustic properties of the vowel itself. This is one example of the many ways in which the sounds of speech are really a question of larger units (the syllable) and the whole system of contrasts.

Consonantal Contrasts

Consonants are by definition those sounds occurring in the syllable margins. It is conventional to divide consonants according to place and manner (of articulation) and voicing. This section will extend this description to the acoustic consequences of the place, manner and voicing taxonomy.

Voicing principally refers in articulatory terms to laryngeal activity, the presence or absence of vibration of the vocal folds. The acoustic consequence is ideally a periodic or aperiodic waveform. The actual acoustic cues to voicing depend upon the consonants involved (stops or fricatives or affricates), syllable stress, syllable position, and whether or not the consonant is part of a cluster. Manner categories are shown in Table III.

Manner	*Speech Sounds*
Stop	p t k b d g
Fricative	f θ s ʃ v ð z ʒ h
Affricate	tʃ dʒ
Nasal	m n ŋ
Approximant	w j r l

Table III: Manner categories

Place refers to the main place of constriction within the vocal tract, as shown in Table IV. These categories are the main positions for English, working from the front of the mouth to the back. For detailed phonetics more places and multiple articulatory gestures may actually be involved, but those given will suffice for a discussion of the system of contrasts.

Place	*Speech Sounds*	*Examples*
Bilabial	p b m	pie, buy, my
Labiodental	f v	face, vase
Dental	θ ð	thin, then
Alveolar	t d n s z	to, do, new, Sue, zoo
Palato-alveolar	ʃ ʒ tʃ dʒ	shy, measure, church, judge
Velar	k g ŋ	back, bag, bang
Laryngeal	h	hooray, Henry

Table IV: Place categories

Syllable Initial Consonant Contrasts

Decoding of consonants within the syllable-initial margin is considerably simplified by a very strong constraint: only ONE yes/no decision is required for each manner possibility for the whole cluster, and ONE voicing decision for the whole cluster. For instance, a cluster AS A WHOLE is either nasal or non-nasal. There is no possibility of combining nasals into clusters to require two or more decisions about nasality. The same simplification applies to all the manner contrasts, and to voicing.

Thus a syllable initial consonant cluster (as a whole) is either:

1 — nasal or not
2 — fricative or not
3 — stop or not
4 — approximant or not
5 — voiced or voiceless

Further there are usually only one or two and at most three initial consonants, and their sequence is subject to further strict constraints. The result is that,

although there are about 24 consonants, there are nowhere near $24 \times 24 \times 24 = 13{,}824$ possible clusters involving three consonants. In fact there are less than ten (some variation according to dialect); they all begin with /s/, the second consonant is a voiceless stop, the third is an approximant, and nearly half of the stop+approximant pairs are not allowed (/tl/, /kl/, /pw/ and /tw/). Now these five manner and voice decisions can be discussed in turn.

NASAL. The main acoustic cue to nasality is presence of a relatively strong nasal formant at about 300 Hz. The F1 for the related non-nasal articulatory configuration is eliminated, and F2 and higher resonances decay more quickly because of the extra loss through the nasal tract. This extra loss can also be described as a widening of formant bandwidths and a reduction of formant amplitude. (See Ladefoged, 1962 for a discussion of loss and bandwidth.)

FRICATIVE. Acoustic cues to frication, on the other hand, lie in the frequencies above 1 kHz, and spreading as high as 8 to 10 kHz for /s/. Frication is the audible consequence of air turbulence at a constriction. This signal is aperiodic, though it may be added to a periodic signal from the larynx in the case of voiced fricatives. In either case, the presence of random noise above 1 kHz will be the cue to frication.

STOP. A stop has a temporal cue, an abrupt interruption or gap in the signal. This definition is sensible for words within an utterance, but not for the beginning of an utterance. Although no gap as such occurs in this position, the release from vocal tract closure is still cued by the abrupt onset (half a gap!), the rapid energy rise as the syllable begins. Also associated with stops is rapid formant motion as the vocal tract moves from an obstructed to a non-obstructed shape. In particular, F1 will rise. The change in F2, though ultimately of great significance, is related to place, and so can be ignored for the purposes of manner decisions.

