Chapter 1

Introduction

1.1 Justification

The study of the acoustic structure of languages is already a mature field. Certain influences keep it in flux though. There are new analysis techniques being developed continuously and faster computers now allow us to study at a more complex and in depth level than before. Another important factor is that although the large (in terms of speakers) languages of the world (English, French, German, Mandarin etc.) have been studied in depth, there are many smaller (yet acoustically interesting) languages still awaiting careful study. In this study we concentrate on one of these languages, namely Afrikaans, and analyse the acoustic structure of its long vowels and diphthongs as spoken by first language Afrikaans speakers and first language South African (SA) English speakers. The vowels and diphthongs of SA English are also studied and a comparison is made between first and second language speech.

There are many reasons why we would want to study the acoustic structure of a language and know what influence different mother tongue accents would have on the acoustic structure. Some of the fields which would benefit from such research are:

- Automatic speech recognition
- Automatic accent recognition
- Phonetics
- Synthetic/computer speech or text-to-speech (TTS) systems

Speech recognisers use the statistical probability of occurrence of a sequence of acoustic observations to determine what a speaker is saying. The modelling of these acoustic observations can take many forms such as Gaussian mixtures of mel scaled cepstral coefficients or linear predictive coding coefficients[1]. These "acoustic models" are then used in recognition algorithms such as Viterbi decoding used in conjunction with hidden Markov models (HMM's), dynamic time warping comparators or neural networks to perform recognition. These technologies are not directly relevant to this study, but rather the fact that they are all in some way based on a type of acoustic model of the spoken language.

The acoustic modelling entities we will use in our study are the well-known formants - the resonant frequencies of the vocal tract[2]. We will also spend some time looking at prosodic modelling (more specifically, pitch modelling) of the vowels and diphthongs as it is often in this respect that languages and accents may differ significantly. We pay special attention to the diphthongs and formulate a more accurate and informative measure of diphthongization and use this to analyse some controversial vowels.

We concentrate this study on the long vowels and diphthongs for the following reasons:

- The short vowels were addressed in a previous study[3].
- The vowels and diphthongs are arguably a larger source of differences in accents and languages than the consonants.
- They are the "glue" which hold the consonants together to form words.

- They lend themselves to acoustic analysis by being voiced and relatively easy to segment.

We aim to form simple acoustic models of the long vowels and diphthongs which can then be used to determine if and where differences occur between SA English and Afrikaans mother tongue pronunciation of these. We then determine how large these differences are. The diphthongs in particular are studied in a new way to clarify the status of certain sounds which are classified as vowels by some phoneticians and as diphthongs by others.

The envisioned uses of this knowledge in the fields mentioned above could be the following:

- Speech therapists and language teachers may use the differences in pronunciation of the vowels and diphthongs in elocution lessons to teach different accents.
- Knowledge of the differences between the acoustic models can be used in speech synthesisers to create a pleasing or different accent.
- During training of speech recognisers, certain vowels and diphthongs can be targeted for re-estimation or retraining to improve recognition rates for different accents.
- Pronunciation dictionaries[4] contain valid phonemic transcriptions of words for a particular language and dialect. This is useful in both automatic speech recognition and TTS systems. A better understanding of the vowels and diphthongs used by South African speakers will result in more accurate pronunciation dictionaries.

It is not sufficient to prove that HMM's are capable of distinguishing and recognising various accents. We need to know the specific acoustic differences between accents so that we can make justified decisions when choosing training sounds for a speech

recognition data base. "Black Box" accent/dialect recognition tests such as performed by Miller and Trischitta[5] and Teixeira et al.[6] help us to determine the flexibility of HMM's, but they do not assist us in choosing word lists or structure for the text material of future databases. By knowing which sounds differ significantly between languages and accents we can, in principle, endeavour to collect only the required adaptation data required to retrain a recogniser, thus reusing the expensive data we have already collected for an alternative accent or language. Some studies[7] propose dialect recognition using shibboleth words, but this only demonstrates that HMM's can be used to model accents. Other studies[6][5] demonstrate that HMM's can be used to recognise phonemes as belonging to a certain dialect, but this does not tell us what makes the phoneme unique, or how we can approach improving recognition by adapting existing models which may have been generated at great expense. We must analyse the structure of languages and see how they differ at a phonemic level.

We further justify studying second language structure with reference to speaker adaptation and quote from Digalakis and Neumeyer[8]:

> "Adapting the parameters of a statistical speaker-independent continuous-speech recogniser to the speaker and the channel can significantly improve the recognition performance and robustness of the system. We have recently proposed a constrained technique for Gaussian mixture densities. The recognition error rate is approximately halved with only a small amount of adaptation data, and it approaches the speaker-independent accuracy achieved for native speakers."

The hypothesis that we are going to test, is that there are measurable and significant differences between the first and second language pronunciations of Afrikaans and SA English long vowels and diphthongs. We test this under the assumption that knowledge of these differences can be used to improve the recognition rates of automatic speech recognition systems.

A side issue that we address is the issue of diphthongization of the long vowels - also in the framework of an L1-L2 comparison.

1.2 Background

The acoustic structures of British and American English have been studied intensively over the last hundred years. Perhaps one of the most famous researchers in this field, Daniel Jones[9] is largely responsible for the International Phonetic Association (IPA) vowel chart still used today. His research into the location of the extreme cardinal vowels is an important reference work.

Working with more realistic (natural) speech, Peterson and Barney[10] analysed the locations of the formants of male, female and child speakers of American English. This work is often used today as a reference of vowel locations and how to plan and carry out speech analysis studies.

Following on the work of Peterson and Barney, Holbrook and Fairbanks[11] analysed the paths followed in formant space by the diphthongs of American English. Although the experiments were not carried out as carefully as those of Peterson and Barney, and the analysis techniques were relatively primitive, the work is an important reference of diphthong analysis. The technique they used is explained in Chapter 2: Theoretical Framework. We do not attempt to compare their results with ours as we would not be able to determine if any differences are as a result of the different analysis technique of accent differences.

Afrikaans has remained quite unresearched in terms of acoustic modelling until 1988 when Taylor and Uys[12] with some insightful writing but inaccurate modelling (due to rounding errors and the use of a single speaker [and thus a biased pronunciation]) plotted one of the first formant maps of the Afrikaans vowels. Although Taylor and Uys did perform diphthong analysis, the techniques used consisted simply of mean

formant locations at the initial and terminal points of the diphthongs with simple linear interpolation between these points. We claim therefore that their technique was too primitive to generate any conclusive results and only indicates general trends.

More recently, Van der Merwe et al.[13] performed a more in depth study of the acoustic structure of Afrikaans vowels. A large percentage of their study revolves around the analysis of formant ratios which is controversial representation of the vowels. The formant ratio theory speculates that although the resonant frequencies (formants) of speech correlate for voiced speech sounds (and therefore appear to have relevance) there may be the possibility that voiced speech structure (and possibly understanding) results from the spacing of the formants (i.e. their ratios). This appears to have some intuitive justification, but there has not been much scientific evidence to support it.

Analysis techniques have improved since the early nineties, and the greater employment of computers to perform the formant extraction, analysis and visualisation has greatly improved the accuracy and repeatability of acoustic modelling experiments.

Perhaps one of the most recent scientific works on the acoustic structure of many of the Afrikaans vowels has been performed by Botha and Pols[3]. This study is based on a relatively large data set and the formants have been carefully extracted as stationary frequencies. These authors also emphasise the apparent relevance of the formant ratio theory. In many ways our work is a continuation of this study where we are concentrating on the long vowels and the diphthongs while paying careful attention to their dynamic nature.

The most recent work performed on certain of the aspects of some of the Afrikaans vowels and diphthongs is the work performed by Raubenheimer[14][15]. Due to the limited publication of master's dissertations and doctoral theses, our attention was only drawn to this work after our own work had been completed.

1.3 Method

This section deals with the experimental protocol of our analysis. The data that was used in the study is first described. Then the experiments which were performed on the data and the methods employed to achieve our aims are introduced. The actual details of these steps are discussed in greater detail in the respective chapters later in this dissertation.

We recorded spoken first and second language data of 17 male speakers from the two language groups (Afrikaans and South African English). The data was listened to and all poor recordings of incorrect utterances were discarded. The remaining data was then segmented and labelled (tagged) for the vowels and diphthongs of interest. We extracted the formants and pitch contours from each labelled segment and once again cross checked this by superimposing the formants on spectrograms. Where possible, incorrect formant trajectories were corrected, and where it was not possible, they were discarded. Pitch trajectories which were sporadic or disjoint or obviously incorrect were also discarded.

The final formant data was then used to calculate the mean locations of the vowels in formant space, and the trajectories of the diphthongs in formant space. The means and the trajectories were then subjected to analysis of variance statistical significance tests to determine whether the two language/accent groups produce equivalent or noticeably different vowels and diphthongs.

The diphthong trajectories were fitted using cubic splines, and the cubic spline coefficients were compared using analysis of variance calculations. This metric for measuring diphthongization is an important step in clarifying and classifying the status of various vowels and diphthongs.

As a further study the mean pitch contours were compared to determine if there were significant intonational (prosodic) differences between the groups.

Importantly, we are not only studying the acoustic structure of Afrikaans first language, but also Afrikaans as second language and similarly for SA English.

1.4 Contributions of this dissertation

The major contributions of this study are those defined by the goals, namely, the modelling of the acoustic structures of the long vowels and diphthongs of Afrikaans and South African English, both in first and second language context, and a statistical comparison of these models.

We therefore contribute:

- Static formant models of most of the Afrikaans long vowels
- Dynamic formant models of most of the Afrikaans diphthongs
- Static formant models of most of the South African English vowels
- Dynamic formant models of most of the South African English diphthongs
- Analysis of variance statistical comparisons between these sets of models where relevant
- A new measure of diphthongization and a resulting clarification on the status of certain vowels which have long been the subject of speculation.

We specifically concentrate on the long vowels and do not duplicate work already performed by Botha and Pols[3] which concentrates on the short vowels of Afrikaans and SA English.

Knowledge of the models above can be used to improve automatic speech recognition. Adaptation to speaker dependent recognition can be carried out more efficiently if we

know which speech sounds are prone to accent shift. Cross-language training of ASR systems can also be facilitated using this knowledge. Recognition databases may also be trained with prior knowledge that certain sounds may be pooled for training as they are common to both language groups whereas other sounds are characteristic of a particular group[16].

