# An acoustic comparison of the vowels and diphthongs of first-language and African-mother-tongue South African English

by

**Janus Daniël Brink**

Submitted in partial fulfilment of the requirements for the degree
Master of Engineering (Computer Engineering)
in the Faculty of Engineering
UNIVERSITY OF PRETORIA

April 2002

# Abstract

Speaker accent influences the accuracy of automatic speech recognition (ASR) systems. Knowledge of accent based acoustic variations can therefore be used in the development of more robust systems. This project investigates the differences between first language (L1) and second language (L2) English in South Africa with respect to vowels and diphthongs. The study is specifically aimed at L2 English speakers with a native African mother tongue, for instance speakers of isi-Zulu, isi-Xhosa, Tswana or South Sotho. The vowel systems of English and African languages, as described in the linguistic literature, are compared to predict the expected deviations of L2 South African English from L1.

A number of vowels and diphthongs from L1 and L2 speakers are acoustically compared and the results are correlated with the linguistic predictions. The comparison is firstly made in formant space using the first three formants found using the Split Levinson algorithm. The L1 vowel centroids and diphthong trajectories in this three-dimensional space are then compared to their L2 counterparts using analysis of variance.

The second analysis method is based on simple hidden Markov models (HMMs) using Mel-scaled cepstral features. Each HMM models a vowel or diphthong from one of the two speaker groups and analysis of variance is again used to compare the L1 and L2 HMMs.

Significant differences are found in the vowel and diphthong qualities of the two lan-

guage groups which supports the linguistically predicted effects such as vowel substitution, peripheralisation and changes in diphthong strength.

The long-term goal of this project is to enable the adaptation of existing L1 English recognition systems to perform equally well on South African L2 English.

**Keywords**: Speech recognition, accent classification, acoustic analysis, formants, cepstra, vowels, diphthongs, second language South African English, African mother tongue.

# Uittreksel

Sprekeraksent beïnvloed die akkuraatheid van outomatiese spraakherkenningstelsels (OSH). Kennis van die akoestiese variasies as gevolg van aksent of dialek verskille vanaf die moedertaal kan dus gebruik word in die ontwikkeling van meer robuuste stelsels. Hierdie projek ondersoek die verskille in uitspraak tussen eerstetaal (L1) en tweedetaal (L2) Engels in Suid Afrika met spesifieke verwysing na vokale en diftonge. Die studie is spesifiek toegespits op L2 sprekers met Afrika moedertale, byvoorbeeld sprekers van isi-Zulu, isi-Xhosa, Tswana of Suid-Sotho (Sesotho). Die vokaalstelsels van hierdie Afrikatale word vergelyk met dié van L1 Engels soos wat dit beskryf word in die taalkundige literatuur om verwagte verskille in uispraak te voorspel.

'n Aantal vokale en diftonge van die twee moedertaalgroepe is akoesties vergelyk en geanaliseer in beide die formant- en Mel-geskaleerde kepstrum ruimtes, waarvolgens gemete verskille vergelyk is met die taalkundige voorspellings. Die vergelyking word eerstens in formantruimte getref waar die eerste drie formante gebruik is, soos onttrek deur die Split Levinson metode. Die eerstetaal vokaal gemiddelde posisie en diftong trajekte is dan vergelyk met dié van tweedetaal sprekers deur middel van variansie analise.

Tweedens is eenvoudige verskuilde Markov modelle gebruik met kepstrale koeffisiënte as kenmerke. Elke model stel 'n vokaal of diftong voor in een van die twee spreker groepe. Variansie analise is dan weereens gebruik om hierdie modelle te vergelyk.

Beduidende verskille tussen die L1 en L2 Engelse klanke is gevind wat die taalkundig-voorspelde effekte soos vokaal vervanging, meer periferiële uitspraak en veranderings in diftongsterkte ondersteun en bevestig.

Die langtermyndoelwit van hierdie projek is om die aanpassing van bestaande L1 Engelse herkenningstelsels moontlik te maak sodat hulle net so goed presteer op L2 Suid-Afrikaanse Engels.

**Sleutelwoorde**: Spraakherkenning, aksentklassifikasie, akoestiese analise, formante, kepstra, vokale, diftonge, tweedetaal Suid-Afrikaanse Engels, Afrika moedertale.

# Acknowledgements

I would like to thank the following people for their help and support, without which this project would not have been possible:

- Professor Liesbeth Botha, my supervisor for this and many other projects.
- My wife, Anita, for all her support and especially for her help with an immense data collection task.

# Contents

# Abbreviations

| | |
|---|---|
| **ANOVA** | Analysis of Variance |
| **ASR** | Automatic Speech Recognition |
| **BSAE** | Black South African English |
| **DOF** | Degrees of Freedom |
| **EAE** | East African English |
| **HMM** | Hidden Markov Model |
| **L1** | First Language |
| **L2** | Second Language |
| **LPC** | Linear Predictive Coding |
| **MFCC** | Mel-frequency scaled Cepstral Coefficients |
| **PCM** | Pulse Coded Modulation |
| **RMS** | Root Mean Square |
| **RP** | Respectable Pronunciation |
| **SAE** | South African English |
| **SPP** | Singularity Predictor Polynomial |
| **TE** | Tswana English |
| **WSAE** | White South African English |
| **ZE** | Zulu English |

# Chapter 1

# Introduction

## 1.1 Introduction

People communicate in many different ways, of which speech is certainly the most important. People are used to communicating with one another using verbal interaction as a fast, efficient and robust communications medium. It is therefore natural to wish to extend this means of interaction to a wider spectrum of listeners, such as computers and other automated systems. This was made possible by the development of automatic speech recognition (ASR) systems, which enables machines to interpret spoken language, creating a more efficient and natural man-machine interface.

The performance of ASR systems have greatly improved in recent years, but their performance is still affected by many factors. One of these is variation in dialect, or accent, of the language used to interact with the system, compared to the speech data on which the system was trained [1, 2]. Accented speech causes changes in acoustic quality of the speech sounds used in communication, while dialects include new vocabulary, which is often specific to a particular region or ethnic group. When a language is used as a common communications medium between speakers of different mother-tongues,

these native languages often have a marked influence on the pronunciation of a second language [3, 4].

Numerous cultural and ethnic groups are present in South Africa and even though 11 official languages have been declared, English is still used as the main language for business and public communication. The pronunciation of second language (L2) English therefore varies significantly from the first language (L1) norm. A study was therefore performed to test if the acoustic differences between first and second language English, as predicted by the linguistic literature [5, 6, 7, 8], can be quantitatively measured using speech processing techniques and acoustic analysis.

The second language speakers of interest were native speakers of African languages, for instance mother tongue isi-Zulu, isi-Xhosa, Sesotho (Southern Sotho) or Tswana speakers, since these accents tend to be of the most persistent, i.e. non-standard pronunciations are retained even in acrolectal (more educated) second language speakers [3].

This study first compares the phonetic structure of these languages with that of English and then investigates the relevance of general linguistic predictions as to second language English learning within the African native language context. These linguistic findings are then compared to experimental results obtained from a detailed acoustic and spectral analysis (including Mel-scaled cepstral coefficients and formants) of the vowels and diphthongs of both first and second language South African English. More details on the approach used for this analysis are discussed below.

## 1.2 Approach

In order to test the linguistic predictions mentioned above, speech data was required from South African English speakers, some speaking first and others second language English. Enough speakers per language group were needed to ensure that we are

not merely measuring speaker dependent variation. To eliminate even more variation within a given language group, it was decided to compare phonemes originating from the same words where possible. The data was recorded from local television and subsequently digitized and stored on computer for analysis.

Five types of experiments were conducted:

1. Vowels were compared using formant analysis, thereby directly comparing phonemes from one language group to that of the other in a measurement space closely aligned with the linguistic representation, viz. two-dimensional formant-plots compare well with phonetic vowel charts.

2. Diphthongs were compared using the same formant approach, but since diphthongs are dynamic in nature, the extent of movement within formant space needs to be captured. This was accomplished by using spline interpolation of the formant values over the duration of the phoneme, followed by a comparison of the spline coefficients for the two language groups.

3. To assess the impact of these accents on automatic speech recognition (ASR) systems, the Mel-frequency scaled cepstral coefficients (MFCCs) of the vowels of each language group were used to train hidden Markov models (HMMs) for each context-dependent phoneme in each language group. The parameters of each pair of HMMs were then compared.

4. The same Mel-scaled cepstral analysis was repeated for the diphthongs, using multi-state HMMs.

5. Finally, diphthongisation of vowels and vice-versa (monophthongisation of diphthongs) were examined. The former occurs when a phoneme classified in first language (L1) English as a vowel or weak diphthong, is uttered as a pronounced diphthong in the second language (L2) accent. In the latter case an L1 diphthong is weakened or altogether simplified as a vowel in the L2 pronunciation.

## 1.3 Goals and contributions

One aim of this study is to determine if the perceptual differences between L1 and L2 South African English can be quantitatively measured. This requires the determination of suitable and prominent acoustic parameters that describe the differences between L1 and L2 South African English vowels and diphthongs. Diphthongs need to be modelled in a way that accurately reflects their dynamic nature and enables direct comparison between accent groups. The primary goal however, is to form a statistical model of the acoustic differences in pronunciation of several speech sounds in L2 English as compared to their L1 counterparts with the long term goal of improving the performance of ASR systems. This is accomplished by determining how the differences found using acoustic analysis translate to differences in the phoneme models used for speech recognition.

This work contributes to the creation of a continuous speech database of South African English and indicates some of the acoustic differences between first and second language South African English [9]. The use of accent dependent models for robust English speech recognition systems in South Africa depends on accurate knowledge of the acoustic differences between first and second language pronunciation. The influence of these differences on Mel-scaled cepstral based vowel and diphthong models are therefore illustrated to aid in the adaptation of such systems [10, 11].

## 1.4 Organisation

Chapter 2 gives the theoretical background for this study. Section 2.1 examines some linguistic background and related work on the English vowel system as compared to that of the African languages and makes some predictions as to what can be expected in African second language English. The speech processing techniques used in analysing the data are outlined in Section 2.2.

Chapter 3 describes the data and experiments. Section 3.2 describes how data was obtained and how it was subsequently structured to enable logical analysis. Five types of experiments were conducted: Firstly vowels were compared using formant analysis, as described in Section 3.3. Secondly diphthongs were compared using the same approach, with additional spline interpolation of the formant tracks in Section 3.4. As a third experiment hidden Markov models were trained on the mel-scaled cepstra of the vowels and on the diphthongs (fourth experiment) of each language group and compared as detailed in Sections 3.5 and 3.6. Finally, in the fifth experiment, diphthongisation of vowels and vice-versa (monophthongisation of diphthongs) were examined as described in Section 3.7.

A global overview of the results and concluding remarks are given in Chapter 4.

# Chapter 2

# Background theory

In this chapter the theoretical background for this study is given. A brief introduction to the linguistics and phonetics relating to the vowels and diphthongs of English is given and the classification of African languages in Southern Africa is discussed. The vowel system of English is compared to that of some of the African languages to gain insight into the framework used by the second language English student.

This is followed by a summary of the findings of some related studies to define the linguistic and acoustic framework for South African English. We also look at previous work on the characteristics of African forms of English and discuss some relevant language accent classification methods.

An overview of the speech processing techniques used is given in order to clarify the approach used in the experiments and to familiarise the reader with the relevant feature spaces. A brief description is given of formant analysis, cubic splines, mel-scaled cepstra, hidden Markov models and analysis of variance.

## 2.1 Linguistic issues

This section deals with the linguistic issues involved in studying the accents of English in South Africa, specifically with respect to accents influenced by mother tongue speakers of native African languages. Section 2.1.1 looks at the phonetic structure of the English vowel system, which forms the basis for comparison of the dialects of English, while Section 2.1.2 looks into the classification of the African languages found in South Africa from both the linguistic and the purely phonetic perspectives. The vowel structures of these languages are compared to that of English to get insight into the process of second language English learning, which is further explored in Section 2.1.3.

Before continuing with linguistic, phonemic and phonetic analysis, it would be prudent to briefly define these fields and clarify the relationship between them.

Linguistics is the study of the high-level morphology (form) and functioning of language, it defines the broader environment within which languages exist and explores their grammatical structure. Phonology, or phonemics, represents the study of phonemes within the language [12], contrasting sounds that give meaning to thoughts, words and sentences. Phonemics help bridge the gap from linguistics to the physical formation and characteristics of the basic sound units in speech as studied in phonetics [13]. Although it is here, at the phonetic level, where the physical measurements are made, the results must be viewed in their phonemic and linguistic contexts to enable a meaningful and systematic study of accents and language.

### 2.1.1 Phonetics and the English language

Phonetics is the study of the basic sound-units forming the basis of spoken language [14, 15]. These sound-units, or phones, are the building blocks for syllables which, in turn, are strung together to from words, phrases and sentences. Phones may be categorised into four broad sound classes: vowels, diphthongs, semivowels and consonants. Vowels

and diphthongs are relatively long duration voiced sounds, i.e. produced with vibrating vocal folds and thus have well defined spectral characteristics [16]. Consonants tend to be better defined in the temporal domain and are usually indicated by sudden changes in sound characteristics between vowels. While vowels remain fairly constant in sound quality throughout the duration of the phoneme, diphthongs are formed by the smooth transition from (or near) one vowel position to another [17]. Semivowels are also voiced sounds, but tend to experience a more constricted air flow through the mouth and often have a nasal quality.

Although many schemes have been proposed to represent vowel phone-space [13, 18], the most popular one today is the framework introduced and developed by Daniel Jones [19] and which was adopted by the *International Phonetics Association* (I.P.A.). The English vowels can thus be represented according to this I.P.A framework as shown in Figure 2.1.

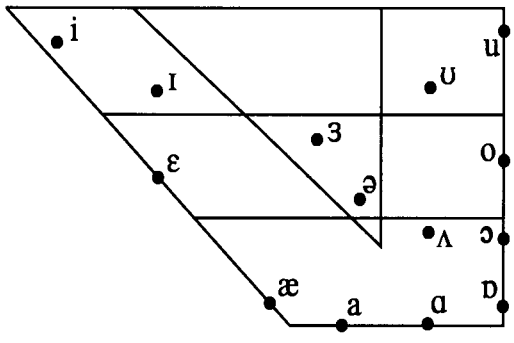


Figure 2.1: Basic vowel system of English (from [14]).

The dots in the figure indicate the positions on the I.P.A. chart of the vowels (phonemes) used in standard English. While word pronunciations differ according to dialect, Tables 2.1 and 2.2 show a few examples to give an indication of how these vowels and diphthongs are articulated.

The "high" vowels such as /i/ and /u/ are pronounced with a higher tongue position in the mouth, while "low" vowels such as /a/, /ɑ/ and /ɔ/ are pronounced with a lower tongue position. The arch of the tongue, i.e. the highest point of the tongue, is also

| *Vowel* | *Example* | *Vowel* | *Example* |
|---|---|---|---|
| [ɪ] | h*i*t | [ʌ] | c*u*t |
| [i] | h*ea*t | [ɒ] | c*o*t |
| [ɛ] | b*e*d | [ɑ] | c*a*rt |
| [ɜ] | b*ir*d | [ɔ] | c*augh*t |
| [ə] | *a*bout | [ʊ] | p*u*ll |
| [æ] | b*a*d | [u] | p*oo*l |

Table 2.1: Articulation of vowels in English (from [17]).

| *Diphthong* | *Example* | *Diphthong* | *Example* |
|---|---|---|---|
| [aɪ] | b*uy* | [aʊ] | d*ow*n |
| [eɪ] | b*ai*t | [əʊ] | b*oa*t |
| [ɔɪ] | b*oy* | | |

Table 2.2: Articulation of diphthongs in English (from [17]).

frequently used to describe vowels. The "front" vowels (/i/, /ɛ/, /æ/) are articulated with the tongue arched in the front of the mouth, while "back" vowels (/ɒ/, /ɔ/, /o/, /u/) are produced with the tongue arched more to the back.