An affricate is acoustically a stop with a slow release, so the gap is followed by frication produced at the same 'place' as the closure (homorganic fricative).

APPROXIMANT. An approximant is vowel-like but with slowly varying formants (but not so slow and not with the same pattern of motion as for diphthongs).

VOICING. Voicing in syllable initial position is complicated. There is no voicing contrast for nasals and approximants. Similarly many clusters have no voicing contrast. Fricative + stop is always 'voiceless', regardless of the actual articulation or acoustic manifestation. Similarly fricative + nasal has a 'voiceless' fricative. All that really remains are single stops, and stop + approximant clusters; even here a voicing contrast may not be required except on stressed syllables.

For unconjoined stops and stop + approximant clusters, on stressed

syllables, in syllable-initial position, one must consider acoustic cues to voicing. The cue is generally characterized by *Voice Onset Time (VOT)*, which is the asynchrony between release of the closure and initiation of larynx vibration. In English the sounds /b,d,g/ begin to have laryngeal vibration at about the time the constriction opens (is released), or shortly (less than 25 msec) thereafter. If larynx vibration does not begin until well after release (well over 25 msec) the sound is categorised as a /p,t,k/.

The same holds for stop + approximant, except the VOT may become very long indeed and an entire /r/ or /l/ (for instance) may be produced without laryngeal vibration (as a voiceless fricative). The perception of this difficult contrast (a few milliseconds either way from 25 msec) is aided by concomitant effects: larynx activity produces low frequency (below 1 kHz) spectral content, so the tilt of the spectrum is a cue. Lack of low frequency energy means there is little excitation for F1, so absence (cutback) of F1 is a related cue.

Syllable Final Consonant Contrasts

Syllable final contrasts start off by following the same simplification as for syllable initial: only one manner and voicing decision for the whole cluster. Unfortunately this is then complicated by the affix system in English which can add markers for plural or possessive (/s/, /z/, /ə/) and for past tense and past participle (/t/, /d/, /əd/). Only one voicing decision still must be made, because these affixes 'agree' for voicing; but multiple fricative and stop combinations are introduced. Also longer sequences are possible (up to four), and fewer constraints upon combinations.

One simplification, however, is that the approximant possibilities are reduced to just /r/ or /l/ in English, and reduced even more (though with a compensatory increase in diphthong possibilities) to just /l/ in some dialects (non-rhotic; no /r/ sound).

The basis for fricative and stop decisions is essentially as for syllable initial position. For nasality (in clusters) and voicing, however, the decision has much to do with the preceeding vowel.

NASALITY. A vowel preceeding a cluster involving nasality will itself be nasalized, and this will acoustically be cued by the 'nasal' formant at about 300 Hz and a general widening of the other formants. Whether an actual nasal consonant is produced is less important. 'Granted' will be heard the same whether pronounced /grăntId/ or /grătId/. Of course if the pronunciation changes to /grănId/ then the nasal consonant must occur; but then it is no longer a consonant cluster. One might even say that the vowel *always* carries the nasality information, and the /n/ is only required to mark the syllable as closed.

VOICING. The syllable final voicing contrast is almost a misnomer, because the periodicity of the signal during the final portion of the syllable is somewhat irrelevant. The strong cue is the length of the preceeding vowel: long for voiced, short for voiceless. The syllable initial VOT and the syllable final vowel length decisions are both temporal; thus decisions about the time domain must be made correctly in order to properly decode the 'periodicity' distinction voiced vs voiceless.

Differentiation Within a Manner Group: Place Contrasts

Speech perception is now almost complete. Given all the decisions about prosodic shape and stress pattern and syllables and syllable structure and the decisions about which manner categories are involved in consonant clusters, only differentiation within a manner group remains. This differention consists of place distinctions. It deserves note that there are strong visual place cues. In fact, the visual cue can be stronger than the acoustic cues for bilabial stops (McGurk and Macdonald, 1976).