1.5 Organisation of this dissertation

The next chapter continues with an explanation of the background to this study. It details the concepts that we are working with and explains the mathematics behind the analysis techniques used.

Vowels are dealt with first. We explain what they are and how we represent them. We also summarise some of the research performed on vowels in the past fifty years.

The logical continuation of vowels, namely diphthongs, are then explained. Although not much research has been performed on diphthongs, we describe some of the ground-breaking work performed in this field of phonetics.

We then go on to explain a graphical technique used to visualise speech in the spectral domain known as a spectrogram. The spectrogram has been an integral part of this study in terms of labelling and data checking.

We then describe the first of the abstract concepts used in this study, namely formants. For various reasons which are explained in this section, formants were used as the primary means of acoustic modelling in this study. The algorithm for calculation of formants is also given.

It was decided that we should examine the intonation (prosody) of the vowels and diphthongs between the two groups to determine whether acoustic differences were only

visible at a phonemic level, or whether accent differences were possibly due to pitch differences. Pitch extraction is easy to achieve in conjunction with formant extraction as they are both based on the calculation of linear predictive coefficients.

We next describe the concept of equivalence classification, in other words, the concept that speakers learning a language at a late age tend to use the phonemes they already know from another language, to pronounce words in the new language. Evidence of this may be difficult to establish in the framework of the current study due to the bilinguality of the speakers, but it is an important concept which must be considered.

The last two sections of the second chapter discuss the use of cubic splines to form a low dimensional representation of the diphthong trajectories and pitch contours. We then discuss the use of analysis of variance statistical tests to determine mathematically how significant the difference in mean or mean trajectory is between the two accent/language groups.

We begin Chapter 3 that deals with our experiments with a discussion of the objectives of the study, that is, what it is we are trying to achieve with this research.

The data we have recorded and the structure of the database are discussed next. We also discuss the speakers and problems encountered with the data recording procedure. The words used and the selection process are discussed.

The "Method" section is second in importance only to the results. This section describes the way we went about verifying the recorded data before extracting the formants and pitch for analysis. It then explains the software that was written to visualise the data in useful ways and how the vowel formant means, diphthong formant trajectories and pitch contours were compared. It also explains how we determined whether a speech sound was a vowel or a diphthong. This is particularly important in sounds which are surrounded by some controversy. This matter is discussed in later sections.

The "Results" section shows the graphs generated by using the software we have written and then discusses vowel by vowel and diphthong by diphthong the conclusions we may draw by observing these graphs and analysis of variance results. We also discuss the level of diphthongization of both vowels and diphthongs.

We conclude with a summary and conclusions regarding the study in Chapter 4.

Chapter 2

Theoretical Framework

In this chapter we discuss the algorithms, techniques and principles employed to model the acoustic differences between Afrikaans and South African English first (L1) and second language (L2) speech.

We begin by explaining what we mean exactly by the terms "vowels" and "diphthongs". With these explanations we include summaries on some of the research performed in the fields of phonetics and phonology pertaining to vowel and diphthong modelling.

Once we have explained what it is we are studying we describe a useful 2-D visualisation tool we have used, known as a *spectrogram*. Spectrograms are easy to plot and they are extremely useful for the labelling (tagging) of speech data where it is often impossible to see phoneme transitions on a 1-D energy versus time speech signal alone.

Once we have labelled our data we need to extract the relevant features from it that we need for the "acoustic modelling". We have chosen as features the resonance peaks of the vocal tract, known as *formants*. We explain the mathematics and algorithms required to extract formants from the speech signal based on linear predictive coefficients (LPCs).

We also study whether intonation (prosody) has a large influence on the perceived accent of the speaker[17]. To this effect we have extracted the pitch contours of the utterances (vowels and diphthongs) of the speakers.

In order to put the analysis and comparisons of our data in a theoretical framework, we consider the work by Flege[18] on the concept of "equivalence classification".

As long vowels and diphthongs have dynamic formant and pitch directories, we choose to model these using cubic splines. With this technique we take multiple samples and fit them to a curve which we can represent with relatively few parameters.

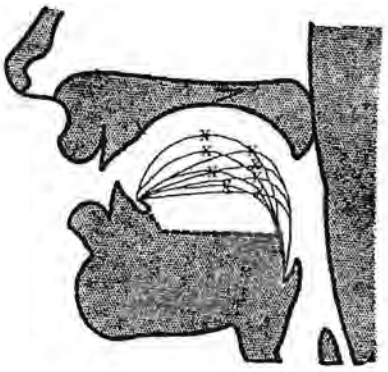
The actual method of comparison is finally explained. We have used the analysis of variance (ANOVA) test developed by Fisher[19] to perform tests which will indicate the significance of differences between the means of two sets, taking the variance into consideration.

2.1 Vowels

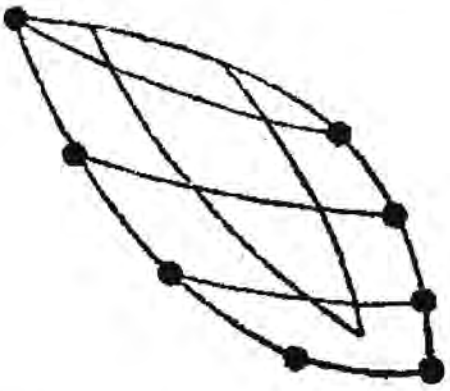
There is no simple definition of what constitutes vowels, but they are generally classified as follows[20]:

> "In ordinary speech a vowel is a voiced sound in the pronunciation of which the air passes through the mouth in a continuous stream, there being no obstruction and no narrowing such as would produce audible friction. All other sounds are consonants."

The difference in quality between vowels is caused by the movements of the tongue and lips which result in a change in the shape of the resonance chamber of the mouth. Vowels are usually classified by the part of the tongue which is raised: front, middle or



Tongue positions of the Eight Primary Cardinal Vowels.



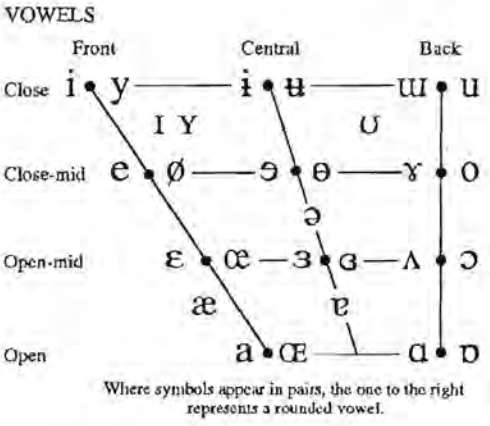


Figure 2.1: The cardinal vowels as organised by the placement of the tongue in the oral cavity. The top diagram from Ward[20] indicates the positions of the tongue which give rise to the eight cardinal vowels. This results in the middle figure (also from Ward) which has been simplified to the current IPA vowel chart seen at the bottom(from the IPA).

Short-vowel		Long-vowel	
i	h**i**t	i:	h**ea**t
u	f**u**ll	u:	f**oo**l
i	*w**ie**l (wheel)*	i:	*spi**eë**l (mirror)*
u	*k**oe**l (cool)*	u:	*b**oe**r (farmer)*
ɛ	*n**è** (not/no[inq.])*	ɛ:	*w**e**ns (wish)*
ɔ	*p**o**nt ()*	ɔ:	*p**o**nd (pound)*
a	*m**a**n (man)*	a:	*m**aa**n (moon)*

Table 2.1: Examples of short vowels as opposed to long vowels in both English and Afrikaans.

back, and according to the degree of raising which takes place, namely: close, half-close, half-open and open. This is clearly illustrated in Figure 2.1.

In this study we have concentrated on the long vowels as opposed to short vowels (which have been analysed in a previous study by Botha[21][3].) The long vowels differ from short vowels not only in their duration but also in their quality and thus in their formant structure. Therefore, the short vowel <i> as in the word "hit" will differ significantly from the long vowel <i:> or <ɪ> found in the word "heat". Some examples of short vowels and their long vowel counterparts are given in Table 2.1. Long vowels are also considered to be prone to diphthongization and we have measured this to determine the validity of such a statement.

Peterson and Barney

Peterson and Barney[10] performed important vowel research in 1952 by recording two lists of ten vowels from 33 men, 28 women and 15 children, thereby creating a database of 1520 words. These were all in consonant-vowel-consonant (CVC) context, and h-vowel-d was the preferred structure where possible as it was found that the consonants in this particular CVC structure were not as prone as other consonants to influencing the integrity of the vowels. Using calibrated Plexiglas templates they read the formant frequencies off spectrographs. The sounds they used are given in Table 2.2.

Vowel	Word	Vowel	Word
i	Heed	ɔ	Hawed
ɪ	Hid	ʊ	Hood
ɛ	Head	u	Who'd
æ	Had	ʌ	Hud
ɑ	Hod	ə	Heard

Table 2.2: The vowels studied by Peterson and Barney[10] with source words

It is important to note that Peterson and Barney only extracted instantaneous formant frequency values at a single point in a vowel sound. All their plots are also based on only these instantaneous formant frequencies. Plots of their results are given in Figure 2.2.

Taylor and Uys

Until recently (1988) no one had performed any in depth study into the acoustic structure of the Afrikaans vowels. Taylor and Uys[12] created a small data set consisting only of vowels uttered by Uys. Using this they created a (speaker dependent) vowel map for Afrikaans. Although not a conclusive study, it is an important reference. Plots of their results are given in Figure 2.3.

Van der Merwe et al.