### 2.1.2 South African native languages

In this section a classification of the African languages found in Southern Africa is portrayed. Firstly a linguistic scheme is given, followed by a phonetic classification based on the languages' vowel systems.

### Linguistic classification

This section is offered as additional background in order to place the second language English speakers used in this study in context. In the linguistic scheme, the term 'group' indicates an aggregation of languages possessing similar phonetic and grammatical features, where members of different languages within the group can often understand one another without too much effort. The term 'cluster' indicates a grouping of dialects stemming from the same literary form [5].

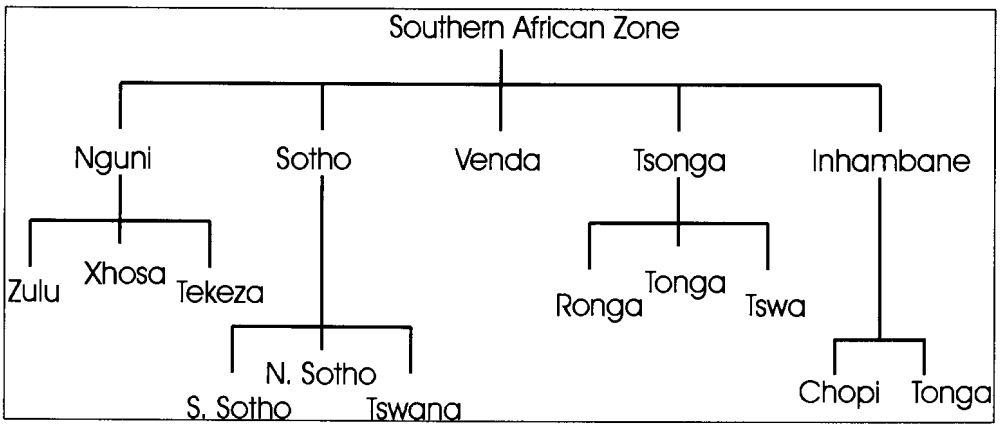


Figure 2.2: Classification of African languages in Southern Africa (from [20]).

In Figure 2.2 the language groups found in South Africa are shown. The most prominent among them are the Nguni and the Sotho groups. The Nguni group contains the Zulu, Xhosa and Tekeza language clusters. The Sotho group includes the Northern and Southern Sotho, as well as the Tswana clusters. The Tsonga group have constituents Ronda, Tonga and Tswa, the Inhambane groups contains Chopi and another variation of Tonga, while Venda does not show much dialectal variation.

The South African dialects of the language clusters in Nguni are depicted in Figure 2.3, and can be briefly outlined as follows: The Zulu cluster includes Zulu (originally from Natal), Ndebele (found in the former Transvaal) as major members. The Xhosa cluster contains a number of dialects, but the literary form is based on the Gcaleka-Ndlambe-Gaika group of dialects. The notable dialect of the Tekeza cluster is Swazi, which is spoken in Swaziland and Mpumalanga (former Eastern Transvaal), while Phuti is

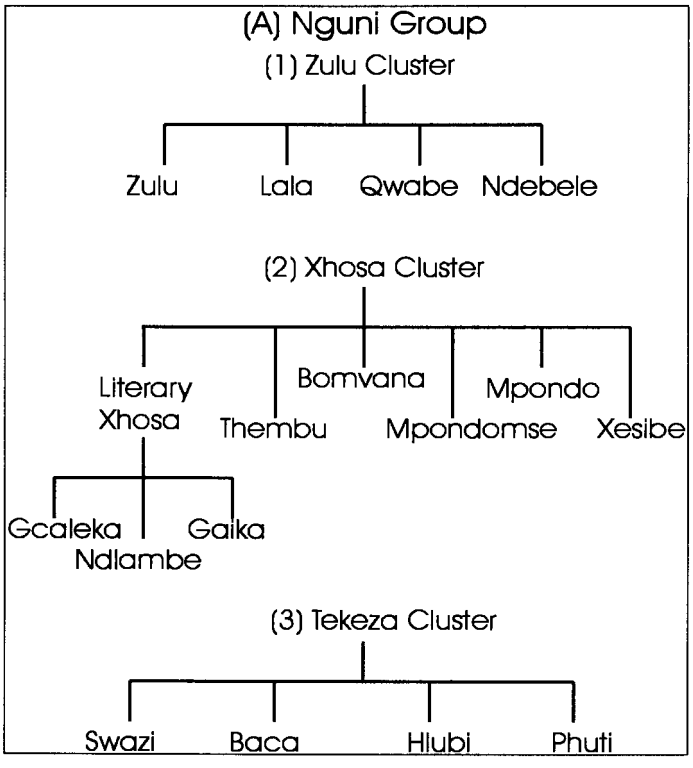


Figure 2.3: Clusters and dialects of the Nguni group (from [20]).

heavily influenced by Sotho.

The Sotho group contains three clusters (Figure 2.4), namely Southern Sotho (Sesotho), Northern Sotho (Sepedi) and Tswana. Southern Sotho contains the Southern Sotho dialect found in the Free State on which the Sotho literature is based, as well as Lozi, which is spoken in Lesotho. The Northern Sotho cluster, spoken in the former Transvaal contains a number of dialects including Ndebele-Sotho and Pedi. Finally the Tswana cluster contains a multitude of dialects spoken from the North West (former Western Transvaal), the Northern Cape to the Free State.

Venda does not contain any distinctive dialectal variants, while the Tsonga group contains the Ronga, Tonga and Tswa clusters each with their dialects (Figure 2.5), while the Inhambane group contains the Chopi and Tonga clusters.

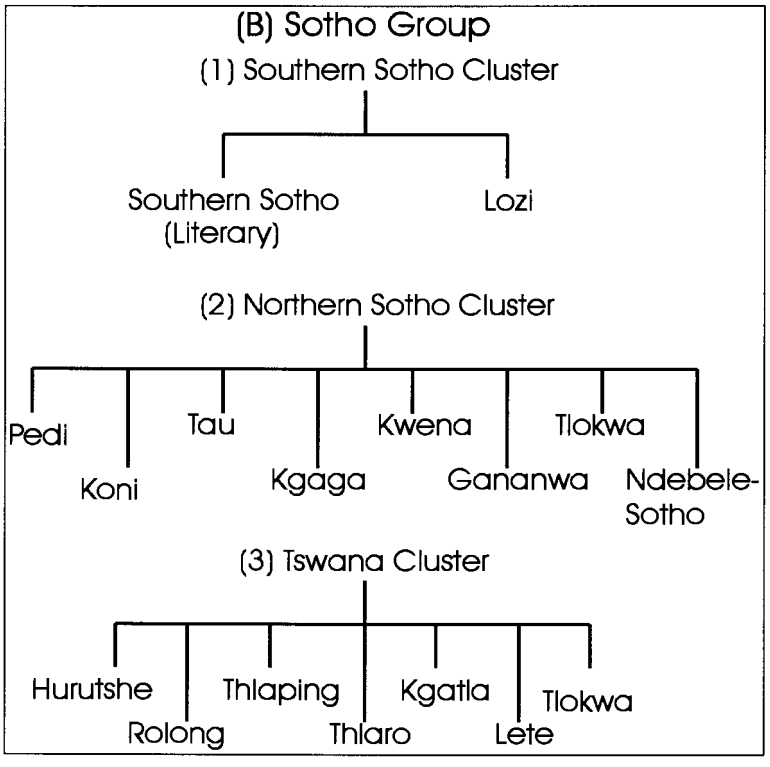


Figure 2.4: Clusters and dialects of the Sotho group (from [20]).

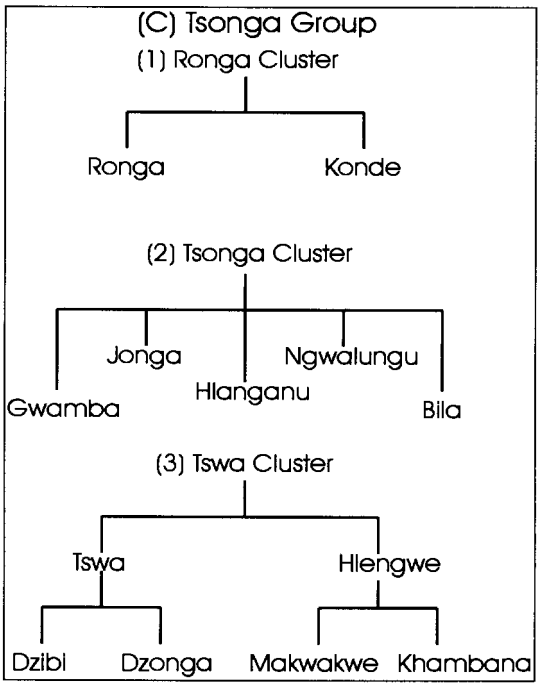


Figure 2.5: Clusters and dialects of the Tsonga group (from [20]).

### Phonetic classification

When these languages are classified solely according to their vowel systems, they can all be divided into two main groups, the Nguni and the Sotho vowel groups [5, 7]. The South African native languages are typical of Bantu languages as they have a perfectly balanced vowel system, which is characterised by one low vowel /a/, and an equal number of front and back vowels which mirror each other in height.

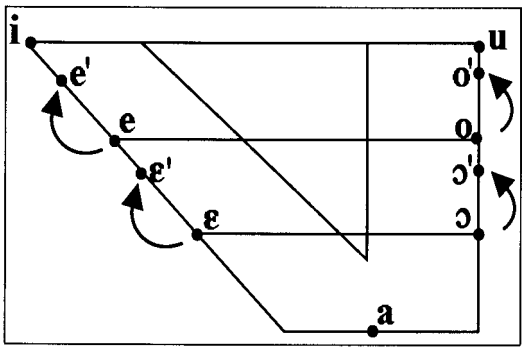


Figure 2.6: The seven main vowels used in Sotho (from [21, 22]).

It is found that all the languages within the Sotho group contain seven main vowels [21, 22] /i/, /e/, /ɛ/, /a/, /ɔ/, /o/ and /u/ (Figure 2.6). Apart from these there are also four raised vowels in Sotho. Raised vowels are phonetic variants of the same phoneme, caused by vowel assimilation [23]. Assimilation takes place when a vowel with higher tongue position in an adjacent syllable "raises" the pronunciation of a vowel with a lower tongue position. This usually happens when a "lower" vowel is followed in the next syllable by a "higher" vowel, for example the "e" (/ɛ'/) in "molemi" is raised (by the /i/) compared to the "e" (/ɛ/) in "lema".

The languages contained in the Nguni group as well as the Tsonga and Venda groups all display a simpler structure with five main vowels /i/, /ɛ/, /a/, /ɔ/ and /u/, and two raised vowels /ɛ'/ and /ɔ'/ as shown in Figure 2.7 [24, 25].

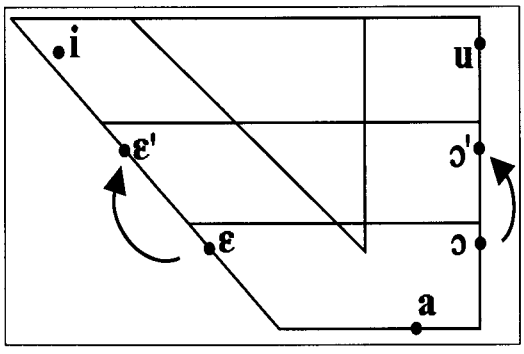


Figure 2.7: The vowels used in the Nguni, Tsonga and Venda languages (from [24, 25]).

### 2.1.3 English in Southern Africa

Previous studies on African accents of English have focused both on defining the South African English (SAE) accent as well as the reasons behind its existence. More general studies on language learning also give some insight into the formation of SAE, while studies on how accent is perceived points to the most prominent features of accented speech and how they can be measured.

Linguistic studies such as those done by Lanham [26], and Wells [6] examine so-called *white* South African English (WSAE) by indicating the deviations of this accent from that of British Received Pronunciation (RP) English. This helps to define what is called first language (L1) English in this study. As they state, WSAE ranges in pronunciation from a more conservative (RP) to a broader (more extreme) accent. So L1 English as used in this study lies somewhere between these two extremes and the effects reflected in these studies are also visible when comparing L1 English to L2 as influenced by an African mother tongue.

Other studies from the linguistical view point such as those performed by Schmied [3] and Jacobs [8] specifically examine the accents of English in Africa when spoken by African mother-tongue speakers. These results help to define the second language (L2) English used in this study.

While the abovementioned studies are fairly general in their treatment of language

accent as a speech pathology or linguistic problem, more recent studies focus on the acoustic properties of accented speech. Studies such as those done by Arslan and Hansen [27, 28, 1] and Flege [2] define some features that can be used to measure accent and quantitatively determine the effects of foreign accent.

### Characteristics of SAE

Lanham [26] states that the strongest factor influencing perceived accent is the pronunciation of vowels (and diphthongs) and proceeds to give 6 characteristic features of SAE according to a 1967 study:

1. Diphthongs are often replaced with long vowels or pronounced as very weak diphthongs. The diphthongs most seriously affected are /aɪ/ (as in *ride*, *high* or *wise*) and /əʊ/ (as in *road*, *low*, *close*). The /aɪ/ diphthong is replaced by either /o/ or /ɑ/, while /əʊ/ tends to lose the lip rounding and become neutral /ə/.

2. The vowel /ɪ/ as in *sin*, *did* or *this* is pronounced as [ə] as in the Afrikaans words *sit*, *dik* or *wil*. Also, the replacement of the /ɪ/ in *wind* or *pin* with [ə] is uniquely South African. This is caused by spelling out the word as it is written, instead of learning the pronunciation orally.

3. In *yes*, *men* or *get* the /ɛ/ is pronounced with a higher tongue position in SAE in the direction of cardinal /i/.

4. In extreme SAE lip rounding is given to long vowels /ɔ:/, /ɑ:/ and /ɜ:/ and the tongue is raised to an abnormally high position.

5. Another SAE effect is the shortening of the long vowels mentioned above and in addition shortening and monophthongisation of the diphthong /ɛə/ as in *fair* or *there*.

6. The lateral approximant /l/ influences the vowel quality of a preceding /ɪ/, /ɛ/ or /ʊ/.

The vowel space movements caused by these hallmarks of SAE are summarised in Figure 2.8, where the dotted lines represent the monophthongisation of diphthongs and other arrows indicate vowel movement from British English to SAE.

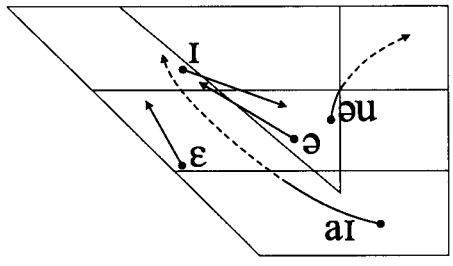


Figure 2.8: Vowel space changes in SAE according to Lanham [26].

Lanham also first (1967 [29]) argued for the existence of three distinct phonemes in SAE for the vowel (/ɪ/) in *kit*. He argued that not only are /ɪ/ and /ə/ distinct, but that /ə/ itself is pronounced as either a '*high schwa*' or '*low schwa*'. Later (1978 [30]) however, he opts for unifying all three as a single phoneme.