At this lowest level the decisions are made which finally unambiguously decode the 'sounds' of English. These decisions do not 'recognize' sounds; they decide amongst a few alternatives. Once all the decisions about the structure of a syllable margin have been made (manner and voicing contrasts) only decisions within a manner group (roughly, decisions as to place) remain.

stop:	/ptk/ or /bdg/
nasal:	/mnŋ/ (just /mn/ syllable initially)
approximant:	/wjrl/ (just /rl/ or just /l/ syllable finally)
fricative:	/fθsʃ/ /vðzʒ/ (/h/)
(affricate:	/tʃ/ /dʒ/)

Affricate manner could be treated as detecting stop followed by fricative.

The status of /h/ requires discussion. It is only possible syllable initially, has no voicing contrast, and is acoustically similar to a fricative but produced by 'aspirant' turbulence at the larynx rather that fricative turbulence at another constriction. In articulatory and phonological terms it might be considered a separate manner, but for perception it is close to the other sounds produced by turbulence, the fricatives. So decoding /h/ can be considered part of the determination of place, along with /fθsʃ/ (presenting an asymmetry in that it has no voiced counterpart, owing to laryngeal 'place).

The differentiation within these groups will now be discussed in detail.

STOP. The F2 transition must now be used. Bilabial /pb/ lower F2, alveolar /td/ raise it. Because actual place of articulation varies according to the neighbouring vowel for /kg/ (closure may occur along the hard palate rather than at the velum), the acoustic cue for /kg/ is modified by vowel context.

These transitions are among the most difficult cues in speech. The

durations can be less than 50 msec, and the amplitude is low, 30 dB or more below the level of an adjacent vowel.

NASAL. The same F2 transition cue as used for stops will separate /m/ from /n/. Syllable final /ŋ/ has the same variation according to neighbouring vowel as has /kg/. There is no syllable initial /ŋ/.

FRICATIVE. The sibilants /sʃ/ /zʒ/ and non-sibilants /fθ/ /vð/ have a sizeable amplitude difference. The voiceless non-sibilant fricatives are the weakest of the sounds of English, roughly 30 dB below the loudest vowel sounds. Separation by amplitude and voicing leaves the four pairs listed above.

All that remains is the decision within these pairs, which is mainly cued by the frequency at which the noise energy is concentrated. For the sibilants, /s/ and /z/ will have appreciable energy between 4 to 8 kHz, while /ʃ/ and /ʒ/ will have a lower frequency concentration. This distinction is very much a matter of a contrast between relatively higher and relatively lower; absolute values vary greatly across speakers.

Similarly, the dental fricatives /θ/ and /ð/ will have a higher frequency concentration of energy than for /f/ and /v/.

Finally, the aspirant /h/ does not have a characteristic energy concentration, and this fact becomes its distinctive cue. The energy will be determined by vocal tract shape, thus making the spectrum for /h/ very much conditioned by any associated vowel. Indeed in whispered (all aspirant) speech, the /h/ isn't really distinguishable from a neighbouring vowel in purely acoustic terms.

In all cases fricative energy is mainly above 1 kHz, and in many cases, as mentioned, the sound is quite weak. Thus problems with fricatives provide an early warning system for hearing impairment.

As with stops and nasals, visual cues are useful for partly differentiating fricatives: /θ/ and /ð/ are clearly different from /f/ and /v/; and liprounding is a cue to /ʃ/ and /ʒ/ vs /s/ and /z/.

APPROXIMANT. In general, the approximants have an amplitude somewhat below that for vowels. This is especially true for /r/ and /l/. A strong cue for /r/ is provided by F3. In a spectrogram a very clear dip in F3 is observed for an intervocalic /r/. Another clear cue to /r/ is liprounding. Both /r/ and /l/ produce a lowering of F2, though the significance of this effect varies according to the vowel involved (vowels with a high F2 are affected more than those with a low F2).