Van der Merwe et al.[13] recognised the lack of any in depth study into the Afrikaans vowels and their state of change due to foreign linguistic effects. Working with a smallish corpus of 10 male, first language, middle aged speakers, they recorded 3 utterances for each of 8 Afrikaans vowels (<i>, <ɛ>, <æ>, <ə>, <a>, <u>, <ɔ> and <œ>) and processed them. They extracted the first three formants and the fundamental frequencies (pitch). Plots of their results are given in Figure 2.3. They do not state by what means they indicated to the speakers how they should know which

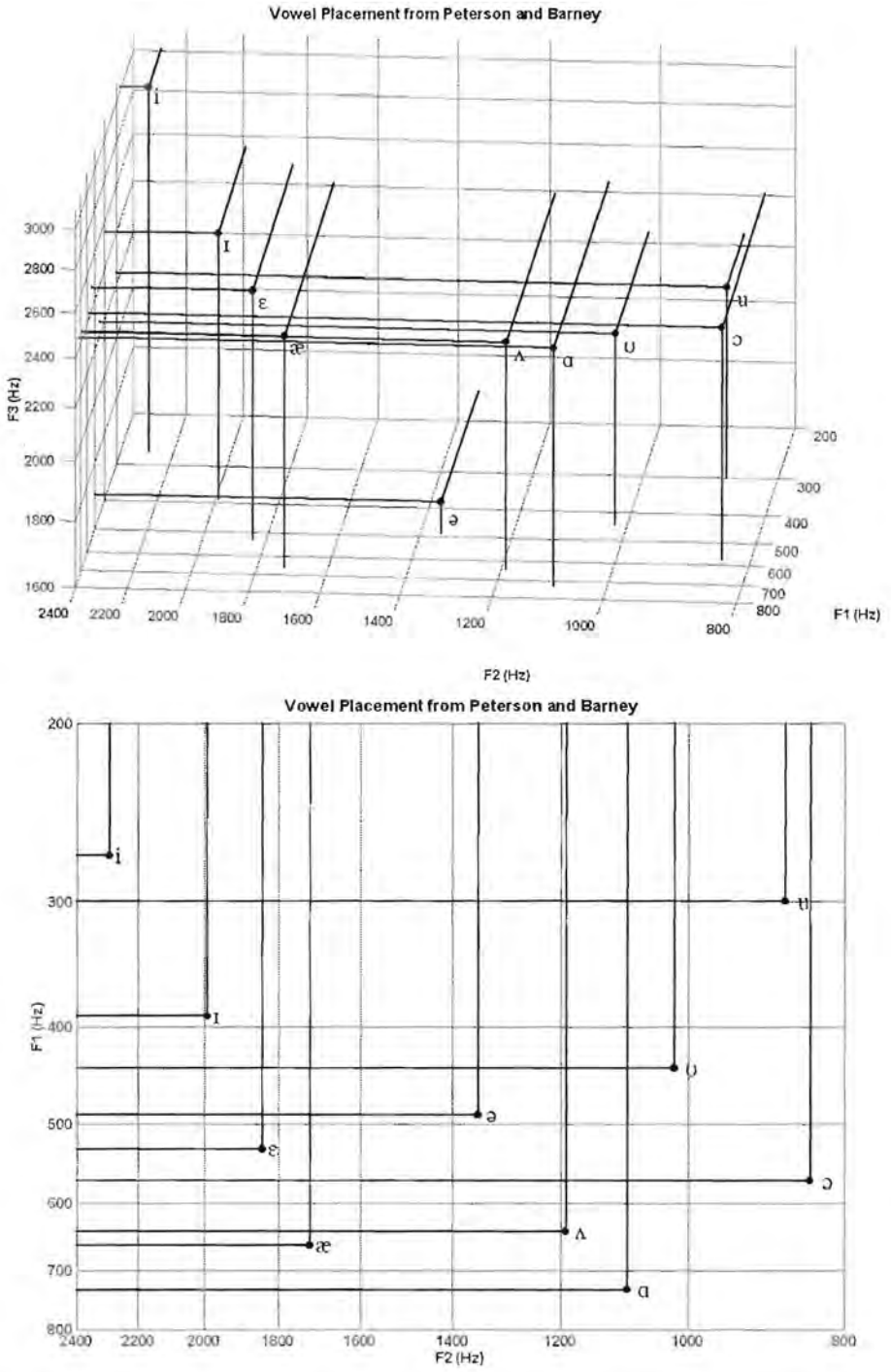


Figure 2.2: The Peterson and Barney[10] vowels in 3 dimensions (F1,F2,F3) and 2 dimensions (F1,F2). Note the similarity between the 2 dimensional plot and the cardinal vowel chart given in Figure 2.1.

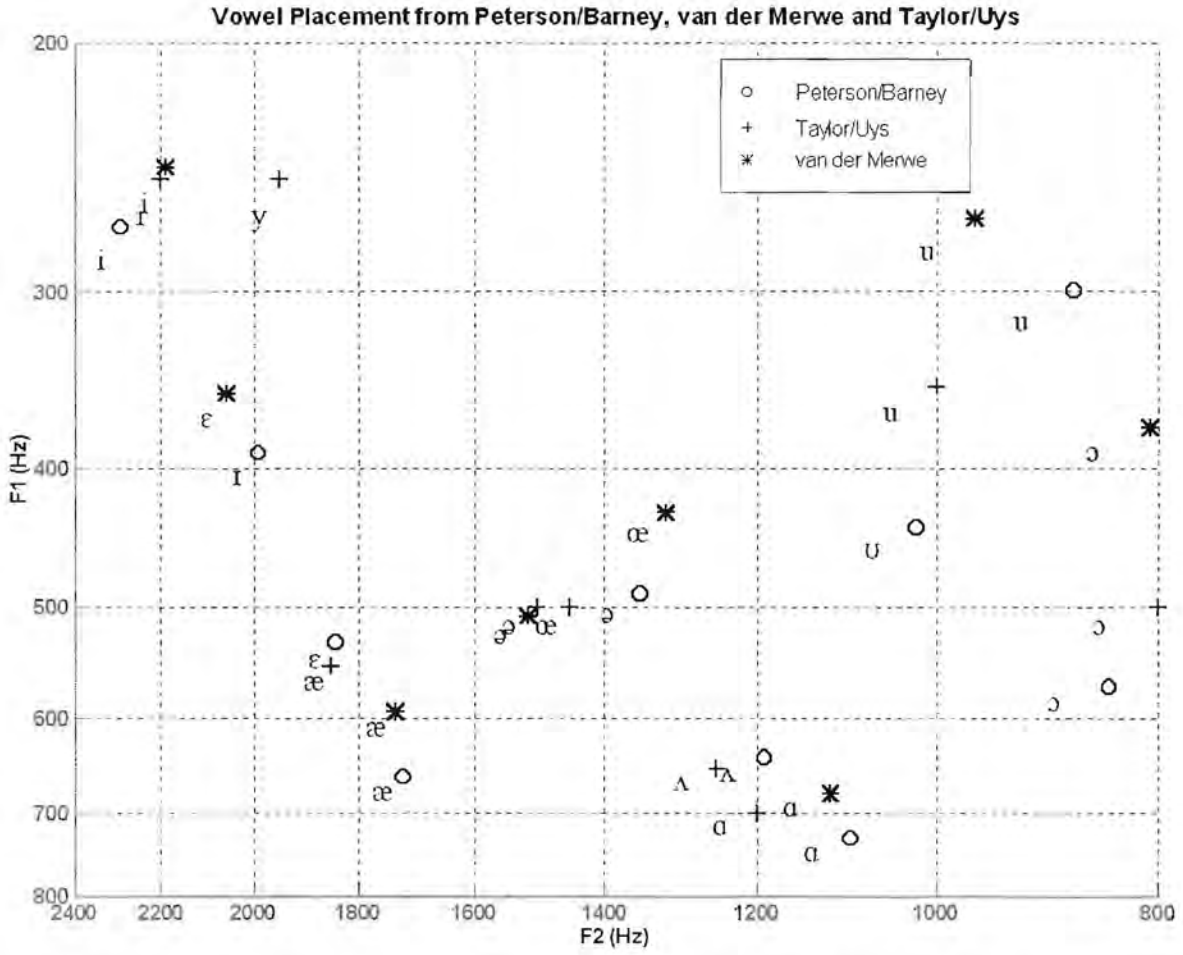


Figure 2.3: The Peterson and Barney[10], Taylor and Uys[12] and Van der Merwe[13] vowels shown in 2 dimensions (F1,F2).

of the 8 isolated vowels they were to utter. It may be as a result of this that they found no clear clustering of <æ> as was found with the other vowels.

Importantly, the authors are mostly of audiological training and found it important to analyse the formant ratios. Although there exists some controversy over the validity and usefulness of the formant ratio theory, there does, to the eye (which is a fairly good pattern recognition device) appear to be sufficient importance to warrant further study into the matter[22].

Botha and Pols

Botha and Pols[3] performed what is probably one of the most recent studies on the Afrikaans vowel system. Their research focused on the short vowels <a>, <æ>, <ɛ>, <i>, <ə>, <œ>, <u>, <y> and <ɔ>. In particular, they studied the mean formant locations of the stationary vowels (Plots of their results are given in Figure 2.4) and the formant ratios. An important distinction of this paper from other phonetic studies is that it examines not only mother-tongue Afrikaans, but also the pronunciation of Afrikaans vowels by mother-tongue South African English speakers. Our study is a continuation of this work, with the emphasis on the long vowels and the dynamic diphthongs.