A later phonetic study on English in South Africa was performed by Wells [6] in 1982. Although he finds a number of typical vowel changes in SAE, the most striking is the *split* in the so-called *KIT* vowel (/ɪ/).

In contrast with Lanham's three (or one) phoneme view, he tentatively proposes the existence of two distinct phonemes. He states that the historical (RP) /ɪ/ has undergone a phonemic split in SAE, which can be clearly distinguished if a word such as *sing* ([siŋ ∼ sɪŋ]) is compared with *limb* ([lïm ∼ ləm]). Here "∼" indicates pronunciation ranging from the former to the latter version. The phoneme /ɪ/ as in *big* is pronounced as [i ∼ ɪ], while /ə/ as in *bit* is pronounced as [ï ∼ ə] and is a stressable vowel in SAE. Lanam's *high* and *low* schwa are therefore merged as a single phoneme. There is also a tendency to raise and front the /ɪ/ (as in *kiss*) making it virtually cardinal, thus [kis] or [lik].

The other front vowels are also changed in SAE. The /ɛ/ in *dress* and /æ/ in *trap* tend to be closer in SAE than in RP. The vowel in *dress* has an SAE pronunciation ranging from the RP [ɛ] to almost a cardinal [e], while the vowel in *trap* is spoken from a conservative [æ] to a broad-accent [ɛ]. A typical SAE pronunciation of *yes man* would therefore be ['jes mɛn].

The vowel of *fair* (/ɛə/) is also pronounced closer, but also more monophthongal. It ranges from a conservative [ɛə] to a general [e:], thus [ðe:] *there*.

The vowel changes defined by Wells are summarised in Figure 2.9.

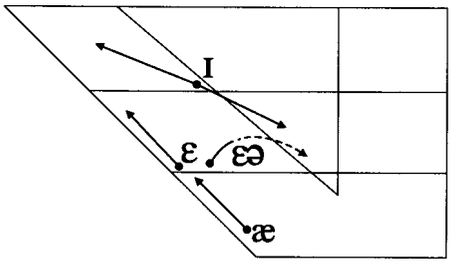


Figure 2.9: Vowel space changes in SAE according to Wells [6].

As far as diphthongs and long vowels are concerned, Wells agrees with Lanham that the latter half of diphthongs tend to be weakened, while long vowels are pronounced very back and sometimes rounded.

### African forms of English

The accent of English as spoken by African mother tongue speakers was assessed in a study by Schmied [3]. He states that the pronunciation of English in Africa is of particular importance, since non-standard pronunciation seem to be the most persistent in African accents as they are retained by even the most educated speakers. He attributes a number of reasons for African forms of English, such as the general learning strategies used by children and exposure to the written language. However, he downplays the

effect of African mother tongue influence and states that in many cases speakers are exposed to a number of African languages in addition to English. He therefore rather proposes that second language English is affected by a common substratum of African languages known to the speaker.

Exposure to the written language causes "spelling pronunciation" to take place, where the literary form is used as a phonetic guideline. This causes, for example, the articulation of /b/ and /h/ in *debt* and *heir*. This also causes various deviations in the pronunciation of /ə/, where it is pronounced as the vowel suggested by the orthographic symbol in the text, for instance resulting in [ageɪn] for *again*.

More specifically he states that central English vowels (/ʌ/, /ɜ:/ and /ə/ as in *but*, *bird* and *a*) are avoided and tend to more peripheral pronunciation ([e], [ɛ], [i], [ɔ], [ɑ]).

He also agrees with the other studies that diphthongs are monophthongised, especially shorter, closing diphthongs (/eɪ/, /əʊ/) where the second element is hardly heard. Longer diphthongs are preserved, but both elements receive the same emphasis, where in standard English emphasis on the second element is diminished. The mentioned de-centralisation of vowels affects all centering diphthongs (/ɪə/, /eə/, /ʊə/) and tends to make them opening diphthongs ([ɪa], [ea], [ʊa]).

A comparison of RP, East African English (EAE) and WSAE according to Schmied is shown in Figure 2.10. Here one can clearly see how groups of RP and WSAE vowels are treated as a single vowel in EAE. According to Schmied, Southern African English corresponds closely with EAE, which also uses a five-vowel structure as in the case of the Nguni languages.

Van Rooy and Van Huyssteen [4] summarise the characteristics of so-called *black* South African English (BSAE) based on formant analysis. Their vowel results are based on Tswana English (TE) speakers, while they used data from both Sotho- and Nguni-language speakers in their diphthong analysis.

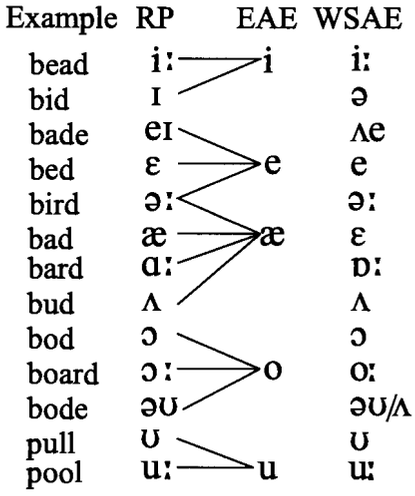


Figure 2.10: Comparison of the vowel spaces of RP English, East African English (EAE) and White South African English (WSAE) according to Schmied [3].

They found that the vowel system of BSAE contains fewer vowel contrasts than WSAE and state that the main reason for this is the lack of discrimination between the long and short forms of vowels in BSAE. A neutralisation in contrast is found between /ɪ/ and /i:/ (as in *kit* and *fleece*), /ʊ/ and /u:/ (as in *foot* and *goose*) as well as /ɔ/ (or /ɒ/) and /ɔ:/ (as in *not* and *caught*). They state that central vowels generally retain their height in vowel space, but are pronounced more to the front, therefore /ʌ/ and /ɜ:/ become /a/ and /ɛ/, respectively. They also find that the schwa (/ə/), which is often used in unstressed syllables in WSAE, is replaced by a full (stressed) vowel in BSAE. They agree with Schmied [3], stating that this substitution frequently takes place according to the orthographic spelling of the word. Vowel substitution as described by Schmied is also supported by their results, where English phonemes which do not exist in the mother-tongue are substituted by the closest native vowel, specifically /æ/ is replaced by /ɛ/.

In their diphthong analysis, Van Rooy and Van Huyssteen found that the only WSAE diphthong which is reproduced in BSAE is the diphthong /ɔɪ/ as in *choice*. All other diphthongs are monophthongised, specifically /aɪ/ and /aʊ/ reduce to /e/ and an extremely far back /o/ respectively. Of interest is their finding that the monophthongised

diphthongs in BSAE often result in vowels which do not exist as phonemes in many African languages, such as /ɪ/ and /e/. They also recognise the need for further study of the pronunciation of BSAE vowels and diphthongs to aid in the development of automatic speech recognition technology in South Africa.

From a perceptual viewpoint, a study was performed by Van Rooy, Wissing and Van den Heever [31] to specifically investigate the effect of under-differentiation (vowel substitution) in Tswana English (TE). They compared the perception of "known" vowels (/i/, /ɛ/, /ɔ/ and /ɑ/), which are valid phonemes in Tswana, with that of "new" vowels (/ɪ/, /ï/, /ɜ:/, /æ/ and /ʌ/), which are unfamiliar to Tswana speakers. Their results are summarised in Figure 2.11, where solid lines indicate the majority classification of a vowel and dashed lines the second most frequent classification. From this, it is clear that vowel contrasts are neutralised in the TE vowel system. In the case where "new" vowels are pronounced by WSAE speakers, TE listeners score significantly lower classification accuracy than WSAE listeners. This seems to indicate TE pronunciation may be influenced by a coarser perceptual framework, which impairs L2 English learning.

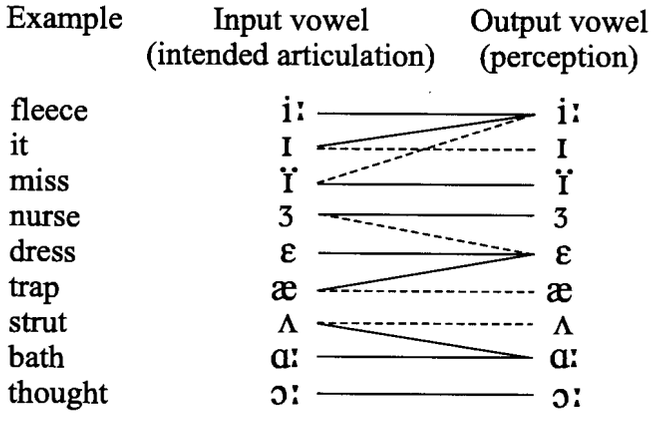


Figure 2.11: Vowel substitution in Tswana English, adapted from Van Rooy et al. [31].

### Intelligibility of L2 SAE

Jacobs [8] studied the effects of consonant variation on the intelligibility of English speakers with Zulu as mother-tongue from a socio-linguistic point of view. She showed

that other phonemic level deviations occur in African forms of English, for instance where minimal pairs (i.e. words where a single phoneme change causes a change in meaning, e.g. *bad*→*bed*→*bird*) are not retained. The elements in such minimal pairs are not distinguished clearly and tend to be pronounced the same (homophony), making them difficult to distinguish.

She states that this leads to unintelligible English speech in many educated second language speakers, which hinders their interaction with the broader English speaking community. She argues that the anti-contrastivity argument, which states that ambiguous pronunciation of minimal pairs does not affect intelligibility since the context will reveal the correct meaning, is not valid. It oversimplifies the problem since it is only concerned with isolated words whereas spoken discourse relies on strings within a social context. Often the contextual factors are not well defined and she warns that the divergence of Zulu English (ZE) from the South African English norm is accelerated due to the accented English taught to students.

This view is contested by many in the language teaching field. Adendorff and Savini-Beck [32], for instance, state that attempting to teach first language English pronunciation is unrealistic and regards mother-tongue English as an elitist variety. They also promote that a second language speaker often whishes to speak with a marked accent as a symbol of the person's heritage and that it should be retained as a symbol of solidarity. On the other hand they also place strong emphasis on intelligibility and advocates that accent should be kept within bounds to ensure effective communication.

We would therefore like to characterise the acoustic differences caused by such an accent in order to aid further research on the adaptation of automatic speech recognition (ASR) systems, enabling them to successfully recognise both first and second language SAE.

### Equivalence classification

The *equivalence classification hypothesis* states that speakers of a certain region or ethnic group tend to develop the same phonetic reference frame when learning language as a child and that this framework fossilizes long before the age of 12 years. When a new language is then learnt at a later age, the theory states that the person will tend to substitute second language (L2) phones with similar first language (L1) phones.

Flege [33] explains that as a result, second language vowels are often pronounced with specific deviations which are influenced by the person's mother tongue. He finds from formant results that the L1 counterpart is often substituted in the case of similar sounding L2 phones, while wholly new and unfamiliar L2 vowels are recognised as such and are appended to the speaker's phonetic framework during the learning process. Speakers of second language English therefore often mispronounce similar phones, while new phones are learnt authentically.

### Automatic speech recognition of accented English

Another team who has completed numerous studies on empirical accent and language classification is Levent Arslan and John Hansen. They state that speaker accent is an important issue in developing robust speaker independent recognition systems and that knowledge gained from reliable accent classification can help improve overall recognition performance. They agree with other language education studies [34] that a speaker acquires a specific speaking style up to the age of 16, which is generally preserved when a second language is acquired. As a result speakers tend to substitute phonemes from their native language when they encounter a new phoneme in a second language. As an example the phoneme /æ/ as in *cat* is not a valid phoneme in most languages other than English.

In one study [27] where they used a number of prosodic features, as well as formants,

to classify foreign English accent, they found that the most distinct features of accent are at the phonemic level, with phoneme substitution being a consistent identifier of language accent.

In another case [28] they compared the performance of an automatic language accent classifier, using Mel-scaled cepstral coefficients and energy features, against that of humans using a subjective listening test. Their results show that their computer based accent classification consistently achieves better performance that humans. They state that although other prosodic features, such as pitch (intonation) contour also vary significantly with accent, inter-speaker variability makes it difficult to isolate what portion of the variability is due to accent alone. Variation in other spectral features and energy are a more reliable indication of language accent. Their results also indicated that whole word models capture accent information better than monophone models - i.e. word-context affects pronunciation of the same phoneme - and that some words are better relayers of accent that others, due to their phonemic contents. Finally they illustrated a dramatic decrease in error rate for an automatic speech recognition system using accent dependent hidden Markov models (HMMs), compared to that of a first language based system.

In a later study [1] Arslan and Hansen performed a comprehensive acoustic study of foreign English accent using temporal features, intonation patterns and frequency characteristics. This included a detailed frequency analysis of accented speech, where they propose that the problem of speech recognition and accent (or language) classification rely on different regions of the frequency spectrum. They propose that since the human auditory system is highly sensitive to low frequencies [35], non-native speakers tend to concentrate on correcting the low-frequency characteristics of their speech (which is governed by the overall shape of the vocal tract and corresponds to the first formant, F1). The midrange frequencies (1500 - 2500 Hz) on the other hand differ from that of native speech, since these deviations are not as easily perceived by the learner. These frequencies (F2-F3 range) also represent detailed tongue movements which are more difficult to adopt for non-native speakers. They therefore propose a new frequency

scale which is more sensitive to accents than the linear or Mel-scaled scales. Their results indicate that high frequencies (2500 - 4000 Hz) do not impact much on the performance of either speech recognition or accent classification. Midrange frequencies (1500 - 2500 Hz) on the other hand contribute more to accent classification, while low frequencies (100 - 1500 Hz) are more significant for speech recognition.

This concludes our discussion of the linguistic background. We will refer to the findings discussed here in Chapter 3, where we will compare them to our experimental results. The following section details the speech processing techniques used in our analysis.

## 2.2 Speech processing and analysis

In this section the acoustic and signal processing techniques used in the comparisons are briefly described. Section 2.2.1 examines a robust algorithm for determining formants, while Section 2.2.3 details the cepstral features used for hidden Markov modeling of the L1 and L2 phonemes. Finally Section 2.2.2 gives a brief overview of cubic splines which were used to model the dynamic nature of diphthongs in formant space.

### 2.2.1 Formant analysis

When a person speaks, the vocal tract continuously changes shape and modulates the voice signal. The resultant vowel sound then consists of the *resonant frequencies* of the vocal tract, called *formants*. The variation of these formants with time forms formant tracks, which can easily be seen as the peak energy intensities in a spectrogram image of a voiced sound (Figure 2.12, *left*). The spectrogram is formed by computing the Fourier transform within a small time window (usually in the order of tens of milliseconds) and shifting this window along the time signal. For each position of the window a new spectrum is calculated. By vertically plotting these successive spectra,

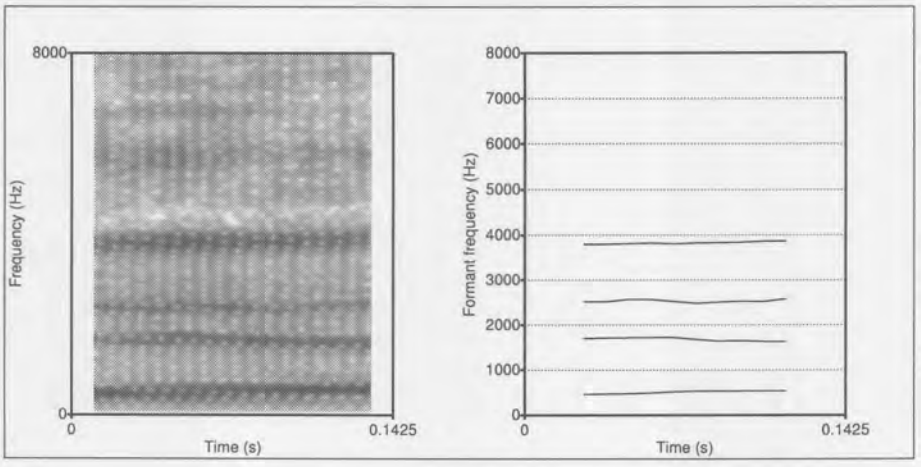


Figure 2.12: The spectrogram (left) and first four formants (right) as extracted by the Split Levinson algorithm for the vowel /i/.

using intensity to indicate energy, a spectrogram is created. Usually the first three or four formants (F1 through F4 from lower to higher frequencies) are used in acoustic analysis as these contain the majority of the information when studying vowels.