The 'semivowels' /w/ and /j/ are similar to diphthongs, with two differences:

1. formant motion is faster (though still slower than for stop and nasal transitions) and
2. formant motion is greatest at the beginning of the sequence (a /w/ or

/j/ followed by a vowel), whereas for diphthongs the motion is faster toward the end.

Note that cue (2) can only be used (in English) because /w/ and /j/ are restricted to syllable initial position.

With completion of the place decisions the decoding of the speech signal is complete.

Speech Perception

Auditory Space

The implications for speech perception are a main concern in hearing loss. It is common in audiology texts to summarise certain acoustic properties of the speech signal on a sort of audiogram (level vs frequency), with areas marked out for the various categories of speech sound. A typical diagram is Figure 10.

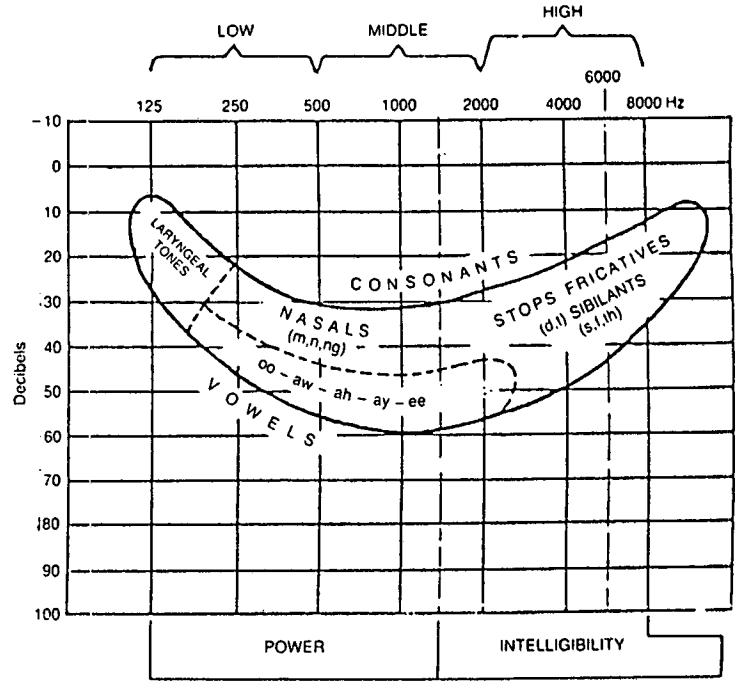
Such a chart is informative but incomplete. It is a map only of the short-term spectral content of speech superimposed on the auditory dimensions of response to the level and frequency of pure tones. It provides mainly a summary of the parts of 'auditory space' involved in vowel contrasts and place/manner decisions.

Pitch detection is usually neglected, or consigned misleadingly to the part in auditory space occupied by the 'laryngeal tone'. As has been discussed, strength of the fundamental is not the determiner of pitch, and a more informative diagram would show pitch related information as dependent upon the entire spectral content of the signal, although a voiced sound would ordinarily have most of its energy below 1 kHz.

The real problem with conventional auditory space, for pitch perception and generally for all of speech processing, is that the time dimension is neglected. Speech perception depends upon a three dimensional auditory space (amplitude vs frequency vs time), and the conventional diagram is just a slice through this space. The time we selected for Figure 10 is at the short-time end of the complete picture. Consideration of a time dimension is particularly relevant to pitch perception and sensory-neural hearing loss, because of the existence of a temporal mechanism for frequency discrimination.

Audiologists are familiar with the frequency selectivity along the basilar membrane within the cochlea. Hair cell loss or other damage impairs the capabilities of this frequency analyser. Stronger signals must be used to get a response. Minimum detectable frequency change is increased. The rate at which frequency changes can be followed is reduced. Finally, frequency selectivity is reduced (Moore, 1982).