2.2 Diphthongs

The diphthongs are considered to be a combination of two vowels, so pronounced as to form a single syllable. A list of common diphthongs and words in which they are commonly found is given in Chapter 3 on page 55 in Table 3.2. These gliding sounds are generated on a single impulse of breath. English and Afrikaans diphthongs are of the falling type, having the greater prominence at the beginning. They are called de-crescendo diphthongs[20]. They are generally written phonologically as two

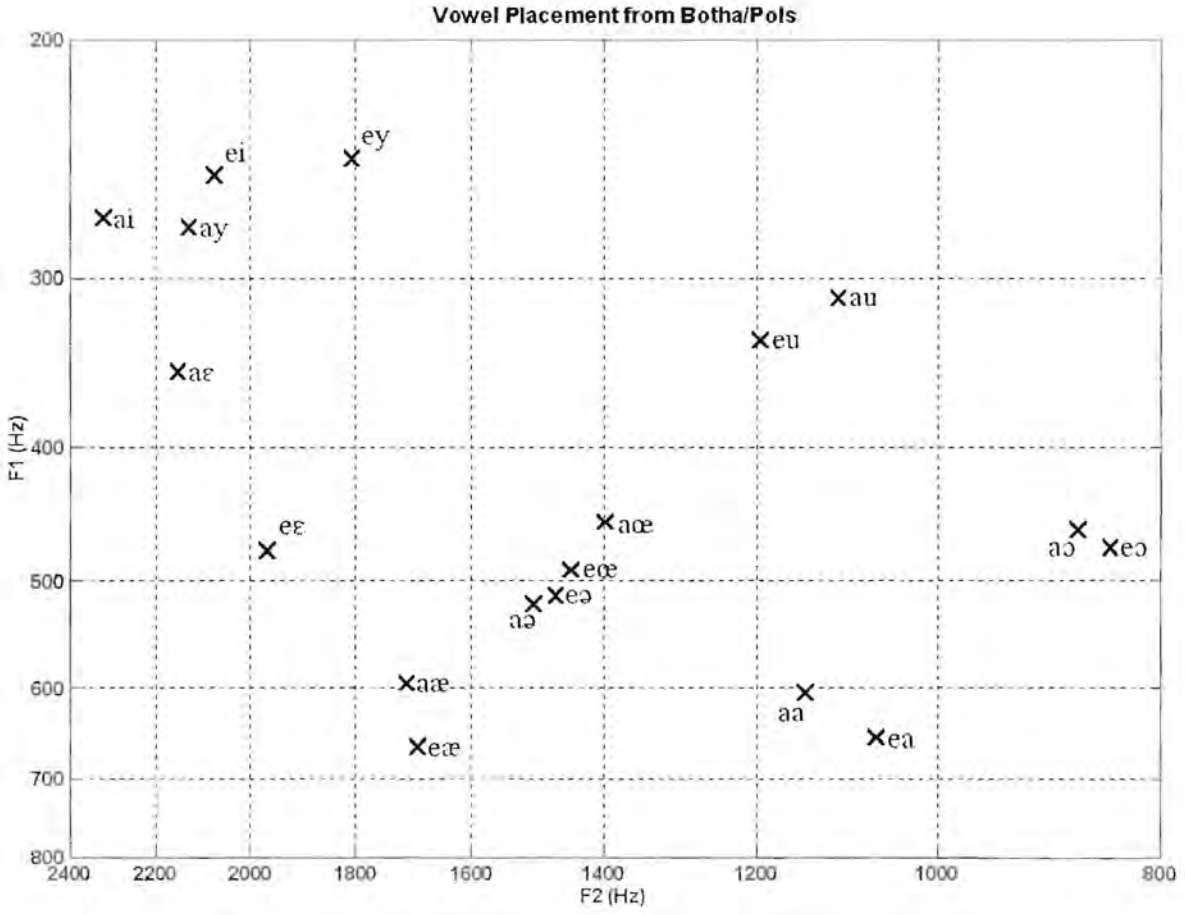


Figure 2.4: The Botha and Pols[3] vowels shown in 2 dimensions (F1,F2).

Diphthong	Word
eɪ	Lady
oʊɪ	Home
aɪ	Time
aʊ	Now
ɔɪ	Boy
ɪə	Here
ɔə	More
ʊə	Your

Table 2.3: The diphthongs commonly found in English with example words (from Ward[20]).

orthographic symbols, the first being the starting point (vowel) of the tongue and the second being the terminating point.

In principle it is possible to move from any vowel to any other and thus the number of diphthongs would seem immense. In reality, however, certain sounds are either too complex (tongue tying) or awkward sounding to be used. According to Ward[20] the majority of English speakers possess nine diphthongs. These are given in Table 2.3.

Afrikaans has a more complex diphthong structure. In Afrikaans phonologists traditionally only recognise three true diphthongs, all others are pseudo diphthongs[23]. We concur with Taylor and Uys's[12] definition of diphthongs. They go to great effort to explain their reasoning and critically evaluate the arguments (or lack thereof of others). We summarise their comments here:

Phonologists (e.g. Coetzee[23]) state:

- The three "true" diphthongs recognised are <əi>,<œy> and <œu>.
- In these true diphthongs both vocal components are of equal length, being lengthened by an equal degree when lengthened expressively.
- In pseudo diphthongs only the initial component can be stressed and only this

component can be lengthened expressively.

Taylor could find no empirical evidence to support this but do concede that this form of classification may be valid on phonological grounds.

Taylor summarises with:

- True diphthongs consist of: [VV] - where each component has short vowel status.
- Pseudo diphthongs consist of: [V:VC] - an initial long vowel [V:], another short vowel [V] and [C] representing the final [j] or [w] glide.
- Diminutive diphthongs: [CV] - There is only a single diminutive diphthong which occurs i.e. the Afrikaans "-jie" or [-ci]. Both components are very short.
- Diphthongised long half-close vowels: Seen as monophthong "vowels" by phonologists, namely [e:,o:,ϕ]. Afrikaans linguists seem to downplay this phenomenon which produces the only centring diphthongs in the language and call them "potential diphthongs". They call the "vowels" "*swak gesnede*"(unchecked) and regard the process as of a purely mechanical and perceptually irrelevant contaminant of vowel length. Taylor labels them as [iə,uə,yə]. See Figure 3.6(bottom left) and Figure 3.7(bottom left) for confirmation of this classification for $< e :>$ and $< o :>$.

Figure 2.5 shows some of the English diphthongs indicating their origins and terminating points in relation to the standardised vowel chart.

Holbrook and Fairbanks

Holbrook and Fairbanks[11] continued with the work of Peterson and Barney by analysing five of the common diphthongs found in American English. They are given in Table 2.4 with example words in which these diphthongs are found and the diphthongs are

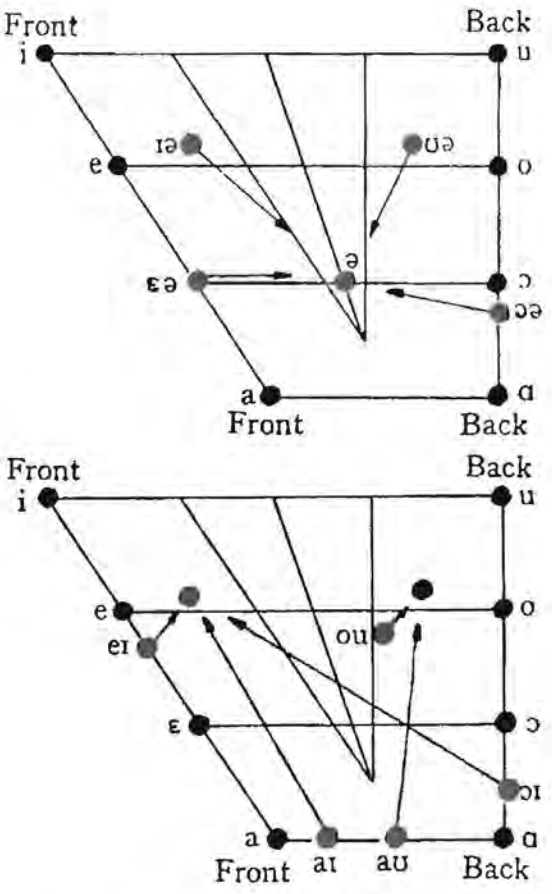


Figure 2.5: Some common English diphthongs and their movement in vowel location as described by the standardised vowel chart (from Ward[20]).

Diphthong	Word
eɪ	Hay
aɪ	High
ɔɪ	Hoy
oʊ	Hoe
aʊ	Howe
jʊ	Hugh

Table 2.4: The diphthongs used by Holbrook and Fairbanks[11] with source words

graphed in Figure 2.6. Although they used slightly more modern equipment, their technique was similar to that of Peterson and Barney. The formants were measured at five points over the period of diphthong voicing. The means of these formant points were then plotted. Essentially this was Peterson and Barney's technique at multiple points along the sound. Their results show reasonably clearly the formant movement as articulation moves from one vowel to the next.

Taylor and Uys

Although Taylor and Uys[12] did process some of the Afrikaans diphthongs, their analysis methods were somewhat rudimentary and we can make no comparison between their results and the results found in this study. Their analysis technique consisted of averaging the formant values of the first quasi-stationary section of the diphthong and then plotting a linear interpolation to the average of the last quasi-stationary section of the diphthong. We therefore make no further mention of their diphthong analysis.

2.3 Spectrograms

To effectively label the long vowels and diphthongs within the speech segments we have recorded, and in order to check formant extraction, we require a simple way of