A number of problems present themselves when one attempts to determine the formant values via numerical analysis. The two most common ones are:

1. Not all voiced sounds explicitly contain the first three or four formants - the voice may be low-pass filtered by the vocal tract or formants may merge, forming a single resonance frequency.

2. On other occasions the formula used may fail due to numerical reasons and fail to converge.

The method used in this study was proposed by Willems [36] to robustly determine the formant values and is described below.

The first step in formant analysis is usually to estimate the spectral envelope of the speaker's vocal tract by an all-pole filter using linear predictive coding (LPC) analysis. LPC provides a good model of the speech signal, especially for voiced segments, where

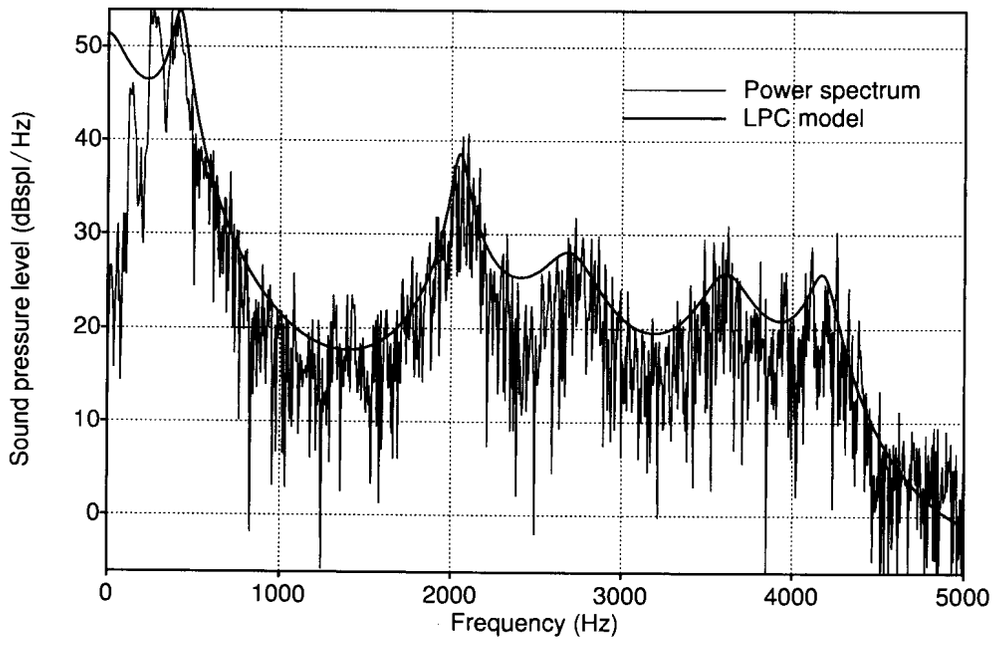


Figure 2.13: Fitting the spectrum of the vowel /i:/ with an LPC model.

the spectral envelope contains well-defined resonances (Figure 2.13). LPC essentially models the speech signal as a scaled generator function ($u(n) \times G$) that is modulated by a time-varying all-pole filter ($H(z)$), as shown in Figure 2.14. The generator function can be seen as the speaker's vocal chords, which produce a quasi-periodic signal, while the all-pole filter represents the speaker's vocal tract, which modulates the signal to produce different voiced sounds.

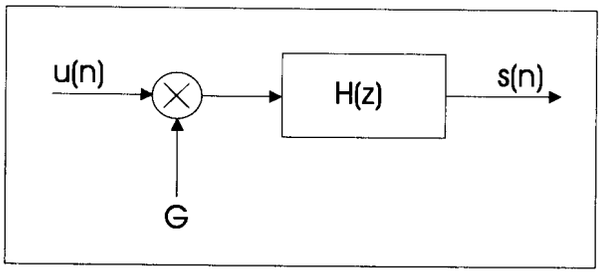


Figure 2.14: The LPC model of voiced speech.

The all-pole filter is defined as a polynomial of order $m$ as follows:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + a_2 z^{-1} + \ldots + a_{m+1} z^{-m}}. \tag{2.1}$$

The all-pole filter therefore models the spectral characteristics of the speaker's vocal tract during a specified time interval. The spectral peaks, indicating the resonance frequencies or formants, can be found by analysing this LPC model in second order sections as discussed shortly. Let us first examine LPC analysis in more detail.

In LPC analysis a speech section of $N$ samples (usually around 25ms in duration) is multiplied by a Hamming window, from which the autocorrelation coefficients are calculated as

$$r_k = \sum_{j=1}^{N-k} s_j \cdot s_{j+k}. \tag{2.2}$$

Using this autocorrelation sequence, the filter coefficients $(a_2, a_3, \ldots, a_{m+1})$ can be determined from the set of so-called Yule-Walker equations

$$\mathbf{Ra} = \mathbf{b}, \tag{2.3}$$

where $\mathbf{R}$ is a Toeplitz matrix (which means it is persymmetric) and $\mathbf{b}$ is identical to the first column of $\mathbf{R}$, shifted by one element and with opposite sign:

$$\begin{bmatrix} r_1 & r_2^* & \cdots & r_m^* \\ r_2 & r_1 & \cdots & r_{m-1}^* \\ \vdots & \ddots & \ddots & \vdots \\ r_m & r_{m-1} & \cdots & r_1 \end{bmatrix} \begin{bmatrix} a_2 \\ a_3 \\ \vdots \\ a_{m+1} \end{bmatrix} = \begin{bmatrix} -r_2 \\ -r_3 \\ \vdots \\ -r_{m+1} \end{bmatrix} \tag{2.4}$$

where $r_j^*$ is the complex conjugate of $r_j$.

The *Levinson* algorithm [37] can then be used to solve this set of linear equations. It is a recursive algorithm, where for each recursion $k$ an A-polynomial $A_k(z)$ of next higher order is calculated with all zeros *within* the unity circle.

The zeros of the polynomial can be either complex-pairs or real. These zero-pairs can

be written as:

$$N(z) = 1 + pz^{-1} + qz^{-2}. \tag{2.5}$$

Using these zero-pairs, the polynomial can be represented as second-order sections:

$$A(z) = \prod_{j=1}^{M/2} N_j(z^{-1}) = \prod_{j=1}^{M/2} (1 + p_j z^{-1} + q_j z^{-2}). \tag{2.6}$$

These $(p_j, q_j)$-pairs are then determined using the so-called Bairstow algorithm.

Each complex pair zero represents a resonance (formant) and the $p_j$, $q_j$ values give the formant's frequency and bandwidth respectively from:

$$p_j = -2e^{-\pi B_j T} \cdot \cos(2\pi F_j T) \tag{2.7}$$

$$q_j = e^{-2\pi B_j T}, \tag{2.8}$$

where $T = 1/F_s$ is the sampling period. From this $B_j$ and $F_j$ can be found.

Real zeros cannot be converted to formant values as these do not describe resonances, but rather give the spectrum a certain slope.

The two problems identified earlier can now be defined as:

1. Existence of real zero's from which no formants can be derived.
2. Failure of the Bairstow algorithm to converge.

In order to overcome these problems, a more robust method, called the *Split Levinson* algorithm, was developed by Genin and Delsarte [36], requiring approximately half the multiplications when performing the LPC analysis, compared to the normal Levinson method. This is possible because instead of the A-polynomials, singularity predictor

polynomials (SPPs) are used. These are symmetric polynomials whose zeros lie *on* the circle of unity and therefore contain half the amount of meaningful coefficients.

The SPPs are defined as:

$$P_k(z) = A_k(z) + \hat{A}_{k-1}(z), \tag{2.9}$$

where $A_k(z)$ is the $k^{th}$ recursion of the Levinson algorithm and $\hat{A}_k(z)$ the reciprocal polynomial of $A_k(z)$.

The recursive form can be shown to be:

$$P_{k+1} = (1+z)P_k(z) - \alpha_k z P_{k-1}(z) \tag{2.10}$$

where $\alpha_k$ is a parameter determined from the autocorrelation coefficients.

The A-polynomial at a given moment $M$ can be calculated from the SPP as:

$$A_M(z) = (P_{M+1}(z) - \lambda_M z P_M(z))/(1-z) \tag{2.11}$$

where $\lambda_M$ a parameter which, like $\alpha_k$, can also be calculated from the autocorrelation coefficients.

The robust formant extraction algorithm can therefore be summarised as follows:

1. Speech segment $(1 \cdots N)$ is multiplied with a Hamming-window.
2. Autocorrelation coefficients $(r_k)$ for the segment are calculated.
3. The Split Levinson algorithm is used to find the zero pairs $(p_j, q_j)$, indicating the resonances of the all-pole filter $H(z) = \frac{1}{A(z)}$.

This technique is used to determine the formant values of the vowels in our analysis. The following section details the methods used to analyse the time-varying formant values of diphthongs.

### 2.2.2 Splines

While vowels are static in nature and formant-space comparison could be made by comparing mean values, diphthongs had to be analysed differently. As the vowel quality of a diphthong changes, it forms a trajectory in formant space. Viewing each formant in isolation, the trajectory can also be represented by the variation of each formant over time. These formant tracks were analysed with the aid of cubic splines which model each formant track over time as a number of cubic splines. This section briefly describes the derivation of a cubic spline from a set of data points.

If $\{(x_k, y_k)\}_{k=0}^{N}$ denotes $N+1$ data points where $x_k$ is monotonically increasing (i.e. determined axis of a function), then the cubic spline fit to the data $S(x)$ is defined [38] as the set of cubic polynomials $S_k(x)$ with the following properties:

I. $S(x) = S_k(x) = s_{k,0} + s_{k,1}(x - x_k) + s_{k,2}(x - x_k)^2 + s_{k,3}(x - x_k)^3$
for $x \in [x_k, x_{k+1}] \quad \forall k = 0, 1, \ldots, N-1.$

II. $S(x_k) = y_k$ for $k = 0, 1, \ldots, N.$
(The spline passes through each data point.)

III. $S_k(x_{k+1}) = S_{k+1}(x_{k+1})$ for $k = 0, 1, \ldots, N-2.$
(The spline forms a continuous function.)

IV. $S'_k(x_{k+1}) = S'_{k+1}(x_{k+1})$ for $k = 0, 1, \ldots, N-2.$
(The spline forms a smooth function.)

V. $S''_k(x_{k+1}) = S''_{k+1}(x_{k+1})$ for $k = 0, 1, \ldots, N-2.$
(The second derivative is continuous.)

The following variables are defined in order to calculate the spline coefficients $s_{k,0}$ through $s_{k,3}$ in (I) from the data points:

$$\begin{aligned} m_k &= S''(x_k), \\ h_k &= x_{k+1} - x_k, \\ d_k &= \frac{y_{k+1} - y_k}{h_k}. \end{aligned}$$

It can be shown [38] that the following relation exists:

$$h_{k-1}m_{k-1} + 2(h_{k-1} + h_k)m_k + h_k m_{k+1} = u_k, \quad \text{where } u_k = 6(d_k - d_{k-1}) \text{ for } k = 1, 2, \ldots, N-1. \tag{2.12}$$

This gives us $N-1$ linear equations, containing $N+1$ unknowns. However, if we choose a relaxed natural spline (the end points are constrained by setting $S''_{x_0} = m_0 = 0$ and $S''_{x_N} = m_N = 0$), we obtain a diagonally dominant linear system of $N-1$ equations with a unique solution. After the unknowns $\{m_k\}$ are determined, the spline coefficients can be calculated from:

$$\begin{aligned} s_{k,0} &= y_k, \quad s_{k,1} = d_k - \tfrac{h_k(2m_k + m_{k+1})}{6}, \\ s_{k,2} &= \tfrac{m_k}{2}, \quad s_{k,3} = \tfrac{m_{k+1} - m_k}{6h_k}. \end{aligned} \tag{2.13}$$

This spline technique is used as the basis for modeling the diphthongs in formant space as further described in Chapter 3. The formant analyses of vowels and diphthongs are augmented with comparisons in the Mel-scaled cepstral domain, as described below.

### 2.2.3 Mel-scaled cepstra

One of the most widespread techniques used in modern speech recognition systems to model the speech units such as phonemes or words, is hidden Markov models (HMMs) with Mel-scaled cepstral coefficients as features. Cepstra are well suited for speech recognition, since the cepstral features are robust against channel effects and tend to separate features of the voice source and the modulation. Psychophysical studies have shown that human frequency perception does not follow a linear scale, which led to the definition of subjective pitch scale, called the "Mel" scale [35]. Using the Mel scale therefore models the perceptual frequency scale used by humans to recognise speech. It has been shown that speech recognition tasks perform better when the cepstrum is based on such a warped frequency scale [39].

An HMM models a chosen speech unit as a number of successive states, where each state represents a time segment in which the cepstral features remain fairly static. As the speech changes over time, one progresses from one state to the next, according to a transition probability matrix. An HMM is trained by estimating the probability density function of the features for each state from a set of training data. One can then discriminate between different speech units by comparing how well each trained model fits some chosen input speech.

In this section we consider the cepstral coefficients. The cepstrum of a sound wave is defined as the Fourier transform of the log of its power spectrum. These coefficients therefore describe the shape of the power spectrum. For the power spectrum $S(\omega)$ of a sampled signal, the Fourier series expansion of $logS(\omega)$ can be written as:

$$logS(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega}, \tag{2.14}$$

where $c_n = c_{-n}$ are real and called the cepstral coefficients.

The distance between a pair of spectra $S(\omega)$ and $S'(\omega)$ can be estimated by the $L_2$ cepstral distance. This distance corresponds to the root mean square (RMS) log spectral distance as follows:

$$\begin{aligned} d_c^2 &= \int_{-\pi}^{\pi} |logS(\omega) - logS'(\omega)|^2 \frac{d\omega}{2\pi} \\ &= \sum_{n=-\infty}^{\infty} (c_n - c'_n)^2, \end{aligned} \tag{2.15}$$

where $c_n$ and $c'_n$ are the cepstral coefficients of $S(\omega)$ and $S'(\omega)$.