But pitch judgements for speech do not have to depend upon this 'place mechanism' of frequency analysis, as a temporal mechanism is available for



The frequency components of English speech sounds.

Figure 10. Auditory Space. From Ballantyne, J. (1970) *Hearing*, 2nd Ed.

signals up to at least 1 kHz. Similarly, the various durational aspects of speech contrasts can also be handled by temporal processing. These temporal capabilities are not essentially properties of the cochlea, as is the place (on the basilar membrane) mechanism of frequency discrimination.

Figure 11 shows normal frequency discrimination. It is a plot of the minimum detectable change in the frequency of a pure tone as a function of frequency. Such a minimum change is referred to as the Just Noticeable Difference (JND). For auditory frequency discrimination, there are essentially two quite different areas on the graph.

1. Discrimination for frequencies below 2 kHz is nearly constant at about 2–3 Hz.
2. Discrimination above 2 kHz follows a ratio scale. The JND is a fixed ratio of about 0.2% of the frequency.

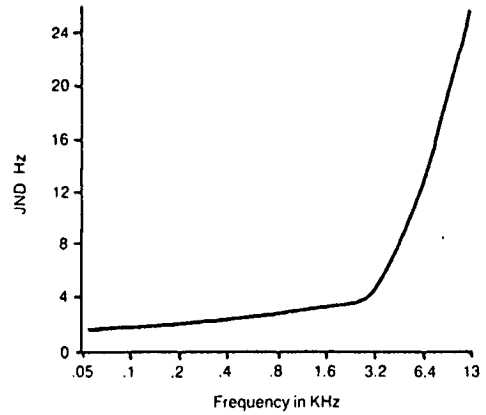


Figure 11. Frequency Discrimination

The difference in slope for the two regions reflects the difference in mechanism. The implication for hearing impairment is that low frequency temporal processing, including the entire range for human voice pitch, may continue when place processing is lost.

Temporal Processing

The only decisions that have no temporal aspect are those relating to vowel quality. Even for vowels we can only establish (in English) five categories before we must consider the time dimension for 'short' vs 'long' and for diphthongs. Everything else directly involves time.

PROSODIC SHAPE. The first decisions require pitch and pitch motion to determine the major parts of the utterance, called tone groups. The pitch itself depends upon the temporal mechanism (periodicity) of pitch detection. The periods involved for the human voice range from 20 msec down to 2 msec. The minimum detectable change in period can be deduced from Figure 11, but is presented directly in Figure 12. It can be seen that the changes in period required to detect a frequency shift are relatively large at low frequencies (and large periods), whereas they are small at high frequencies (small periods). At 2 kHz a JND of 4 Hz represents a change in period of one microsecond. Obviously a temporal mechanism needs to be extremely accurate to work at 2 kHz, and would have to be impossibly refined for higher frequencies. Thus it is very reasonable for the place mechanism to take over for the higher frequencies.

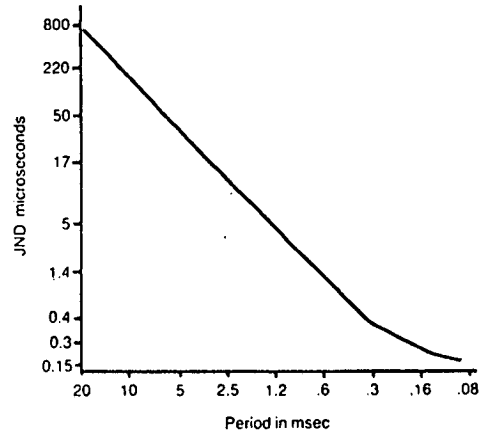


Figure 12. Period Discrimination

The size of a tone group can vary from a single syllable ("No!") to very many words ('I THOUGHT it was Miss Scarlet in the library with a piece of pipe').