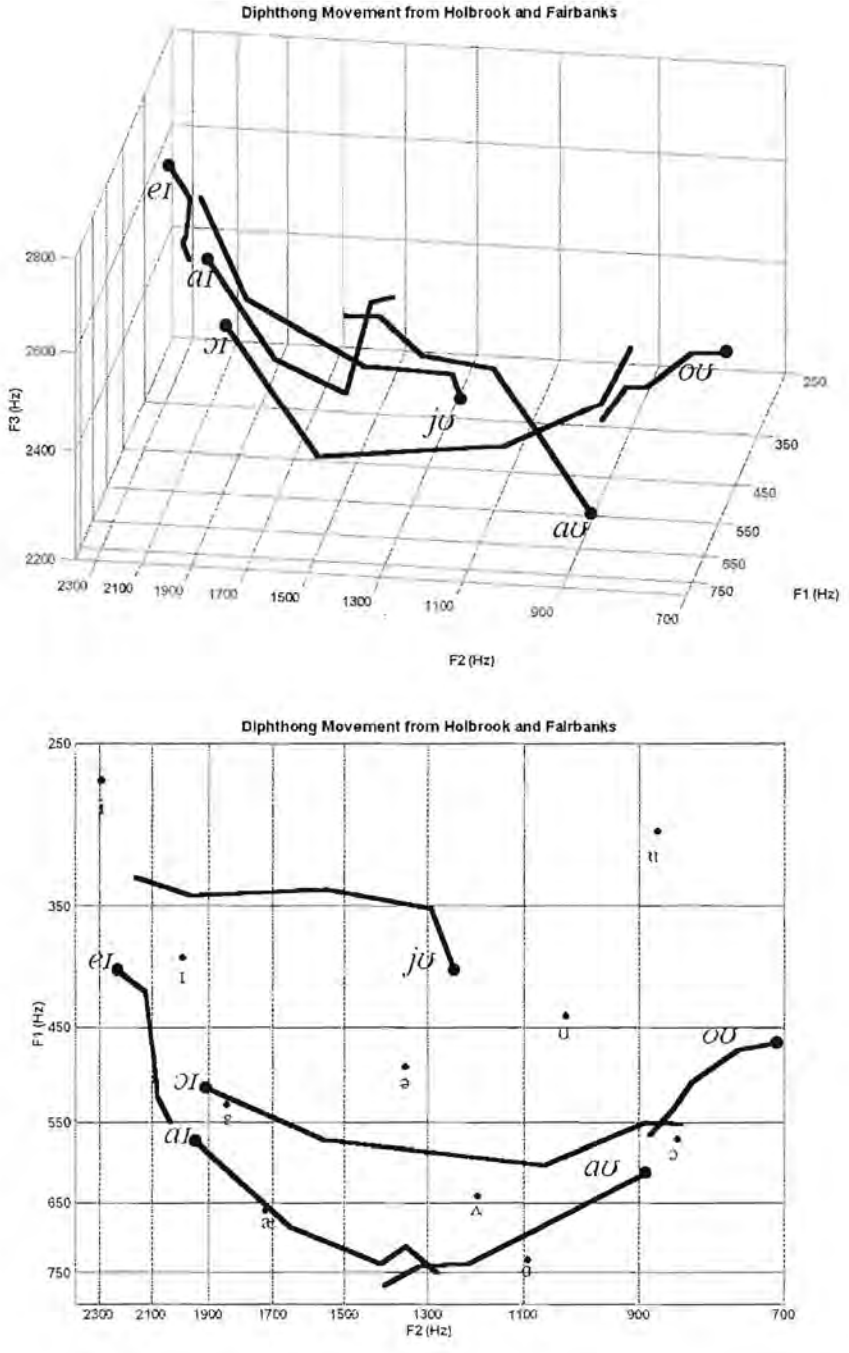


Figure 2.6: The top diagram indicates the Holbrook and Fairbanks[11] diphthongs and their movement in 3 dimensions (F1,F2,F3) and the bottom does likewise in the more traditional 2 dimensions (F1,F2). Also indicated on the bottom plot(as points) are the Peterson and Barney[10] vowels. The terminating point is indicated by a large node(•).

observing the spectral structure and dynamic change of the sound segment over time. This is achieved with the aid of the spectrogram.

Essentially, the spectrogram is a series of Fourier transforms taken over small, overlapping frames of data cut from the original data segment.

The Fourier transform is given as:

$$G(f) = \int_{-\infty}^{\infty} g(t)e^{-j2\pi ft}dt, \tag{2.1}$$

and the short time discrete Fourier transform of samples g_0 to g_{N-1} is:

$$G_k = \sum_{n=0}^{N-1} g_n e^{\frac{-j2\pi}{N}kn} \qquad k = 0, 1, \ldots, N-1 \tag{2.2}$$

If we plot these Fourier transforms vertically, line them up horizontally and then map colour to the magnitude of the spectrum, the image observed from above is the spectrogram. This process is displayed in Figure 2.7.

2.4 Formants

Formants are the resonance peaks of the vocal tract during speech production and they have been used by many researchers as the primary model of voiced speech for many years[2]. Formants can only be extracted (or only have meaning) for voiced speech, such as vowels, where distinct resonance patterns can be associated with a particular

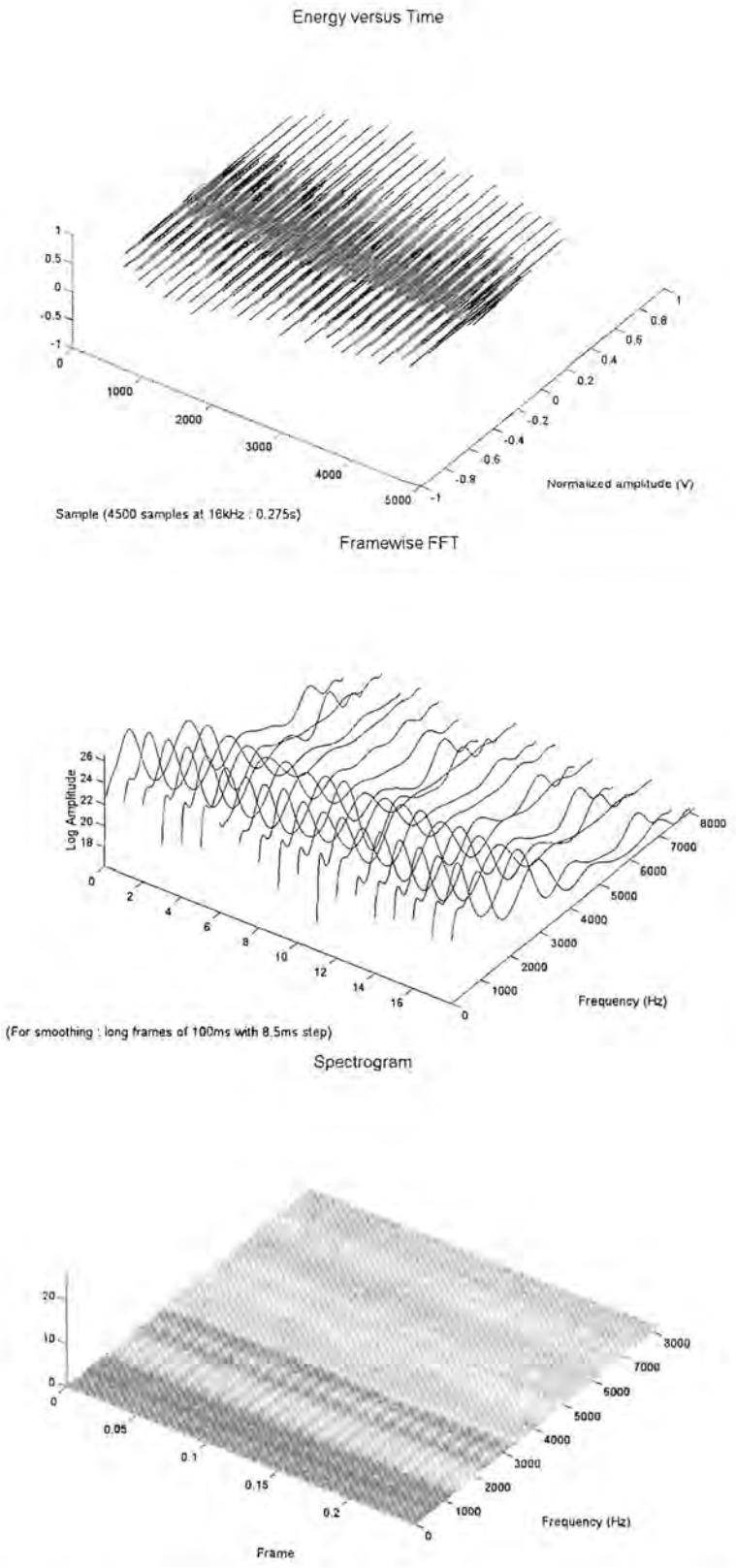


Figure 2.7: Spectrogram extraction illustrated. The time-energy signal is given at the top, the frame-wise spectrum in the middle and the colour-height mapped spectrogram at the bottom.

vowel sound[1]. For this reason they can not be used to analyse consonants (which have no distinct resonance structure which can be associated with any one particular consonant). Formants are used for voiced speech due to their many attractive features, some of which are:

- intuitiveness,
- robustness against channel noise and distortion,
- low dimensionality and hence easily perceived and analysed by a human,
- most immediate source of articulatory information and
- there is a close relation between formant parameters and model-based approaches to speech perception and production.

Formant extraction is the process of determining the most probable resonance frequencies corresponding to peaks in the frequency domain and calculating a temporal path to represent the vocal tract changes (resonant frequency changes) during speech production. It is in principle a very simple process (as will be demonstrated shortly), but it has proven to be complex enough in practice to warrant the efforts of continuing studies. Formant detection becomes a very complex task when the formants merge or lie very close to each other. Excessive noise and signal clipping also pose problems as the spectrum often becomes grossly distorted.

Perhaps the most simplistic means of formant extraction involves spectrum determination, polynomial fitting (or some other spectrum smoothing technique) and then peak picking. Visually this can be represented as in Figure 2.8.

More advanced techniques exist and are commonly used such as:

[1] Whispered speech is the exception to this rule, where, although not voiced, formants may still exist

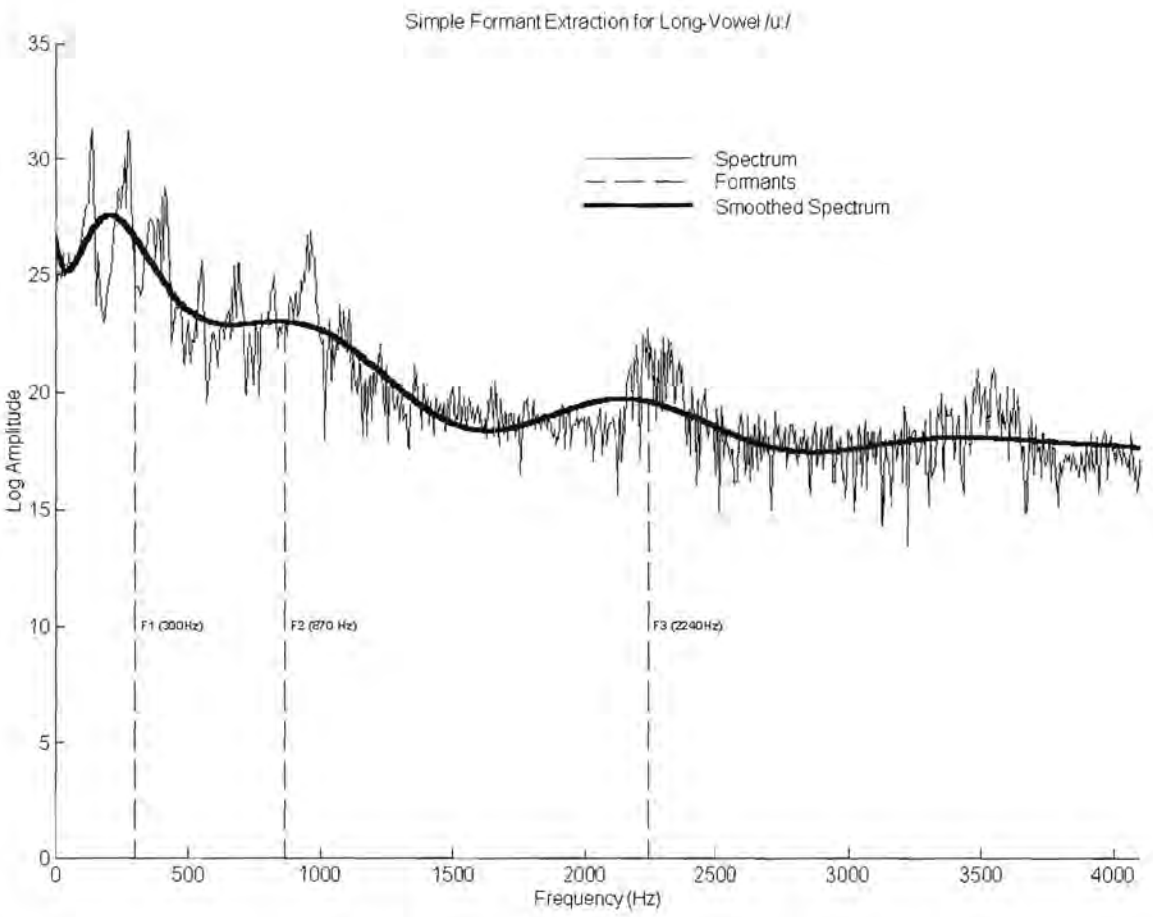


Figure 2.8: A smoothed Fourier Transform demonstrates how simple the concept of Formant extraction(in principle) is.