It can be shown that these coefficients form a decaying sequence [17], and therefore the summation in Eq. (2.15) does not need an infinite number of terms. A truncated cepstral distance can then be defined as:

$$d_c^2(L) = \sum_{n=1}^{L} (c_n - c'_n)^2. \tag{2.16}$$

As a reference, a single tone, 40 dB above the hearing threshold, with a frequency of 1 kHz relates to a subjective pitch of 1000 mels. Below 1 kHz the mel scale follows an approximately linear relationship with frequency, while above 1 kHz the relationship is essentially logarithmic. The mel-scale can therefore be approximated by the following relationship between a frequency $f$ in Hz and its equivalent in mels:

$$mel(f) = 2595 \times \log_{10}(1 + f/700). \tag{2.17}$$

The power spectrum can be converted to such a subjective scale via a filter-bank with center frequencies and bandwidths corresponding to this scale. An example of such a mel-scaled filter-bank is given in Figure 2.15. Here the filters have a triangular band-pass response with a constant spacing of 150 mels and a bandwidth of 300 mels. If the

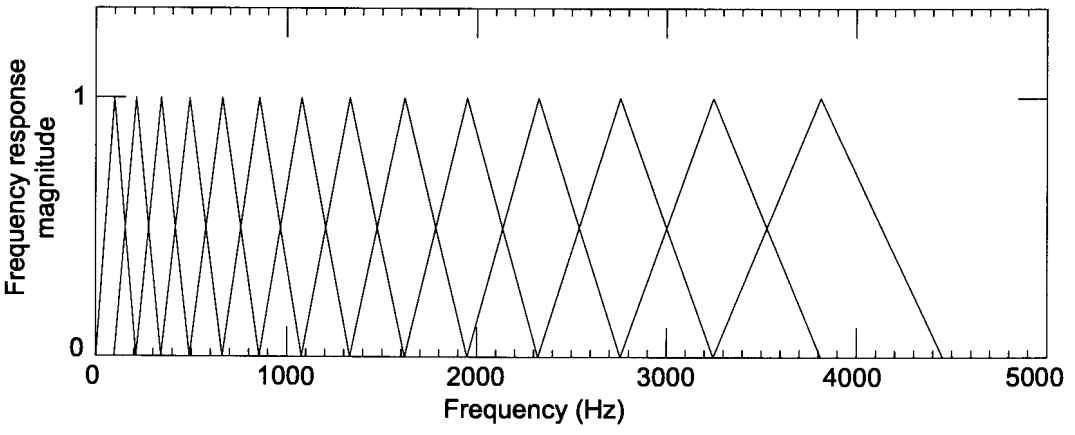


Figure 2.15: A mel-scaled triangular filter bank example.

output of filter $k$ is denoted by $\tilde{S}_k$, then the mel-scaled cepstrum $\tilde{c}_n$ can be calculated as:

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], n = 1, 2, \ldots, L, \tag{2.18}$$

where $L$ is the desired length of the cepstrum.

These features can now be readily used as a robust input in training hidden Markov models (HMMs).

### 2.2.4 Hidden Markov models

Although a full treatment of hidden Markov models (HMMs) is beyond the scope of this document, a brief overview is given here, adapted from Dugad and Desai [40].

Essentially an HMM is a statistical model of a time varying process based on an observation sequence ($O = O_1, O_2, \ldots, O_T$). We cannot observe the actual process, but only its outcome. The temporal nature of the process is modelled by a number of states. The probability of observing a specific outcome (observation symbol) is defined separately for each state and the probability that given a certain state, we move on to

another state is modelled as a transition probability matrix. Mathematically, an HMM model can be defined as follows [17, 41]:

$$\lambda = (A, B, \pi) \tag{2.19}$$

where:

$A$ defines the transition matrix $\{a_{ij}\}$ where $a_{ij} = P(i_t + 1 = j | i_t = i)$, the probability of being in state $j$ at time $t + 1$, given state $i$ at time $t$.

$B = \{b_j(k)\}, b_j(k) = P(v_k \text{ at } t | i_t = j)$, the probability of observing the symbol $v_k$ given state $j$.

$V = \{v_1, \ldots, v_M\}$, the discrete set of possible observation symbols.

Given our HMM model, three main questions can be asked:

- Question 1: Given $\lambda = (A, B, \pi)$, how do we compute $P(O|\lambda)$? That is, what is the probability of observing the observation sequence $O = O_1, O_2, \ldots, O_T$ given our model $\lambda$? (Speech Recognition)
- Question 2: Given $\lambda = (A, B, \pi)$, how do we choose a state sequence $I = i_1, i_2, \ldots, i_T$ in order to maximise $P(O, I|\lambda)$, the joint probability of the observation sequence $O = O_1, O_2, \ldots, O_T$ and the state sequence $I = i_1, i_2, \ldots, i_T$, given the model $\lambda$? (Segmenting speech into phonemes or words)
- Question 3: How do we adjust the model parameters $\lambda = (A, B, \pi)$ so that $P(O|\lambda)$ or $P(O, I|\lambda)$ is maximised? (HMM training on given speech data)

A straight-forward way of solving the first question, would be to find $P(O|I, \lambda)$ for a fixed state sequence $I$, multiply by $P(I|\lambda)$ and then sum over all possible $I$'s. This presents a problem however, since the calculation of $P(O|\lambda)$ would involve in the order of $2TN^T$ multiplications (where $T$ is the number of states in the observation sequence

and $N$ denotes the number of states in the model). A much more efficient procedure to calculate this is called the forward-backward procedure [17] and involves iteratively calculating the probability of a partial observation sequence.

Question 2 states that we need to find the state sequence $I = i_1, i_2, \ldots, i_T$ such that the probability of the observation sequence $O = O_1, O_2, \ldots, O_T$ for this state sequence is greater than for any other state sequence, i.e. maximising $P(O, I|\lambda)$. Again, there exists an inductive algorithm to calculate this, called the *Viterbi algorithm* [42]. In this algorithm one keeps the best possible state sequence at each instant for each of the $N$ states as the intermediate state of the desired observation sequence $O = O_1, O_2, \ldots, O_T$. This results in a best path for each of the $N$ states as the final state of the observation sequence, from which one selects the path with highest probability.

While the first two questions are related to analysis and recognition problems, the third is involved when training an HMM model. The problem is to train an HMM to best fit a set of training observation sequences, given the model in order to be able to recognise a similar observation sequence. Depending on the probability used for discrimination, one of two methods can be used:

1. Segmental k-means algorithm [43]: This method adjusts the model parameters $\lambda = (A, B, \pi)$ in order to maximise $P(O, I|\lambda)$, where $I$ in this case is the optimum state sequence as given by the Viterbi algorithm.

2. Baum-Welch re-estimation formulas [17, 41]: Here the parameters of the model $\lambda = (A, B, \pi)$ are adjusted to maximise $P(O|\lambda)$. This method requires an initial model and is often used as a subsequent step to segmental k-means training. In this case the focus is not on a particular state sequence, but rather the performance of the model across all state sequences, given the training observations.

### 2.2.5 Analysis of variance (ANOVA)

The purpose of analysis of variance is to determine the significance of difference between a number of sampling means [44]. Suppose one has $a$ sets of data collected from different sources $\{X_j\}$, for $1 \leq j \leq a$ with $n_1, n_2, \ldots, n_a$ the number of elements in each data set. The *variance* of all the data from all sets can be estimated from:

$$v = \sum_{j,k} (x_{jk} - \overline{x})^2, \tag{2.20}$$

where $\sum_{j,k}$ is the summation over $k$ from 1 to $n_j$ (data from source $j$) and then over $j$ from 1 to $a$ (summation over all data sources) and $\overline{x}$ is the mean of all the data.

The total variance *between* data sets is given by

$$v_b = \sum_{j,k} (\overline{x}_{j.} - \overline{x})^2 = \sum_{j} n_j (\overline{x}_{j.} - \overline{x})^2, \tag{2.21}$$

where the '.' in $\overline{x}_{j.}$ indicates the mean of the data from source $j$, where we have summed over all values of $k$. The variance *within* data sets is given by

$$v_w = \sum_{j,k} (x_{jk} - \overline{x_j}.)^2 = v - v_b. \tag{2.22}$$

We can see each data set as a random sample of size $n_j$ from the population of that particular data set. Therefore we have for data set $j$ a group of mutually independent, identically distributed random variables $X_{j1}, X_{j2}, \ldots, X_{jn_j}$, which take on the values $x_{j1}, x_{j2}, \ldots, x_{jn_j}$. Each of these variables can be defined as the sum of its expected value and an "error" value

$$X_{jk} = \mu_j + \Delta_{jk} \tag{2.23}$$

where $\Delta_{jk}$ can be taken as independent, normally distributed random variables with zero mean and variance $\sigma^2$. Alternatively, this means $X_{jk}$ can be described by independent random variables with mean $\mu_j$ and common variance $\sigma^2$. We can define $\mu$ as the mean of all the data sets

$$\mu = \frac{1}{a}\sum_j \mu_j, \tag{2.24}$$

which enables us to rewrite Eq. (2.23) as

$$X_{jk} = \mu + \alpha_j + \Delta_{jk} \text{ where } \sum_j \alpha_j = 0. \tag{2.25}$$

If the null hypothesis is true (all data sets are statistically the same) then $\alpha_j = 0; j = 1, 2, \ldots, a$. If we define $n = \sum_j n_j$ as the total number of data samples in all data sets, then from Eqs. (2.20), (2.21) and (2.22) the expected values of the variances can be defined as

$$E(V_b) = (a-1)\sigma^2 + \sum_j n_j\alpha_j \tag{2.26}$$

$$E(V_w) = (n-a)\sigma^2 \tag{2.27}$$

$$E(V) = (n-1)\sigma^2 + \sum_j n_j\alpha_j \tag{2.28}$$

From Eq. (2.27) it is clear that we can define

$$\hat{S}_w^2 = \frac{V_w}{n-a} \tag{2.29}$$

as an unbiased estimate of $\sigma^2$ whether the null hypothesis is true or not. If the null hypothesis is false, however, then from Eq. (2.26)

$$E(\hat{S}_b^2) = \sigma^2 + \frac{1}{a-1}\sum_j n_j \alpha_j^2, \tag{2.30}$$

which is larger than $\sigma^2$ and is linearly dependent on the difference between data set means. The ratio $F = \hat{S}_b^2/\hat{S}_w^2$ is therefore a good statistic for testing the null hypothesis and has a Fisher distribution with $a-1$ and $n-a$ degrees of freedom [44]. The significance of difference between the distributions of the different data sets can therefore be measured according to the single ratio $F$ derived from the data with

$$F = \frac{\hat{s}_b^2}{\hat{s}_w^2}, \text{ where} \tag{2.31}$$

$$\hat{s}_b^2 = \frac{v_b}{a-1} \text{ and} \tag{2.32}$$

$$\hat{s}_w^2 = \frac{v_w}{n-a}. \tag{2.33}$$

Analysis of variance is used in this study whenever normal distributions are compared to determine if two distributions are significantly different.

In the case of formant analysis, each vowel of a particular language group is modeled by a normal distribution of its formant features. Each diphthong is modelled by a cubic spline and each spline coefficient is modelled as a normal distribution. First and second language spline models are then compared using ANOVA. In the cepstral domain, vowels are also represented by normal distributions of their cepstral coefficients, while diphthongs are modelled as three distributions, forming three successive states of an HMM.

In all the above cases, the ANOVA measure is used to grade the distances found between speech sounds. The 95% and 99% significance levels of the $F$ ratio for a

two-class ANOVA test are listed in Appendix A. These levels are used to determine if a significant difference exists between two distributions to a certainty of 95% or 99% respectively.

This chapter discussed some of the linguistic background, such as the basic concepts of phonetics and phonology and looked at the classification and phonetics of native Southern African languages as compared to that of English. Some background on English accents in South Africa was given, which created the framework on which an explanation of the mathematical analysis could be based. These methods are subsequently used in the experiments which are detailed in Chapter 3.

# Chapter 3

# Experiments

This chapter firstly lists the objectives we wish to achieve with our analysis. This is followed by a discussion of the data set that was used. The acquisition, structure and pre-processing of the data are examined.

Finally, the experiments themselves are detailed. Five distinct experiments were performed and for each the method, results and our interpretation thereof are given.

## 3.1 Objectives

As was stated in Chapter 2, the spectral characteristics of second language English phonemes differ in a number of ways from that of the L1 norm. The main objective is therefore to characterise the differences between a number of vowels and diphthongs of first language (L1) South African English (SAE) and second language (L2) SAE as spoken by native speakers of African languages.

We will compare the two language groups by performing experiments to determine the following:

1. The difference in formant space between L1 and L2 vowels.
2. The difference in formant space between L1 and L2 diphthongs.
3. The difference in Mel-scaled cepstral space between L1 and L2 vowels.
4. The difference in Mel-scaled cepstral space between L1 and L2 diphthongs.
5. The trends of diphthongisation and monophthongisation.

## 3.2 Data

In this section the data used in order to perform the experiments is examined. We discuss the requirements the data must fulfill to be applicable to the study (for instance what speakers were included or excluded). The data source (broadcast speech) is discussed, the classification used to structure the data, the processing performed to reduce the data to a manageable set and finally the resulting data set is detailed.

This data is included on an accompanying compact disk set as described in Appendix C.

### 3.2.1 Requirements

In order to be representative of local South African accents, the data had to be obtained from a wide spectrum of South African speakers. Both first and second language English speakers were of interest so that African mother-tongue influence could be compared to a local first-language norm. The second language English speakers were constrained to first language speakers of native African languages, such as isi-Zulu, isi-Xhosa, Sesotho, Sepedi, etc.

Another aim is for the inter-speaker variability of spectral features within a language group to be minimised. When these features are considered, speakers can be categorised

into three groups: male, female and children, where females tend to have higher spectral components and children even higher. For this reason children were excluded from the data set, while male and female data were treated separately.

### 3.2.2 Source

A practically endless source of data, meeting the above requirements was found in local South African television broadcasts. Speech segments were captured from local news and similar broadcasts on VHS video tape and later digitised with a PC sound-card at 22kHz (16-bit resolution) for storage as PCM wave files on computer. A total of $4\frac{1}{4}$ hours (698Mb) of raw speech data was collected.

### 3.2.3 Classification

As the data was digitised and stored on computer, each speech segment was labelled according to the specific video tape and position on the tape it originated from to enable traceability. The speaker(s) within the speech segment were added to a list of globally unique speaker labels. Each speaker's gender was noted and further also subjectively categorised (by the author) according to the strength of the perceived accent on a five-point scale as follows:

- 0 - L1 speaker
- 1 - L1 speaker, with some influence of native language accent
- 2 - Borderline
- 3 - L2 speaker, fair amount of accent perceived
- 4 - L2 speaker, strong accent

Each speech segment was also attributed a quality rating ranging from 0 (very poor) to 10 (excellent), according to the perceived amount of background noise, interference from other speakers or background music and other channel effects. This quality rating was used to aid in selecting the data in order to minimise the labelling effort. All the speaker information was saved in a central speaker database, while a transcription counterpart for the audio file was saved, indicating the speaker labels and other parameters associated with the speech segment. These files are described in more detail in the following section.

### 3.2.4 Processing

This section details the pre-processing that was performed on the raw sound data to create a useful speech database. The requirements of this database is to enable us to gain direct access to the phonemes we wish to compare in the experiments. We must also be able to obtain statistics from the database, such as the number of words present in each language and gender group that are of acceptable quality, the number of utterances present for each unique word in the database and so forth.

The pre-processing steps are summarised by the diagram in Figure 3.1. After the raw speech has been acquired, it is segmented according to speakers and each speaker section is then orthographically transcribed (typing in the text spoken). Using the statistics available from this data, a suitable subset of words is selected and their word labels are automatically generated. The position of these labels (indicating the section of speech file which corresponds to a specific word) is only approximate and subsequently needs to be corrected by hand. When the selected words have been accurately labelled, these word segments are extracted from the original speech data. These segments are then phoneme labelled by hand and the phoneme segments extracted to generate the final context-dependent phonemes to be used in the experiments.