INTONATION. The patterns of pitch change (and related loudness and duration changes) operate at the phrase or tone group level. Divisions may be made within a tone group, though this becomes very much a matter of the 'school' of intonation theory to which one adheres. The important fact is that intonation patterns never operate below the unit of the syllable. Thus the fastest motion we should expect (in terms of frequency change per unit of time) is about an octave over the duration of a single syllable (probably several hundred msec for so large an intonation marker). Ordinarily pitch motion will be rather less emphatic, and have lower rates of change.

STRESSED SYLLABLES. Stressed syllables have been defined in terms of their length and ability to carry an intonation marker. A syllable is roughly 100 to 500 msec. Durational differences between stressed and unstressed syllables are usually quite clear; a two-to-one ratio is typical.

SYLLABLES. Defining the syllable is fraught with difficulties. The easy case is a vowel surrounded by stops. The utterance /bebebeb/ is clearly three syllables, and each syllable division is marked by a 30 dB amplitude change over as little as 30 msec. At the difficult end are words like "power", and words where only the approximants /wrlj/ divide the vowels. For such words there may be no amplitude changes, but there will still be a change of spectrum vs time, occupying something like 50 to 100 msec in the ordinary

case. The hearing impaired person can be expected to have most difficulty with these spectral cues to syllable division.

VOICING. Voicing is the name for the contrast in English between two groups of stops and fricatives which otherwise have identical place and manner. The actual acoustic cues involve periodicity, spectral balance, and vowel length. These decisions are made only once per syllable margin (consonant cluster). Allowing for the open syllables and for those syllables not involving stops and fricatives, there is as a rough average one voicing decision per syllable. The duration of the cues themselves, however, can be very brief: 30 msec VOT difference between a definite /p/ and /b/. Vowel length differences for syllable-final voicing will be roughly twice this long, or more.

MANNER CONTRASTS. As with voicing, there is only one decision per syllable margin for each of the possible manners. So again the decisions are progressing at something like the syllabic rate, though there can be two stops or fricatives in a syllable final cluster (because of suffixes). The manner cues thus can be thought of as occupying the roughly 50 to 100 msec slot that we have allowed for syllable margins, and in worst case will be half that length. Manner cues tend to depend mainly upon lower frequencies, below 1 kHz. The exception is frication, which is characterized by random noise above 1 kHz.

PLACE CONTRASTS. Place contrasts operate at the segmental level, and thus have multiple decisions per syllable margin. The decision making is simplified by various phonological processes. Elision (deleting segments) and assimilation (making adjoining segments agree for place) reduce the number of place decisions. Thus though one might expect at worst to make five or six (or possibly seven) place decisions for a single syllable, the average is probably again something like one per syllable margin.

Place cues are the shortest speech cues, principally involving 20–50 msec transitions in the second formant. These cues will be in the 1 kHz to 3 kHz region, and will be transitions from or to low amplitudes for stops and nasals. Cues to fricative place will be more spread out, from 1 kHz up to 8 kHz or so, but will also be of low amplitude, as much as 30 dB less than the center of a stressed vowel. These temporal considerations are summarised in the three-dimensional auditory space diagram, Figure 13.

How To Use Speech in Speech Audiometry

Two important questions about various approaches to speech audiometry are: (1) the kind of speech to be used, and (2) the kind of results to be collected. The various answers produce a range of possibilities,which will be considered in turn.