- Split Levinson algorithm [24]
- Linear prediction spectra [25][2]
- Gaussian mixture fitting [26]
- Contour integration [27]
- Digital resonators [28]

to name only a few.

The Split Levinson algorithm was developed by Delsarte and Genin[29] and requires about half as many computations to determine the LPCs as opposed to traditional techniques such as the Levinson[30] algorithm. The algorithm makes use of singular predictor polynomials to split the classic Levinson algorithm into 2 simpler algorithms.

Linear prediction techniques are explained in Section 2.4.1 as this is the technique we have chosen to use.

The Gaussian mixture fitting technique developed by Zolfaghari and Robinson[26] makes use of the Discrete Fourier Transform (DFT) and tries to fit a Gaussian mixture distribution to the magnitude spectrum. This is essentially an improvement on the basic peak picking technique described earlier in this section.

Snell and Milinazzo[27] developed an interesting technique for efficiently calculating roots within the unit circle once filter coefficients had already been determined using LPC techniques. By integrating over an arc of predetermined size it is possible to determine the presence of zeros within that arc and thereby, to arbitrary precision, it is possible to determine the location and number of roots. This in turn gives us the location of the formants.

A technique making use of decomposing the short-time power spectrum in segments has been proposed by Welling and Ney[28]. Each segment is modelled by a digital resonator

and the segment boundaries are then optimised using dynamic programming.

Each technique has its merits and failings. As a result of the many failings of these techniques, formant extraction, for accurate modelling purposes, must be an interactive process whereby the formants extracted must be verified manually and often recalculated using the various techniques mentioned above until satisfactory results are achieved.

This does not mean that the formants are recalculated until they fit the presupposed model of the researcher! This merely means that if the extracted formants are superimposed on a spectrogram and the results are seen to be flawed then recalculation may well be called for.

Holmes[31] has argued that formants may be used to significantly improve recognition in automatic speech recognition (ASR) systems by simply adding formant values and their accuracy probabilities to standard hidden Markov model (HMM) recognisers as additional features. Until recently, formants, although they have definite phonetic significance, have generally only been studied by linguists and largely been forgotten by speech recognition researchers. This is probably due to the complexity of reliable automatic formant extraction.

2.4.1 Linear prediction coefficients

In this study we have decided to use linear prediction coefficients (LPC) as our means of formant extraction. Various techniques were evaluated and compared on a subset of the data we have used in this study and LPC was found to extract the most correct formants most of the time.

The algorithm we found to perform almost as well as LPC concerning pitch extraction was the Split Levinson algorithm. Figure 2.9 demonstrates how formant algorithms may fail to locate the spectral peaks if forced to try and fit more formants than they can

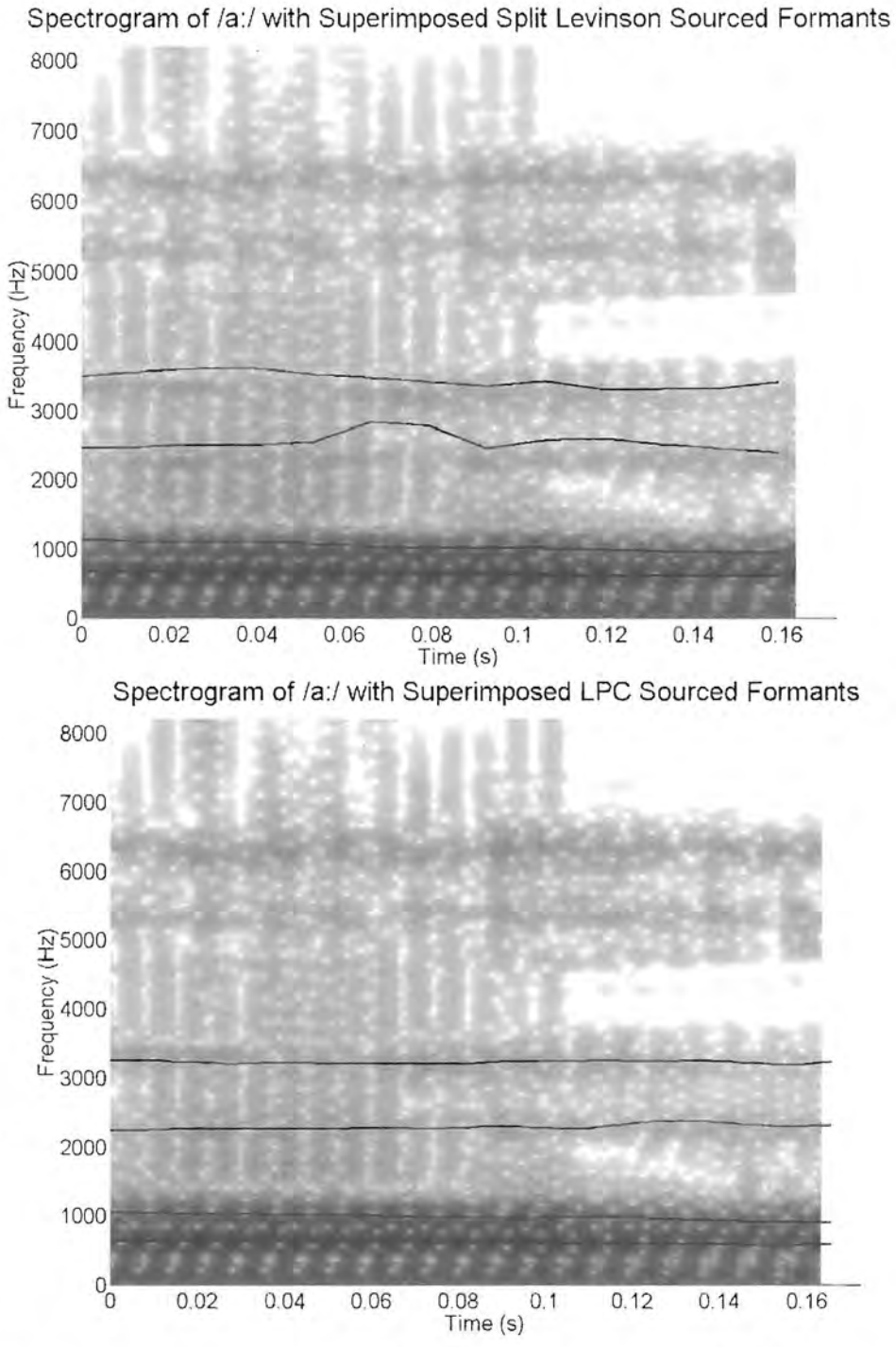


Figure 2.9: Formant extraction using Split Levinson (top) and LPC (bottom). We can see that in this case, LPC has managed to track the formants more accurately.

find. We see that forcing the Split Levinson algorithm to find 4 formants has resulted in incorrect placement of the formants (as shown by the black lines superimposed on the spectrogram). The LPC technique we decided to use is also prone to these errors, but was found to perform consistently well. Its formant extraction for the same piece of speech is shown by the black lines superimposed on the spectrogram in the bottom half of Figure 2.9.

LPC is based on the following principles[2]:

If there is no excitation, then the value of s_n (a speech sample at discrete time n) is correlated with the values of $s_{n-1}, s_{n-2}, ...s_{n-p}$ for some appropriate p. This is as a result of redundancy in the signal representation.

This correlation is due to the limits of how fast the vocal tract can move and change compared to f_s, the sampling frequency. We can therefore write:

$$s_n = f(s_{n-1}, ...s_{n-p}) + x_n. \tag{2.3}$$

where x_n denotes the excitation signal and we assume that x_n doesn't fit the correlation model that we're assuming for s_n.

We assume the f is a linear function of s_n, with p coefficients a_i, so:

$$s_n = \underbrace{\sum_{i=1}^{p} a_i s_{n-i}}_{f} + x_n. \tag{2.4}$$

For a short speech frame we may assume that the "filter" which generates the speech

from the source remains more or less constant. Using our assumption for s_n we have the z-transform of $H(z)$ of s_n given by:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}. \tag{2.5}$$

$H(z)$ has p poles, where the poles are real, or they are complex conjugate pairs; there are no zeros.

The prediction error is defined as:

$$e_n = s_n - \sum_{i=1}^{p} a_i s_{n-i}. \tag{2.6}$$

It is assumed that the error is due to the excitation since the models for excitation do not exhibit the correlation we're assuming for s_n.

For example, for voiced speech shorter or equal to one pitch cycle a simple model would be:

$$x_n = \begin{cases} 1 & at\ pitch\ pulse \\ 0 & elsewhere. \end{cases} \tag{2.7}$$

For unvoiced speech, x_n is modelled as noise, which is by definition uncorrelated.

The squared prediction error is defined as:

$$E = \sum_n e_n^2. \tag{2.8}$$

For the minimum error the partial derivative of E with respect to a_i is set equal to zero for each p, which gives p equations with p unknowns:

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-i} = \sum_n s_n s_{n-i}. \qquad 1 \leq i \leq p \tag{2.9}$$

and the range of n is dependent on the frame size. Then, using equations 2.6 and 2.8 and the a_k's from equation 2.9 we get:

$$E_{min} = \sum_n s_n^2 - \sum_{k=1}^{p} a_k \sum_n s_n s_{n-k}. \tag{2.10}$$

From this the a_k's (LPC's) still have to be determined. There are two ways to determine these: either an autocorrelation or cross-correlation based technique may be used. Each technique has its pros and cons.