The database structure is shown in Figure 3.2. Each raw speech file is associated with

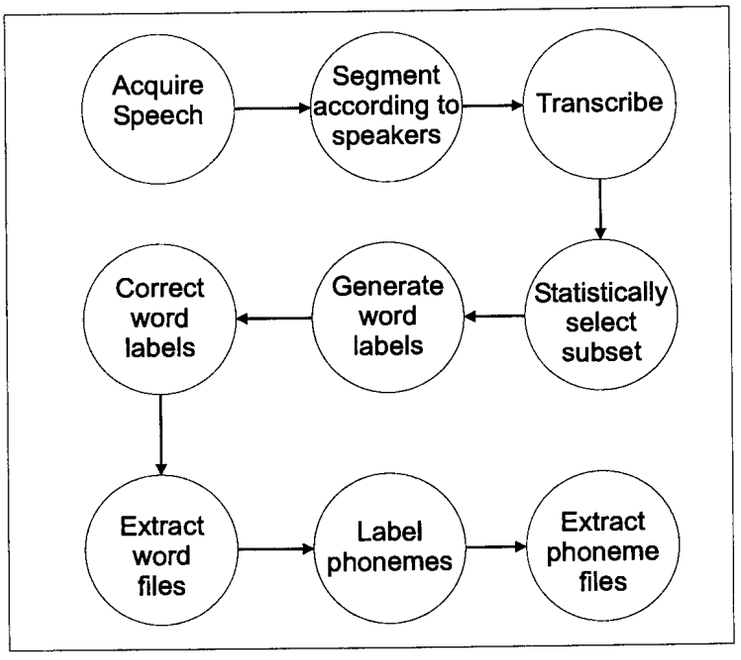


Figure 3.1: Data pre-processing steps.

the speaker database and a text transcription file, which contains the words spoken in text (orthographic) format. Every speech file is also associated with a number of word labels, indicating where in the file specific words are present. Each word label is also associated with an extracted copy of the word segment, stored as a separate file. In turn, each word file is associated with a number of phoneme labels, and each of these with an extracted copy of the phoneme. These phoneme files are then used as input to the experiments.

The data processing and software tools (also included on the accompanying compact disk set) are discussed in more detail in the following sections.

### Transcription

In order to analyse the recorded data, the text transcriptions of each speech segment was required. A software tool called *txscribe* was written to aid in this task, allowing the information associated with each speech file to be saved in an accompanying

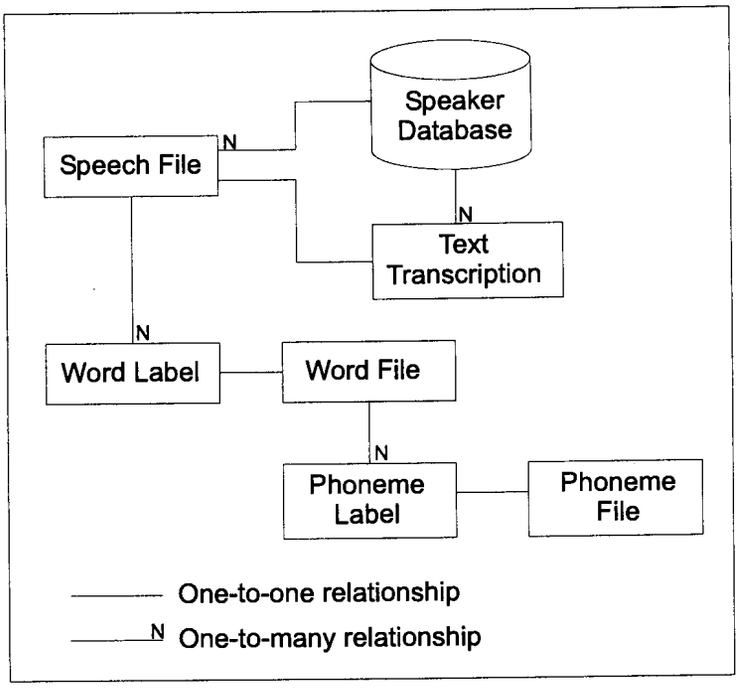


Figure 3.2: Database structure.

transcription file. This tool also managed and updated the central speaker database as data was added. An example window of *txscribe* is shown in Figure 3.3 and the resulting transcription file for the case of a single-speaker speech file is listed below:

```
#TRANSCRIPTION
SP1: He is a very capable and experienced leader and I want his input in everything that I do.
Has been invaded by Uganda and by Ruanda and if therefore, those two countries declare, call
for a cease fire and a peaceful meeting between the three parties that is president president
president, I'm sure  that we'll have taken a very important step towards a peaceful solution.
#NOTES
Some echoing present
#SPEAKERS
SP1: 010
SP2: -01
SP3: -01
SP4: -01
SP5: -01
#TYPE
News - Microphone Speaker
#QUALITY
5
#EOF
```

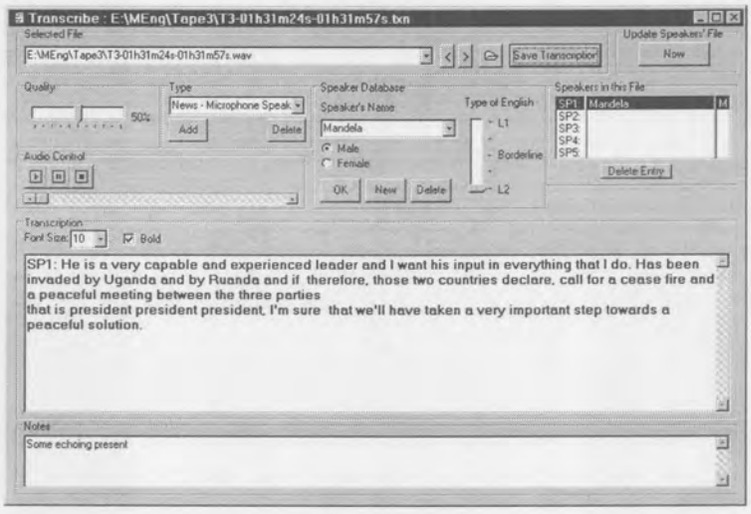


Figure 3.3: The 'txscribe' software tool.

Up to five different speakers are supported per speech segment (the "`010`" indicates speaker number 10, while "`-01`" indicates an empty entry). The text associated with each speaker is preceded by the "`SP1:`" through "`SP5:`" keywords. The quality value of the speech segment is selected, as well as some additional information, including general comments about the speech segment and the type of speech segment (in this case a field interview as part of a news broadcast). Speaker specific information was not stored in this transcription file, but rather referenced to a central speaker database, using the speaker number.

The speaker database consisted of a single table listing the following speaker information:

- Name (if available)
- Gender
- Level of accentedness (0 to 4)
- Speaker ID (database key)

An excerpt from the speaker database containing the speaker of this speech segment is shown below:

```
Name             Gender    Accent  ID
---------------- --------- ------ -----
Manana Makhanya [FEMALE]  [4]    [212]
Mandela          [MALE]    [4]    [010]
Mandi Titi       [FEMALE]  [4]    [211]
Mandla Mtumu     [MALE]    [4]    [210]
```

### Data Reduction

A total of 38,000 words were orthographically transcribed. These transcriptions were used to analyse the relative frequency of occurrence of words within a number of frameworks. A software tool called *WordCount* was written for this purpose, which would parse the transcription files and speaker database to produce statistics based on speaker, speech segment (file), gender as well as language group (L1 or L2). The input could be constrained according to speech-segment quality (as assigned during transcription), speaker gender and language group. Specific words and speakers with accentedness level of 2 (borderline) could also be ignored. The results could further be filtered according to the following rules:

- number of speakers (within gender and language group)
- frequency of occurrence (within gender and language group)
- maximum number of speakers considered
- per-speaker upper bound (number of occurrences considered)

These criteria were used to optimise the data set by minimising the words considered during labelling, while maximising the amount of data useable for analysis. The following section describes how these statistics were used for labelling.

### Labelling

The occurrence (frequency) statistics obtained during data reduction was used to identify the words which could be used in the experiments. The aim was to minimise the manual labelling effort by limiting the total number of words considered, while maximising the cross-section between the two language groups (L1 and L2) for each word. This cross-section was considered within gender groups. The *WordCount* tool was extended to generate initial label files based on its results, which could be refined during labelling. The position of a word label within a speech segment was estimated from the length of the speech segment and the word's position within the transcription. These estimated labels proved very valuable for locating words within large speech files that are sparsely labelled.

The actual labelling was performed via a tool called *Wyre*, which was written for this purpose. An example window of *Wyre* is shown in Figure 3.4. The display includes scalable views of the time-domain signal, the spectrogram representation as well as two label tracks - one for word-level and one for phoneme-level labels. The spectrogram is generated using a Hamming window of 20ms and a step-size of typically 5ms, although this is adapted dynamically to suit the time-scale viewed. A linear frequency scale from 0 to 8kHz was usually selected in order to visualise both the formants of voiced sounds as well as the characteristic higher frequency signature of fricatives.

When a word label is created, the user can select the associated speaker from a list of available speakers present in the current speech file. The word's quality is initially automatically set to that of the speech file as a whole, but can be refined by the user to accommodate short-duration changes in speech quality (for instance a burst of background noise degrading only a few words). This quality rating was used to eliminate low-quality words during further analysis. A total of 3099 words were labelled in this manner and used in the ensuing step of phoneme-level labelling.

The acceptable words (with a quality rating of 4 or higher) were extracted from the

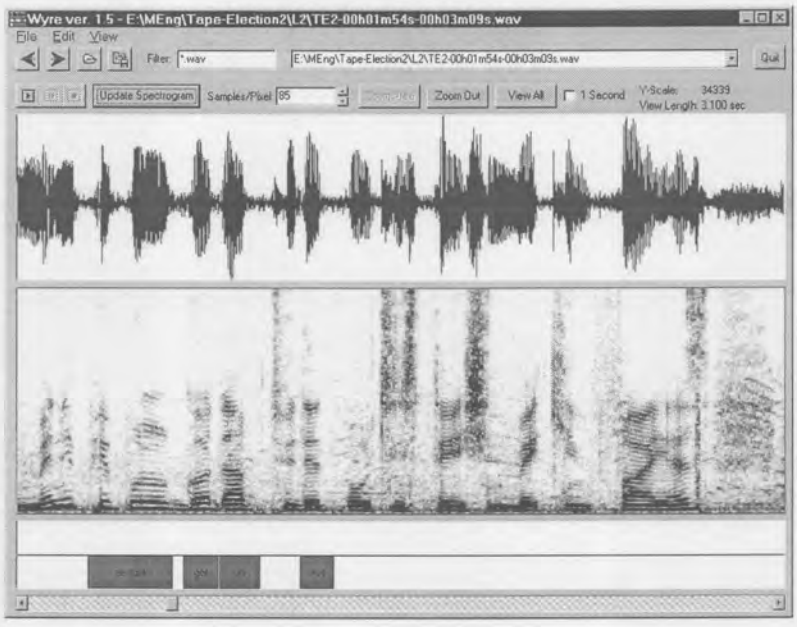


Figure 3.4: The 'Wyre' software tool.

original speech files as a data subset. Each extracted speech file contained exactly one word and had an associated information file indicating where it was extracted from to allow traceability.

Again using *Wyre*, each word was labelled phonemically (as they ought to be pronounced) rather than phonetically (as they are actually pronounced) to make phoneme-level comparison between the accents possible. Only voiced phonemes were labelled, using an adapted ARPABET [17] format as shown in Table 3.1.

Certain voiced segments (vowels, diphthongs, glides and liquids), such as the /ɪɚ/ (ARPABET /IY AXR/) sound in *here* and the /ʊɪ/ (ARPABET /W IY/) in *we*, were labelled as a single 'phoneme' (/IY-AXR/ and /W-IY/). This was possible, since phonemes are compared within word-context, resulting in the comparison of equivalent supra-phonemic structures (and treated as 'diphthongs' due to their paired-vowel structure and pronunciation). These phonemes were then extracted forming context-dependent phoneme files, which were used in the experiments.

| *Phoneme* | *Symbol* | *Example* | *Phoneme* | *Symbol* | *Example* |
|---|---|---|---|---|---|
| /ɪ/ | IH | h*i*t | /ɔɪ/ | OY | b*oy* |
| /i/ | IY | h*ea*t | /aʊ/ | AW | d*ow*n |
| /ɛ/ | EH | b*e*d | /əʊ/ | OW | b*oa*t |
| /ɝ/ | ER | b*ir*d | /ju/ | Y-UH | *you* |
| /ɚ/ | AXR | oth*er* | /ʊa/ | W-AA | *o*ne |
| /ə/ | AX | *a*bout | /ʊæ/ | W-AE | *w*ell |
| /æ/ | AE | b*a*d | /ʊɛ/ | W-EH | *wh*ere |
| /ʌ/ | AH | c*u*t | /ʊl̩/ | W-EL | *will* |
| /ɑ/ | AA | c*a*rt | /ʊə/ | W-IH | *wi*th |
| /ɔ/ | AO | c*augh*t | /ʊɪ/ | W-IY | *we* |
| /ʊ/ | UH | p*u*ll | /ɪɚ/ | IY-AXR | h*ere* |
| /u/ | UW | p*oo*l | /ɛɚ/ | EH-AXR | th*ere* |
| /aɪ/ | AY | b*uy* | /l̩/ | EL | peop*le* |
| /eɪ/ | EY | b*ai*t | | | |

Table 3.1: A list of the phonemes used during labelling and their equivalent ARPABET symbols.

### Resulting data set

This section describes statistics of the data set. We begin with the raw data and the text transcriptions and then proceed to describe the results of the selection process, choosing a sub-section for word and phoneme labelling and finally list the labelled phonemes that are used in the experiments.

The complete data set contained $4\frac{1}{4}$ hours of raw speech data, totalling 37,999 transcribed words from 236 speakers. As shown in Table 3.2, 56% of the data originated from L1 speakers, with the remaining 44% from L2 speakers, excluding any borderline cases. The male:female ratio of the words in the data set was 17 : 10, while the

male:female ratio of the speakers was 19:10 (154:82). Table 3.3 shows the speaker classification, with 34% classified as L1, 10% as L1 with a slight accent, 0% as borderline, 10% with noticeable L2 accent and 46% with a strong L2 accent.

| | *L1* | *L2* | *Total* |
|---|---|---|---|
| Male | 12769 | 11508 | 24277 |
| Female | 8699 | 5023 | 13722 |
| *Total* | 21468 (56%) | 16531 (44%) | 37999 |

Table 3.2: Total transcribed words per language and gender group.

| | *Male* | *Female* | *Total* |
|---|---|---|---|
| L1 | 50 | 30 | 80 |
| L1 - slight accent | 11 | 13 | 24 |
| Borderline | 0 | 0 | 0 |
| L2 - noticeable accent | 16 | 7 | 23 |
| L2 - strong accent | 77 | 32 | 109 |
| *Total* | 154 | 82 | 236 |

Table 3.3: Total speakers per language and gender group.

When the cross-section of words present in both language groups are taken for each gender group the results shown in Table 3.4 are found, indicating 79% of the total data set present in the cross-section.

| | *L1* | *L2* | *Total* |
|---|---|---|---|
| Male | 10156 | 9445 | 19601 |
| Female | 6268 | 4058 | 10326 |
| *Total* | 16424 | 13503 | 29927 |

Table 3.4: Total words in L1/L2 cross-sections within gender groups.

Although these cross-sections may seem large, the random nature of the data (i.e. unconstrained, natural speech) result in only a small number of words being adequately represented within each for the four groups (both gender and both language groups). Figure 3.5 shows the frequency of words for the 50 most abundant words in the L1/L2 cross-section for male speakers. As can be clearly seen, the frequency distribution follows a typical exponential decline when sorted according to word frequency. The graph is shown as a stacked-histogram, where the total count for a specific word is indicated by the bar as a whole, which is then sub-divided into L1 and L2 counts. For instance 1564 utterances of the first word ("the") was counted, with 864 being L1 and 700 L2. By the $50^{th}$ ranking word ("do") the number of utterances had already dropped to 65, with 28 L1 and 37 L2 utterances.