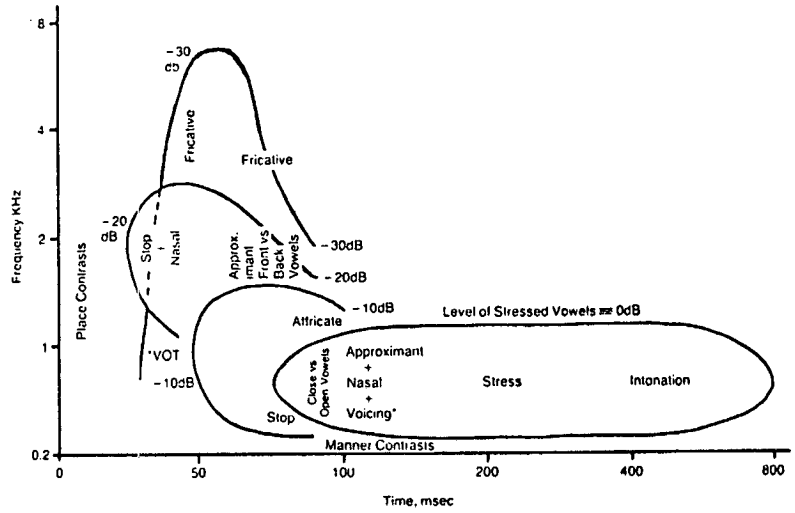


Figure 13. Auditory Space-Time

Speech Detection and Reception Threshold

The simplest possible approach is to:

(a) Dispense with all the linguistic levels except the very lowest and simply use isolated monosyllables.
(b) Ignore all questions of structure. Do not score anything about phonetic type or context. Just score items as right vs wrong.
(c) Ignore confusions between perception and production: the person taking the test 'simply' says what he or she hears. These are called open response tests, because the set of allowed responses is open rather than closed.
(d) Ignore all problems of constraints upon perception and production. Ignore differences in probability of various sounds and sequences; ignore the general tendency to 'grasp after meaning', to prefer a known word to a nonsense word. Ignore the difficulty of producing non-English sequences, even if they were perceived.
(e) Do not enquire into what the person hears. Simply concentrate on the acoustic level at which the sounds begin to be heard (speech detection threshold, SDT), or the level at which sounds are heard with a specified accuracy (such as 50% correct; speech reception threshold, SRT). This approach concentrates on questions of level

rather than questions of speech, which is convenient for reducing speech audiometry to something like pure tone audiometry.

Speech Audiogram

The next simplest form of speech audiometry is to make all the simplifications just mentioned, with the exception of testing at various discrimination levels. A graph can then be made of percent correct vs presentation level, Figure 14. In psychophysics such a graph is called a discrimination function, but audiology uses the term speech audiogram.

A speech audiogram improves upon a simple SDT or SRT measurement in that it recognizes that speech is more complicated than pure tones, and hence there are more questions to ask than simple detection threshold. But all the speech audiogram actually tests is multiple thresholds rather than single thresholds. The whole approach still reflects a preoccupation with levels.

Closed Response Tests

The problems of interaction of production with perception and the constraints of the language involved are largely overcome through 'multiple-choice' tests (see Stevenson and Martin, 1977 for a review). The subject selects from a small set of allowed responses. The selection is usually made by marking a response sheet. A written response simplifies the test, eliminating problems of speech production on the part of the person taking the test (and of perception of that production). Special 'pointing to picture' responses have been used with children.

Closed response tests were originally developed to assess the intelligibility of speech transmission systems (the Rhyme Test and its descendents; Hawley, 1977).

A very important benefit of closed response tests is control over 'errors'. The key to behavioural tests is to restrict the subject just to those decisions being investigated. With only a few error possibilities it becomes possible to analyse errors rather than simply to accumulate them. Such diagnostic tests (DRT, Voiers, 1977; FAAF, Foster and Haggard, 1984) can then be used to add a qualitative aspect to the audiometry. Such tests not only quantify the impairment, but point to specific problem areas.

Continuous Speech Tests

The concept of linguistic levels and their associated constraints can be introduced through the use of continuous speech, sentences and paragraphs. Unfortunately, continuous speech adds a whole new set of problems: responses can now be even more variable than for single words, scoring