For the autocorrelation technique:

- The disadvantages are:
    - The effect of the autocorrelation window (we need to correlate a windowed segment with itself) which must be used:
        - At beginning of the window non-zero values must be predicted from 0

values outside the window.

◇ At end of window, very small values must be predicted from larger values.

◇ Tapering of the signal due to the window leads to slight distortion.

- On the other hand, the advantages are:

  ○ The autocorrelation technique is computationally simple to perform:

  ◇ The matrix is symmetric, and on every diagonal, you get the same element. This is known as a "Toeplitz" matrix.

  ◇ Solution methods are fast - a_i's are calculated using an iterative method of $O(p^2)$, whereas general matrix inversion is of $O(p^3)$.

  ◇ The solution method is not sensitive numerically: can use fixed point (integer) math and the filter you get using the computed a_i's is guaranteed to be stable. Some methods find a_i's that correspond to poles outside the unit circle as an approximation to the true poles. This can't happen with the autocorrelation method.

For the cross-correlation technique:

- The disadvantages are:

  ○ The technique is computationally expensive:

  ◇ The number of computations is of $O(p^3)$ to solve for the a_i's.

  ◇ The technique is numerically sensitive.

  ◇ Can lead to unstable filters.

- On the other hand, the advantages are:

  ○ No distortion due to windowing as no Hamming window is used.

To solve using the autocorrelation technique we first define autocorrelation as:

$$R(i) \triangleq \sum_{n=-\infty}^{\infty} s_n s_{n+i} \tag{2.11}$$

which, if we substitute into equations 2.9 and 2.10 give us:

$$\sum_{k=1}^{p} a_k R(\mid i-k \mid) = R(i). \qquad 1 \leq i \leq p \tag{2.12}$$

and

$$E_{min} = R(0) - \sum_{k=1}^{p} a_k R(k). \tag{2.13}$$

Using the fact that the short term autocorrelation function $R_N(i)$ can be defined as:

$$R_N(i) = \sum_{n=0}^{N-i-1} s_n^{'} s_{n-i}^{'}. \qquad 0 \leq i \leq p \tag{2.14}$$

where: $s_n^{'}$ is the windowed s_n with w_n the windowing function, i.e.

$$s_n^{'} = \begin{cases} s_n w_n & 0 \leq n \leq N \\ 0 & elsewhere. \end{cases} \tag{2.15}$$

Equation 2.12 can be written in matrix form as:

$$\begin{bmatrix} R_N(0) & R_N(1) & \dots & R_N(p-1) \\ R_N(1) & R_N(0) & \dots & R_N(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_N(p-1) & R_N(p-2) & \dots & R_N(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_N(1) \\ R_N(2) \\ \vdots \\ R_N(p) \end{bmatrix} \tag{2.16}$$

Similarly, using cross-correlation we can also determine the LP coefficients. The mathematics is slightly more complex and computationally expensive, but as this technique is generally not used in speech-recognition systems and we have not used this technique we do not go into the details.

LPC is relatively simple to implement as can be demonstrated by a piece of Matlab code written by Levent Arslan[32] and quoted in Appendix A.1. The technique used by him is the autocorrelation technique with Durbin recursion and root finding.

The first requirement (when using the autocorrelation technique) is to find the autocorrelation coefficients and once these have been found, Durbin recursion may be used to calculate the LP coefficients, in other words, solve equation 2.16. Formant extraction then consists of the procedure of calculating the roots of the windowed frames of speech (Equation 2.4) and translating those roots into formant frequencies.

The algorithm in Appendix A.1 makes use of root finding which is relatively expensive computationally, although quite accurate. With modern computers the time spent determining the roots is becoming negligible, but with small devices this may still be an issue. If accuracy is not as important as timing (for example in real time speech communications) we may make use of various other techniques such as one suggested by Markel[2] where we evaluate the estimate of the vocal tract input response at various discrete points and then determine the peaks of the polynomial which fits these points.

We, however, did not use this technique.

Whichever technique we use, we can only expect about 85-90% accuracy for formants lower than 3kHz. This is still acceptable for male voices, but performance degrades significantly for female and child voices. A path tracking algorithm is therefore required to "join the dots" of the most probable of all the possible candidate formants we extract. The is achieved using a number of heuristics such as defining a maximum allowable frequency "jump" from frame to frame and observing that a similar number of peaks should keep appearing between troughs. Cost function techniques such as that used by Boersma[33] (and discussed in Section 3.3: Pitch Extraction) for pitch trajectory tracking may also be used to great effect.

2.5 Pitch

The pitch (also known as the fundamental frequency or F_0) is a very important characteristic to study when evaluating accent and pronunciation differences between language groups. Pitch is a voice characteristic which results from glottal closure and the frequency of this occurrence is known as the pitch of someone's voice. The intonation (or change in pitch with time) may vary greatly between languages, for example, French and Zulu are "musical" or "singing" languages (which results from a modulation of the pitch), Mandarin is a intonational language where a different meaning can be imparted to a word by changing the intonation (pitch). There are many such examples, but most importantly the intonation learnt carries over from a speaker's mother tongue to his second language, especially if the second language is learnt when the speaker is mature. Although from experience it is obvious, it is important to note that there is a great difference in pitch between male (low pitch), female (medium pitch) and child (high pitch) speakers. This implies that we must be careful when comparing the intonation of various speakers. This is one of the reasons why the study was restricted to male speakers of similar age. The effects of gender and

age have far reaching consequences such as poorly defined formants at higher pitched voices[34] and poor hidden Markov model recognition across gender data sets.

Various techniques exist for pitch extraction and there have been attempts to evaluate the effectiveness of these various algorithms[35].

We have already explained autocorrelation in Section 2.4.1 and we now follow up on this with how autocorrelation may be used for pitch extraction.

We have already defined the short-time or windowed autocorrelation function in equation 2.14 as:

$$R_N(i) = \sum_{n=0}^{N-i-1} s'_n s'_{n-i} \qquad 0 \leq \imath \leq p$$

So, if we evaluate $R_N(i)$ for i in the vicinity of $\frac{1}{F_0}$ (i.e. around a reasonable estimate for the inverse of the pitch) then we expect maxima at i=0, $\frac{1}{F_0}, \frac{2}{F_0}, \ldots$, and the pitch is $\frac{1}{F_0}$.

This is one of the oldest and most simple techniques of pitch extraction. This technique can be enhanced by filtering techniques.

Another technique which appears to work well under most situations is the CLIP or centre clipping pitch detection algorithm[35]. This involves pre-processing the speech frame s_n in an attempt to remove the formant information or minimise the vocal tract effects. This is done by low pass filtering the signal to 900 Hz.

We then set a clipping level C_L and centre clip the signal by only retaining samples which exceed $|C_L|$ by subtracting C_L for positive samples and adding C_L for negative samples.

The value of the autocorrelation function for a range of lags using the centre clipped signal is then calculated. The autocorrelation function is then searched for the maximum normalised value and (generally) if it exceeds 0.3 the section is considered voiced and the pitch period is determined from the location of the maximum. We have not used the CLIP technique as experiments by Rabiner et al.[35] seem to indicate that CLIP does not perform as well as LPC techniques, especially on low pitched voices such as male voices (which is what our database is made up of).

2.6 Equivalence classification

The theory of equivalence classification is that all speakers[2] of a certain region or socio-economic grouping, tend to possess equivalent phoneme sets as long as they have resided in that area while learning the language as a child. The theory states that speakers learning a new language at a late age tend to use the phonemes they already know from the first language, to pronounce the words in the new language. This type of study has generally been performed on populations where this is easily determinable, for example, by studying adults who immigrated into a region at various ages, and then studying their phone structures. James Flege has performed many studies on groups like: Italians who had immigrated to America[36] and French speakers living in Canada with various levels of learning immersion at different ages[18].

The age of learning (AOL) has proven to be a critical factor in the phone make-up of speakers. Our study differs significantly from Flege's research in the fact that most white South Africans are familiar with both English and Afrikaans through media such as the radio and television. This is especially true for young first language Afrikaans speakers who may have watched a large amount of British and especially American television series while growing up. The reverse is not necessarily true for young first language English speakers who may not have watched much Afrikaans television. This

[2]Excluding speakers with pathological speech problems.

trend will continue to grow as fewer programs are translated into Afrikaans and as English channels such as subscription and satellite television become more prominent.

It would be inappropriate to make any deductions from the research by Flege on the phone make-up of speakers in a multi-lingual society such as South Africa's. We would assume that if speakers learn multiple languages at a young age that they would be capable of producing native phones for each of those languages. This is in fact confirmed when we hear many young South African children from multi-lingual families switching between languages. This of course complicates our study and we have therefore asked the speakers in our database to ascertain their own fluency in each of the two languages in question (see Figure 3.1 on page 52).

We find that most of the speakers consider themselves to be fairly bilingual. This makes it far more difficult to determine the acoustic differences between the two language groups. If, however, differences are observable with such a marginal group then it bodes well on further research into groups which we know will be acoustically more separated.

2.7 Cubic splines

The dynamic features of the vowels and diphthongs, namely the diphthong formants and pitch contours, have been analysed using curve fitting techniques. In particular, the cubic spline has been used to achieve this[37]. The use of curve fitting is justified by the need to compare the dynamic pitch and formant trajectories. This can not be performed at a point wise level due to the semi-instantaneous jumps which are a result of pitch and formant extraction algorithm shortcomings. These small jumps would unfairly boost the variance of the trajectory and cause analysis of variance tests to judge even similar trajectories as different. We therefore fit a curve to the general trend of the pitch or formant trajectory.

The curve fitting is done in the following way:

- We fit the data using a cubic spline such that the spline fits through every one of the data points we have extracted for the formants or pitch trajectories.
- We resample this cubic spline to give us 128 samples, irrespective of how many points we had originally. We now have a linear time-scaled curve of normalised length.