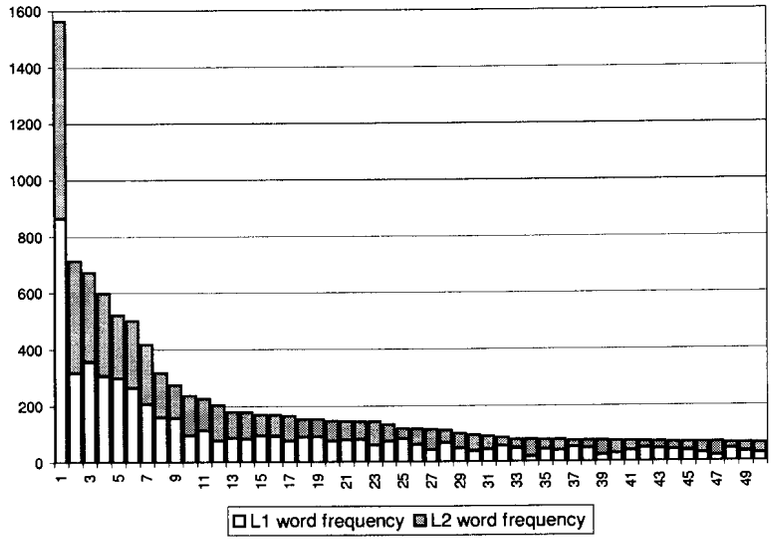


Figure 3.5: L1/L2 Cross-sectional word frequency for male speakers.

Of greater importance than word frequency, was the number of utterances from different speakers for a specific word, which followed a similar trend as the word frequency. For instance the word "the" was uttered by 127 different speakers, 51 classified as L1 and 76 as L2. A general trend was that although more repetitions of a word may have been uttered by L1 speakers, a greater number of different L2 speakers were usually present - indicating that speech segments acquired from L2 speakers were often shorter.

We know that gender and age influences both the fundamental and formant frequen-

cies [45] in a person's speech. Since we wish to focus our analysis on the cross accent variations and therefore minimise the variation within language groups, children were excluded from the data during the original recording phase, and only the male (more abundant) data was considered for labelling and the experiments.

Suitable words for labelling were selected from this male cross section by constraining the number of L1 or L2 speakers to a maximum of 20, while the word had to be represented by at least 10 utterances from different speakers in each language group. Speakers classified as borderline were excluded and an upper limit of 1 utterance per speaker was used to limit the labelling task. With only a single utterance per speaker allowed, a subset of 3081 words was extracted containing 92 different words each with at least 10 utterances from different speakers per language group.

This subset contained a number of unsuitable words though, such as words with multiple pronunciations (like "the" which can be either [ðə] or [ði:]) or short, usually unstressed, indefinite articles (such as "a" or "an"), which tend to be neutralised. With these removed, the subset shown in Table 3.5 was found, totalling 2963 words containing 89 unique words. However, since utterances can easily be of unusable quality, this subset was augmented by another 313 words by allowing 2 utterances per speaker where only 10 to 13 speakers were present in a language group - bringing the total to 3276 words.

Initial word labels were generated for these words using *WordCount* and subsequently hand corrected using *Wyre*. The words of acceptable quality (i.e. not too much background noise, music or cross talk present) were then extracted as separate sound files which are easier to manage.

The data was further reduced in the phoneme labeling stage where only phonemes for which accurate formants could be extracted were retained. This was used as a criterion to ensure that all experiments are based on the same data set, even those which do not rely on formant extraction. The 2668 individual word utterances therefore resulted in a

| *Word* | *# speakers* | *L1* | *L2* | *Word* | *# speakers* | *L1* | *L2* | *Word* | *# speakers* | *L1* | *L2* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| and | 40 | 20 | 20 | you | 40 | 20 | 20 | were | 29 | 16 | 13 |
| are | 40 | 20 | 20 | south | 39 | 19 | 20 | get | 28 | 18 | 10 |
| as | 40 | 20 | 20 | two | 39 | 20 | 19 | had | 28 | 17 | 11 |
| at | 40 | 20 | 20 | which | 39 | 19 | 20 | his | 28 | 13 | 15 |
| be | 40 | 20 | 20 | all | 38 | 20 | 18 | know | 28 | 13 | 15 |
| but | 40 | 20 | 20 | been | 38 | 20 | 18 | more | 28 | 18 | 10 |
| by | 40 | 20 | 20 | their | 38 | 18 | 20 | four | 27 | 15 | 12 |
| for | 40 | 20 | 20 | also | 36 | 19 | 17 | go | 27 | 12 | 15 |
| from | 40 | 20 | 20 | can | 36 | 17 | 19 | said | 27 | 12 | 15 |
| has | 40 | 20 | 20 | if | 36 | 16 | 20 | being | 26 | 12 | 14 |
| have | 40 | 20 | 20 | it's | 36 | 16 | 20 | got | 26 | 10 | 16 |
| i | 40 | 20 | 20 | up | 36 | 20 | 16 | come | 25 | 14 | 11 |
| in | 40 | 20 | 20 | who | 35 | 15 | 20 | like | 25 | 12 | 13 |
| is | 40 | 20 | 20 | just | 34 | 17 | 17 | or | 25 | 15 | 10 |
| it | 40 | 20 | 20 | no | 34 | 14 | 20 | sabc | 25 | 10 | 15 |
| not | 40 | 20 | 20 | what | 34 | 14 | 20 | when | 25 | 12 | 13 |
| of | 40 | 20 | 20 | because | 33 | 13 | 20 | where | 25 | 13 | 12 |
| on | 40 | 20 | 20 | he | 33 | 15 | 18 | out | 23 | 10 | 13 |
| one | 40 | 20 | 20 | here | 33 | 19 | 14 | those | 23 | 10 | 13 |
| other | 40 | 20 | 20 | now | 33 | 20 | 13 | time | 23 | 10 | 13 |
| so | 40 | 20 | 20 | people | 33 | 13 | 20 | cape | 22 | 12 | 10 |
| that | 40 | 20 | 20 | our | 32 | 12 | 20 | first | 22 | 12 | 10 |
| there | 40 | 20 | 20 | do | 31 | 16 | 15 | see | 22 | 11 | 11 |
| they | 40 | 20 | 20 | going | 31 | 16 | 15 | that's | 22 | 10 | 12 |
| this | 40 | 20 | 20 | some | 31 | 16 | 15 | think | 22 | 11 | 11 |
| to | 40 | 20 | 20 | very | 31 | 14 | 17 | back | 21 | 11 | 10 |
| was | 40 | 20 | 20 | them | 30 | 14 | 16 | last | 21 | 10 | 11 |
| we | 40 | 20 | 20 | then | 30 | 17 | 13 | hundred | 20 | 10 | 10 |
| will | 40 | 20 | 20 | well | 30 | 19 | 11 | really | 20 | 10 | 10 |
| with | 40 | 20 | 20 | africa | 29 | 15 | 14 | | | | |

Table 3.5: Subset of 89 words to be labelled indicating total number of speakers per word and speakers per language group

total of 80 context dependent phonemes, each with at least 10 utterances per language group. This final data set is summarised in Table 3.6, using the following format for the phonemes: "`<word context>-<ARPABET phoneme>`". For instance, the /ɔ/ in the word *all* is indicated as "`all-ao`". In the case of the acronym *SABC* (pronounced [ɛseɪbisi:]) the context contains multiple occurrences of the phoneme /i/. In this case the first instance is labelled "`sabc-iy1`" and the second "`sabc-iy2`".

| *Phoneme* | #L1 | #L2 | *Phoneme* | #L1 | #L2 | *Phoneme* | #L1 | #L2 |
|---|---|---|---|---|---|---|---|---|
| africa-ae | 16 | 13 | got-ao | 10 | 13 | sabc-iy2 | 10 | 11 |
| africa-ax | 11 | 10 | has-ae | 11 | 16 | said-eh | 10 | 12 |
| africa-iy | 11 | 10 | have-ae | 19 | 21 | see-iy | 11 | 10 |
| all-ao | 18 | 14 | he-iy | 14 | 10 | so-ow | 15 | 18 |
| also-ao | 12 | 16 | here-iyaxr | 15 | 10 | some-ah | 14 | 14 |
| also-ow | 11 | 14 | his-iy | 11 | 10 | south-aw | 15 | 17 |
| and-ae | 12 | 20 | if-iy | 10 | 14 | that-ae | 17 | 19 |
| are-aa | 19 | 14 | in-iy | 12 | 14 | their-ehaxr | 16 | 17 |
| as-ae | 11 | 17 | is-iy | 14 | 19 | there-ehaxr | 17 | 18 |
| at-ae | 10 | 17 | its-iy | 10 | 14 | they-ey | 16 | 11 |
| back-ae | 12 | 10 | just-ah | 10 | 13 | this-ih | 22 | 21 |
| be-iy | 20 | 15 | know-ow | 12 | 10 | those-ow | 11 | 10 |
| because-ao | 10 | 17 | more-ao | 15 | 11 | time-ay | 11 | 13 |
| been-iy | 19 | 10 | no-ow | 10 | 17 | to-uw | 12 | 16 |
| being-iy | 10 | 12 | not-ao | 19 | 21 | two-uw | 18 | 12 |
| but-ah | 18 | 19 | now-aw | 16 | 12 | very-eh | 11 | 12 |
| by-ay | 15 | 16 | of-ao | 10 | 19 | very-iy | 13 | 14 |
| can-ae | 11 | 13 | on-ao | 17 | 16 | was-waa | 12 | 17 |
| cape-ey | 12 | 10 | one-waa | 20 | 23 | we-wiy | 17 | 17 |
| come-ah | 13 | 11 | other-ah | 20 | 18 | well-wae | 13 | 10 |
| do-uw | 12 | 9 | other-axr | 19 | 17 | what-waa | 10 | 11 |
| first-er | 10 | 10 | our-awr | 11 | 15 | where-weh | 16 | 11 |
| for-ao | 10 | 17 | people-el | 12 | 19 | which-wiy | 10 | 16 |
| four-ao | 14 | 10 | people-iy | 13 | 15 | will-wel | 14 | 15 |
| from-ao | 14 | 21 | sabc-eh | 10 | 13 | with-wih | 12 | 15 |
| get-eh | 19 | 10 | sabc-ey | 10 | 12 | you-yuh | 11 | 12 |
| go-ow | 10 | 12 | sabc-iy1 | 10 | 12 | | | |

Table 3.6: Context dependent phonemes of the final data set and number of L1/L2 utterances of each.

## 3.3 Comparative formant analysis of vowels

In this experiment the vowels of first and second language speakers are compared in formant space. A total of 59 context dependent vowels (listed in ARBAPET format in Table 3.7) are examined. It may be noted that the context-dependent phonemes based on /o/ (/OW/) are included in this list of vowels. Although many phonological references define this phoneme as a vowel, others regard it is a short diphthong ([əʊ]). Due to its disputed status, we include /o/ in both our vowel and diphthong analyses, treating it as a vowel in the former and as a diphthong in the latter. Also, the vowel in *this* is labelled as /ɪ/ (IH) rather than /i/ (IY) in keeping with the split nature of the *KIT* vowel in SAE noted by Wells [6] and Lanham [26]. In our data the [i ∼ ɪ] category defined by Wells is labelled /i/ (/IY/) as in *his*, *if*, *in* and *is*, while [ï ∼ ə] is labelled /ɪ/ (/IH/).

In the next section the specific formant analysis technique used here is described in more detail, followed by the results. We then give our interpretation of these results and some final concluding remarks. Before continuing, however, a few remarks on visualisation of the data.

### Visualisation

If one examines the spectrogram of a typical vowel utterance (Figure 3.6, *left*), it is clear that the energy spectrum remains fairly constant over time. The same therefore holds true for the vowel formants (Figure 3.6, *right*).

One can therefore portray a vowel as a multi-dimensional Gaussian distribution, with a mean value and variance in each formant dimension.

Important work on the mean formant values of English vowels was performed by Peterson and Barney [45] in 1952. They determined the average formant values for 10

| | | | | |
|---|---|---|---|---|
| africa-ae | because-ao | go-ow | more-ao | said-eh |
| africa-ax | been-iy | got-ao | no-ow | see-iy |
| africa-iy | being-iy | has-ae | not-ao | so-ow |
| all-ao | but-ah | have-ae | of-ao | some-ah |
| also-ao | can-ae | he-iy | on-ao | that-ae |
| also-ow | come-ah | his-iy | other-ah | this-ih |
| and-ae | do-uw | if-iy | other-axr | those-ow |
| are-aa | first-er | in-iy | people-el | to-uw |
| as-ae | for-ao | is-iy | people-iy | two-uw |
| at-ae | four-ao | its-iy | sabc-eh | very-eh |
| back-ae | from-ao | just-ah | sabc-iy1 | very-iy |
| be-iy | get-eh | know-ow | sabc-iy2 | |

Table 3.7: Context dependent vowels used for formant analysis.

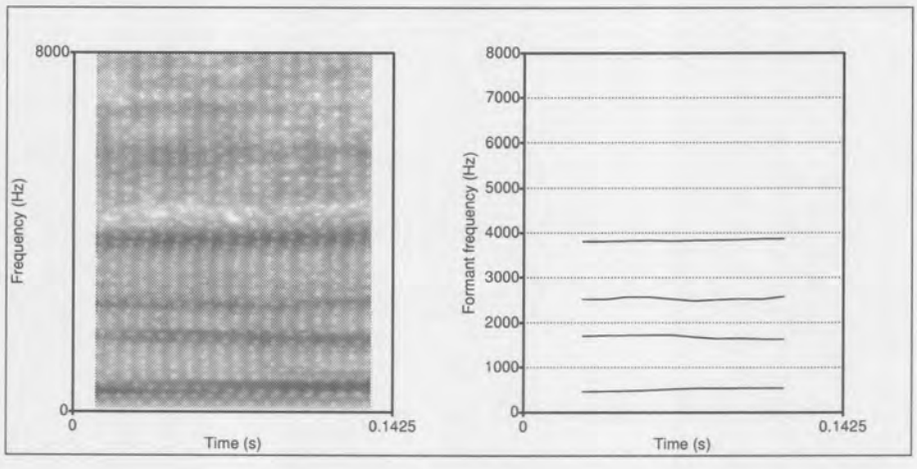


Figure 3.6: The spectrogram (left) and first four formants (right) for the vowel /i/.

English vowels, which they obtained from 33 male speakers. These were used as a reference frame for the formant plots in this analysis. A first versus second formant plot of these values (using reversed axes) are shown in Figure 3.7. Note that this reference frame only indicates the mean values and not the variances. As a comparison, notice the correlation between the I.P.A. vowel chart (repeated in Figure 3.8 for reference) with the F1/F2 formant positions of the reference vowels. We can see that vowel

"height" is reflected in the F1 value, where "higher" vowels, such as /i/ and /u/ (/IY/ and /UW/) have smaller F1 values, while "lower" vowels, such as /æ/ and /ɔ/ (/AE/ and /AO/), lie at a larger F1. The other dimension is represented by F2, with "front" vowels, like /ɛ/ (/EH/), having a larger F2, while "back" vowels, like /ɔ/ (/AO/), have a smaller F2 value.

This similarity makes it possible to interpret our formant results from a linguistic point of view and to compare them with the linguistic findings on South African English accent discussed in Chapter 2.