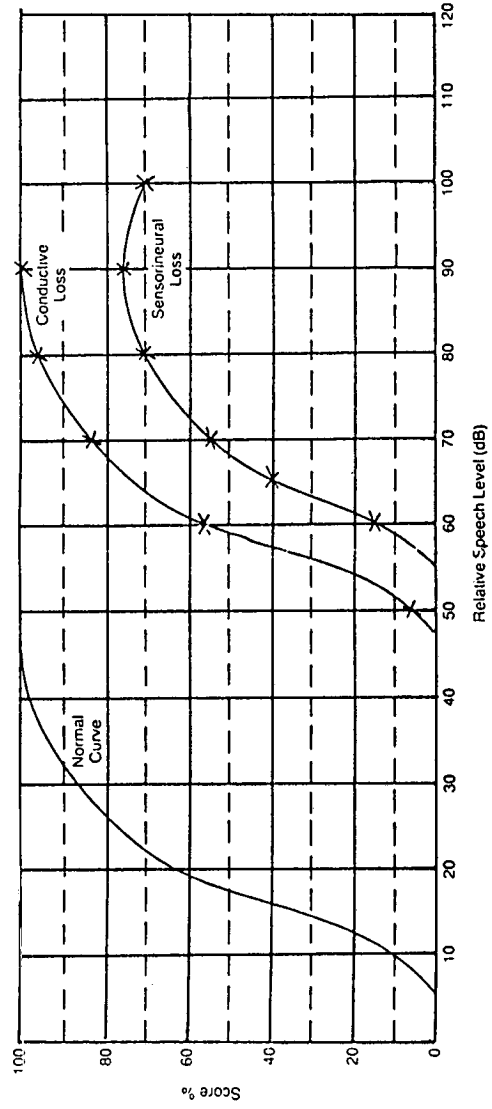


Figure 14. Speech Audiogram with the format recommended by the British Society of Audiology published in the British Journal of Audiology Vol. 9, 1975. A 20% change in score is equal to a 10dB change in level. The shape of the normal curve will depend upon the speech material used. The other two curves illustrate the results from two people one with a conductive hearing loss and the other with a sensorineural loss.

becomes almost impossible, and it is hard to control the difficulty level of materials.

Again, it improves matters greatly to use closed response tests. One particular development (the SPIN test; Kalikow *et al.*, 1977) really only extracts one bit of information out of the whole linguistic structure: how the same acoustic item is perceived in a highly constrained sentence (high predictability) vs a marginally constrained sentence (low predictability). Much research involves linguistic constraints, but very few clinically appropriate tests exist.

Synthetic Speech Tests

Systematic and detailed investigation of auditory time–space requires the use of synthetic stimuli. Much basic research has been performed, but little of direct clinical application. However the advent of the microcomputer and the proliferation of speech synthesis devices at much lower cost than was the case ten years ago eliminate the main technical difficulties with the use of synthesis.

Synthetic speech adds the missing dimensions to a speech audiogram. With synthesis an experimental continuum can be explored step-by-step. The result is a comparison with normal performance on a single acoustic cue to a single speech contrast. Thus in a methodical way the whole pattern of contrasts can begin to be tested. Further, the fact that synthetic speech requires a control computer can be turned to advantage and the stimuli can be computer generated according to the subject's response pattern. These adaptive tests can be made comparatively efficient.

Further, synthetic speech allows a form of speech audiometry which can be made suitable for use with the profoundly deaf, with persons who score at chance level on conventional speech tests. Such testing is very relevant to the screening of candidates for tactile and electrical prostheses, such as cochlear implants and wearable vibrotactile stimulators. (King, this volume; Fisher *et al.* (1983); Pickett *et al.* (1983); Hazan and Fourcin (1985).)

There are many difficulties with speech audiometry. Several have just been mentioned, and other chapters of this book will raise further problems. But we point to these problems in order to find solutions. It should be of great interest for the reader to see how the remaining chapters in this book attempt not just to find problems, but to solve them.

SOUTHAMPTON UNIVERSITY LIBRARY

USER'S DECLARATION

AUTHOR: WRIGHT, R.D. TITLE: An investigation of speech synthesis parameters DATE: 1988

I undertake not to reproduce this thesis, or any substantial portion of it, or to quote extensively from it without first obtaining the permission, in writing, of the Librarian of the University of Southampton.

To be signed by each user of this thesis

DATE	SIGNATURE	ADDRESS