We now want to reduce this into a simple, low-order dimensionality vector for comparison purposes. It was decided to do this by calculating four points which if fitted by a curve would represent reasonably closely the trajectory we began with. So:

- We make the first of the 128 points the first point of the curve. Formant extraction tends to be difficult at the start and end of voiced speech segments. This is why it is important that we make sure that the formant extraction is correct as described in Section 3.3.1 on page 58.
- We divide the 128 point formant or pitch trajectory into three equal sections (as can be seen in Figure 2.10 (middle)). The second and third points are then calculated as being the mean formant or pitch values of a section centred around the first and second divisions (as seen by the horizontal bars in the middle figure).
- The fourth point is equal to the last of the 128 point vector.

We once again perform a cubic spline fit, this time fitting just the four points we have calculated. If we define the cubic spline by:

$$S_k(x) = a_{k,3}(x - x_k)^3 + a_{k,2}(x - x_k)^2 + a_{k,1}(x - x_k) + a_{k,0} \tag{2.17}$$

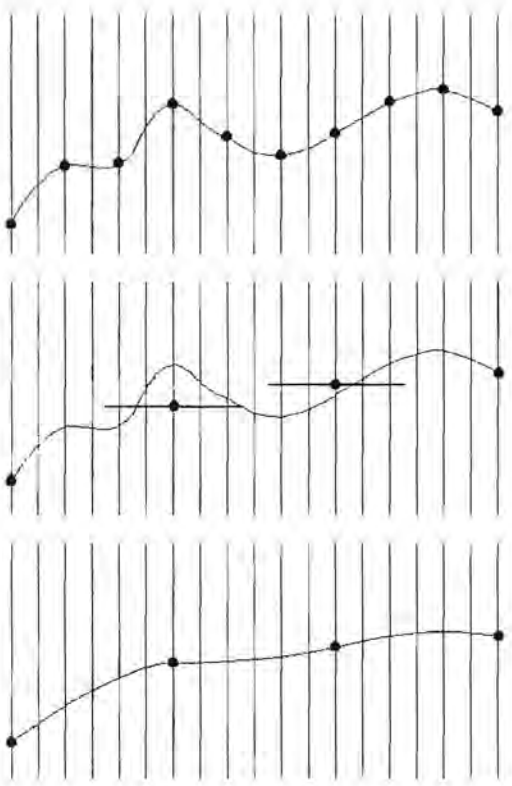


Figure 2.10: Reduction of a multi-dimensional formant or pitch trajectory to a low-dimension cubic-spline for ANOVA comparison purposes.

then, as we are fitting three sections and we have four coefficients per section, we end up with twelve coefficients per formant or pitch trajectory. We are now able to perform ANOVA tests of significance between trajectories of various speakers and using mean trajectories, between the two accent groupings. We have chosen to work with three formants and one pitch trajectory. As it carries little perceptual information to give the exact coefficient which was found to be significantly different, we simply display whether or not we found significant differences within a trajectory. This is displayed in the results tables in Section 3.4 by using dark gray boxes. The magnitude of this difference could only be estimated in an artificial way which we have decided to avoid as we have deemed it sufficient to demonstrate that there is a significant difference between the two language groups. The magnitude of this difference can then be judged by the reader from the trajectory plots, remembering of course that the plots are just mean plots and carry no variance information.

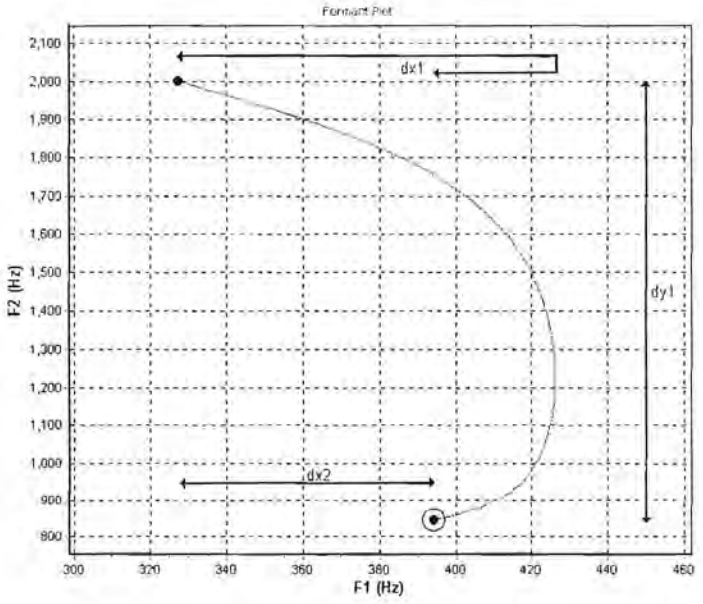


Figure 2.11: We have used two measures of diphthongization. The first is the displacement along the axis between the initial and terminating points(dx2 and dy1) and the second is the cumulative absolute distance traversed along the axis(dx1 where the cumulative absolute value of the arrowed distance is used).

2.8 Diphthongization

We have discussed diphthongs in Section 2.2. We have also discussed the cubic spline in Section 2.7. We can use the cubic spline to form a low order representation of the diphthong formant trajectories. Using this principle we can measure the frequency displacement a diphthong undergoes while moving from the initial "vowel" to the terminating "vowel". We have decided on two measurements of diphthongization, and these are demonstrated in Figure 2.11. We have included the nett formant displacement (dx2 and dy1) and the gross formant displacement (dx1) as our diphthongization metrics.

2.9 Statistics: Tests of hypotheses and significance

Using the notation of Spiegel[38] we state that in statistics we may define a null hypothesis (denoted H_0) which may be used to test the structure of given populations. We may for example make the null hypothesis that the means of the formants for English first language and Afrikaans first language speakers are statistically equal for certain vowel sounds. We may then apply various statistical tests to confirm or deny our hypotheses.

There are two types of errors. Type I errors occur when we reject a hypothesis we should have accepted and Type II errors are said to occur when we accept a hypothesis we should have rejected. Unfortunately we find that when we attempt to minimise Type I errors we ultimately increase our probability of making Type II errors and vice-versa. Usually one of the error types is more critical and this must be taken into account when we define the hypothesis.

The maximum probability with which we are willing to risk a Type I error is called the *level of significance*. We usually specify a level of significance of 0.01 or 0.05. A 0.01 significance level indicates that we are 99% confident that we have made the right decision.

2.9.1 Analysis of variance (ANOVA) test

Fisher[19] developed and used the F distribution to perform "analysis of variance" tests on two or more populations (independent groups of samples).

If x is a sample, then the total variation (variance) of x is defined as:

$$v = \sum_{j,k} x_{jk}^2 - \frac{\tau^2}{n} \tag{2.18}$$

where $j = 1, 2, \ldots, a$ is the number of independent groups (in the sample) of $k = 1, 2, \ldots, b$ measurements each. The variation between the a independent groups is:

$$v_b = \sum_{j} \frac{\tau_j^2}{n_j} - \frac{\tau^2}{n} \tag{2.19}$$

where:

$$\tau = \sum_{j,k} x_{jk} \qquad \text{the total of all the values } x_{jk} \tag{2.20}$$

and

$$\tau_j = \sum_{k} x_{jk} \qquad \text{is the total of the values in the } j^{th} \text{ independent group.} \tag{2.21}$$

Also,

$$n = \sum_{j} n_j \qquad \text{is the total number of observations in all the independent groups} \tag{2.22}$$

where n_j is the number of observations in the j^{th} independent group.

Variation	Degrees of Freedom	Mean Square	F
Between groups, $v_b = \sum_j n_j(\overline{x}_j - \overline{x})^2$	$a-1$	$\hat{S}_b^2 = \frac{v_b}{a-1}$	$\frac{\hat{S}_b^2}{\hat{S}_w^2}$ with $a-1, n-a$ degrees of freedom
Within groups, $v_w = v - v_b$	$n-a$	$\hat{S}_w^2 = \frac{v_w}{n-a}$	
Total, $v = v_b + v_w = \sum_{j,k}(x_{jk} - \overline{x})^2$	$n-1$		

Table 2.5: Analysis of Variance Table

If the group means are not equal i.e. the null hypothesis (H_0) is not true then we can

expect $\hat{S}_b^2$ to be greater than the variance ($\sigma^2 = \sum(x-\mu)^2 f(x)$) and this becomes larger as the difference in means increases. We also know that $\hat{S}_w^2$ (which is given in Table 2.5 and is an unbiased estimate of σ^2) is always equal to σ^2 irrespective of mean differences. It seems therefore that a good statistic for testing H_0 is $\frac{\hat{S}_b^2}{\hat{S}_w^2}$ which we call F in Table 2.5 where a is the number of groups measured. The distribution of this statistic is known as the F distribution in honour of Sir Ronald Fisher.

The calculations required to perform an analysis of variance test are often summarised in tabular form as in Table 2.5. In practice we calculate v and v_b and then deduce v_w. The $\overline{x}$ indicated in the table means the mean value of x. There are $a-1$ degrees of freedom (dimensional elements) between groups and $n-a$ degrees of freedom within the groups. Notice that these formulas are the same as those in the functions mentioned above, with substitutions having been performed and more compact notation being used.

Of course, as we are simply comparing Afrikaans and English, a (the number of independent groups) is only 2 which allows us to simplify things, but for generality we have described a complete analysis of variance, where an analysis of variance consists of calculating the F ratio. To determine whether a particular F ratio indicates a significant difference in means for a particular significance level, we generally use the the F distribution tables published in Fisher's book[19]. The exact value which the F ratio must exceed to indicate a significant difference is dependent on the degrees of freedom i.e. the number of treatments and the total number of observations.

Chapter 3

Experiments

This chapter describes the experiments performed on the data described in Chapter 2. The objectives (in other words, what we are trying to achieve) are described and then the techniques used to meet these objectives are explained. Finally we discuss the results obtained, show graphs of the processed data and discuss our interpretation of the experimental results.

3.1 Objectives

Our primary objective with this study is to create acoustic models of Afrikaans vowels and diphthongs as spoken by mother-tongue speakers. We then want to create acoustic models of Afrikaans vowels and diphthongs for mother-tongue English speakers and compare these models with the Afrikaans models. We would then like to determine whether there are significant differences between the two accent groups.

A further objective which follows from the first is to add South African English vowels and diphthongs to the models and also compare these with the Afrikaans models of the same sounds. This will help us to determine how much of an influence Afrikaans