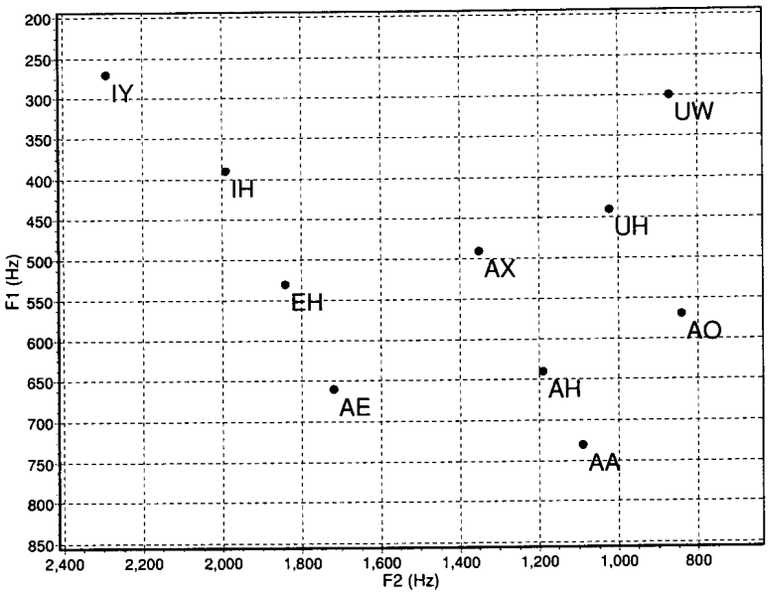


Figure 3.7: Peterson and Barney reference vowels in the F1/F2 plane.

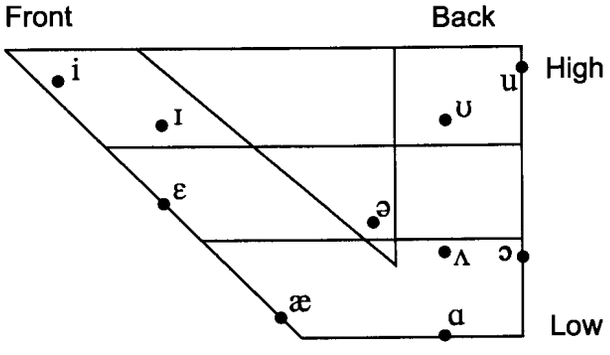


Figure 3.8: I.P.A. chart of selected English vowels (adapted from [14]).

A data visualisation software tool called *GPlot* was jointly developed with Prinsloo [46]

at the University of Pretoria for the purpose of plotting vowels and diphthongs in formant space. This was used to analyse the data and to generate the graphs given here and in the results section.

If we now plot the distribution of a single vowel as spoken by 10 speakers within this framework, we obtain the graph in Figure 3.9, *top left.* Here we can see the mean ($\mu = \frac{1}{N}\sum_{j=1}^{N} x_j$) for the vowel /æ/ (/AE/) indicated by the center of a cross and the individual data points as triangles. The variance, ($\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2$) is depicted by an ellipse with its major axis in the direction of maximum variance. The size of the ellipse in the direction of the major (minor) axis is plotted to be at two times the maximum (minimum) variance ($2\sigma^2$). Figure 3.9, *top right* shows the distribution of the same vowel in the F1/F3 space, with the F2/F3 point of view at *bottom right.*

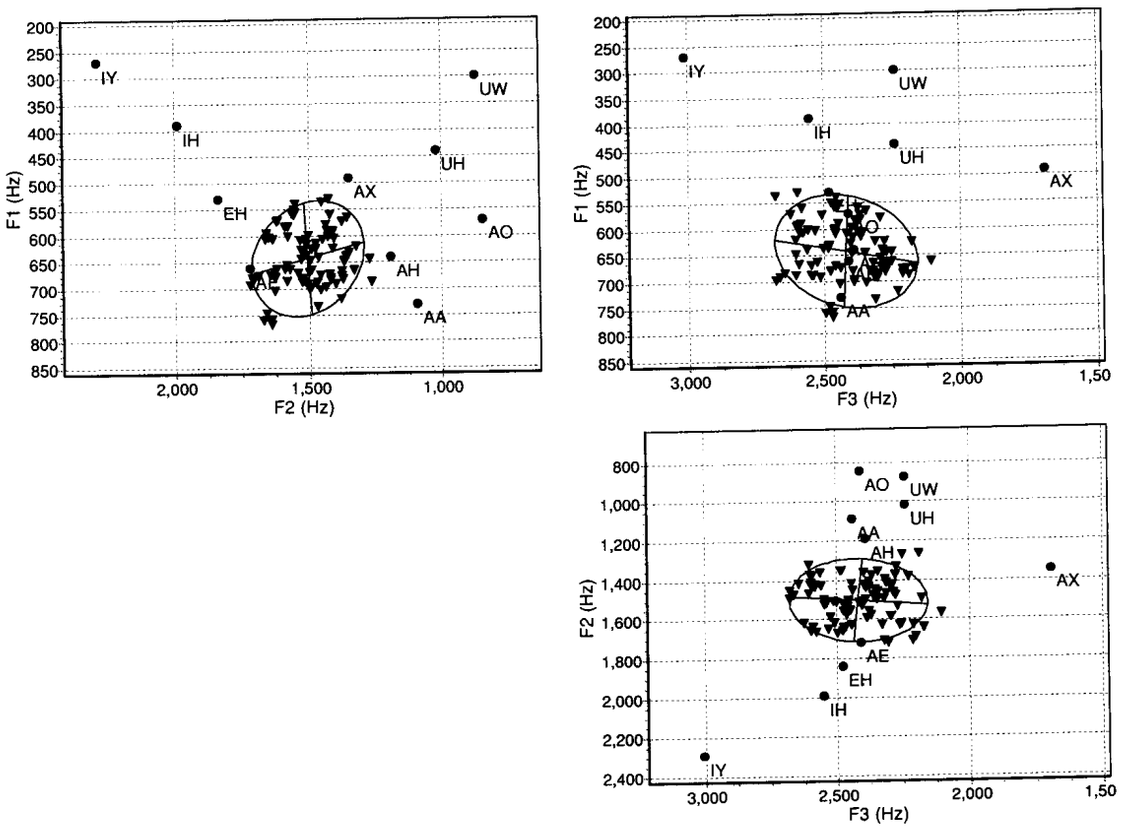


Figure 3.9: Orthographic plots of the distribution of the data obtained for the vowel /æ/.

### 3.3.1 Experimental protocol

In order to analyse the vowels in formant space, the extracted phoneme sound files were processed. For each phoneme, its spectrogram was plotted and formants estimated. The results for each language group were then combined by estimating first and second language Gaussian distributions. These two distributions were then compared using analysis of variance. The following sections define these steps in more detail.

#### Formant extraction

The first three formants were extracted using the Split Levinson algorithm as described in Section 2.2.1 and functions of the *Praat*[1] program with the following parameters:

- Analysis width: 20 ms
- Time step: 10 ms
- Pre-emphasis from 50 Hz

The formants were extracted for each phoneme segment and stored as separate files, which could be used as input to the data visualisation tool. As noted in Section 3.2.4 where the final data set is described, the extracted formants for each utterance were checked against the spectrogram by hand, and cases where they were not correctly estimated were excluded.

#### Comparison

In order to compare the vowels of each language group, each group was modelled by a normal distribution. This was accomplished by first computing the mean value for

[1] *Praat* by Paul Boersma, A system for doing phonetics by computer, IFA, University of Amersterdam (*www.fon.hum.uva.nl/praat*).

the data points of each individual utterance. The global mean and variance for each group was then calculated using these utterance-means. If the mean formant value of an utterance is defined as $\dot{x}_j$, the formant distribution (in each formant dimension) of all the utterances in the group can be described by:

$$\mu_g = \frac{1}{N}\sum_{j=1}^{N}\dot{x}_j \tag{3.1}$$

$$\sigma_g^2 = \frac{1}{N}\sum_{j=1}^{N}(\dot{x}_j - \mu_g)^2, \tag{3.2}$$

where $\mu_g$ is the group mean, $\sigma_g^2$ the group variance and $N$ the number of utterances. It is important to note, that by using the utterance means, each utterance carries the same weight towards the group's distribution. If each individual data point was used, a long utterance with many data points would have been given an unfair contribution, resulting in a skewed group distribution.

Such a group distribution is calculated for both language groups and then compared using analysis of variance. If the first language group contains $N_1$ utterances and the second $N_2$, we can calculate $F = \hat{S}_b^2/\hat{S}_w^2$ (from Section 2.2.5, page 39) and compare this with the significance ratios given for a Fisher distribution with 1 and $N_1 + N_2 - 2$ degrees of freedom.

The results of this comparison for the 59 vowels are given in the following section.

### 3.3.2 Results and discussion

This section details the results found from the formant analysis of vowels described above and we describe the trends present in the data based on these results. A number of tables are presented here, including the actual mean formant values found for each phoneme and the results of the analysis of variance (ANOVA) tests. These are followed by figures where the formant distributions used in the ANOVA tests are plotted to visualise the results. Throughout the tables and figures, phonemes are indicated using

ARPABET symbols. These are augmented in the text by the I.P.A. symbols for ease of reference.

The first two tables (Tables 3.8 and 3.9) list the mean formant frequencies found for each of the context-dependent phonemes. The first column indicates the vowel (for instance "aa" for the vowel /ɑ/ in the first row), while the second column indicates the context-dependent phoneme (in this case the /ɑ/ in context of the word *are* is indicated as "are_aa"). Different contexts of the same vowel are grouped in alphabetical order, for instance the /æ/ in *africa* through the /æ/ in *that* form a single sub-table. The third column indicates the language group ("L1" indicates first language and "L2" second language), followed by the number of utterances used to calculate the average ("n"). The last three columns list the mean formant frequencies for the first through third formants. Grey and white bands are used to visually group the L1 and L2 counterparts of the same phoneme together. Table 3.8 contains results for the vowels /ɑ/ through /ɛ/ (/AA/ through /EH/) and Table 3.9 those for /i/ through /u/ (/IY/ through /UW/). These mean formant frequencies are plotted in a series of sub-plots in Figures 3.10 through 3.14 on pages 76 through 80. These are best viewed in colour.

The analysis of variance results are shown in Table 3.10. The same structure as in the previous two tables is used, where the first column indicates the vowel and the second the context-dependent phoneme. The third column indicates the degrees of freedom ("DOF" = the total number of data points minus 2) of the calculated ANOVA $F$ ratio. The $F$ ratios for the first three formants are given in columns four through six. A light grey block indicates that the $F$ ratio is above the 95% significance level as indicated by the Fisher distribution (i.e. we can be 95% certain that the two distributions are significantly different). A dark grey block indicates that the $F$ ratio is above the 99% significance level. The final column indicates the value of a difference score. This score is calculated and summed across all three formant dimensions as follows: If an $F$ ratio is above the 95% level, the score is incremented by one. If the $F$ ratio is also above the 99% level, the score is incremented again. For instance, /ɔ/ in *all* (first context of the fourth vowel) has $F$ ratios of 36.95, 0.04 and 4.79 respectively for the three formants.

| Vowel | Context | Group | n | F1 (Hz) | F2 (Hz) | F3 (Hz) |
|---|---|---|---|---|---|---|
| aa | are_aa | L1 | 19 | 663 | 1149 | 2528 |
| | | L2 | 14 | 659 | 1355 | 2444 |
| ae | africa_ae | L1 | 16 | 635 | 1505 | 2428 |
| | | L2 | 13 | 634 | 1257 | 2446 |
| | and_ae | L1 | 12 | 615 | 1621 | 2476 |
| | | L2 | 20 | 534 | 1700 | 2501 |
| | as_ae | L1 | 11 | 645 | 1674 | 2472 |
| | | L2 | 17 | 552 | 1544 | 2485 |
| | at_ae | L1 | 10 | 665 | 1673 | 2448 |
| | | L2 | 17 | 549 | 1717 | 2514 |
| | back_ae | L1 | 12 | 646 | 1614 | 2441 |
| | | L2 | 10 | 580 | 1730 | 2425 |
| | can_ae | L1 | 11 | 596 | 1709 | 2435 |
| | | L2 | 13 | 493 | 1663 | 2331 |
| | has_ae | L1 | 11 | 557 | 1589 | 2434 |
| | | L2 | 16 | 488 | 1606 | 2422 |
| | have_ae | L1 | 19 | 625 | 1535 | 2342 |
| | | L2 | 21 | 542 | 1558 | 2417 |
| | that_ae | L1 | 17 | 571 | 1587 | 2480 |
| | | L2 | 19 | 575 | 1606 | 2485 |
| ah | but_ah | L1 | 18 | 596 | 1443 | 2494 |
| | | L2 | 19 | 631 | 1377 | 2435 |
| | come_ah | L1 | 13 | 668 | 1357 | 2353 |
| | | L2 | 11 | 641 | 1263 | 2463 |
| | just_ah | L1 | 10 | 527 | 1548 | 2482 |
| | | L2 | 13 | 570 | 1442 | 2540 |
| | other_ah | L1 | 20 | 674 | 1386 | 2508 |
| | | L2 | 18 | 645 | 1293 | 2468 |
| | some_ah | L1 | 14 | 669 | 1285 | 2442 |
| | | L2 | 14 | 632 | 1291 | 2439 |

| Vowel | Context | Group | n | F1 (Hz) | F2 (Hz) | F3 (Hz) |
|---|---|---|---|---|---|---|
| ao | all_ao | L1 | 18 | 472 | 995 | 2592 |
| | | L2 | 14 | 603 | 1004 | 2446 |
| | also_ao | L1 | 12 | 464 | 976 | 2673 |
| | | L2 | 16 | 531 | 1064 | 2474 |
| | because_ao | L1 | 10 | 456 | 1343 | 2370 |
| | | L2 | 17 | 490 | 1166 | 2495 |
| | for_ao | L1 | 10 | 491 | 1097 | 2532 |
| | | L2 | 17 | 491 | 1037 | 2621 |
| | four_ao | L1 | 14 | 464 | 933 | 2538 |
| | | L2 | 10 | 516 | 927 | 2440 |
| | from_ao | L1 | 14 | 603 | 1124 | 2346 |
| | | L2 | 21 | 522 | 1095 | 2354 |
| | got_ao | L1 | 10 | 558 | 1245 | 2388 |
| | | L2 | 13 | 502 | 1143 | 2558 |
| | more_ao | L1 | 15 | 472 | 912 | 2629 |
| | | L2 | 11 | 552 | 977 | 2396 |
| | not_ao | L1 | 19 | 600 | 1198 | 2534 |
| | | L2 | 21 | 548 | 1203 | 2457 |
| | of_ao | L1 | 10 | 587 | 1128 | 2558 |
| | | L2 | 19 | 526 | 1076 | 2534 |
| | on_ao | L1 | 17 | 616 | 1160 | 2559 |
| | | L2 | 16 | 532 | 1089 | 2416 |
| ax | africa_ax | L1 | 11 | 564 | 1561 | 2374 |
| | | L2 | 10 | 587 | 1423 | 2309 |
| axr | other_axr | L1 | 19 | 507 | 1454 | 2477 |
| | | L2 | 17 | 539 | 1339 | 2509 |
| eh | get_eh | L1 | 19 | 446 | 1885 | 2564 |
| | | L2 | 10 | 547 | 1709 | 2468 |
| | sabc_eh | L1 | 10 | 455 | 1933 | 2505 |
| | | L2 | 13 | 558 | 1647 | 2525 |
| | said_eh | L1 | 10 | 477 | 1785 | 2517 |
| | | L2 | 12 | 532 | 1669 | 2470 |
| | very_eh | L1 | 11 | 460 | 1675 | 2359 |
| | | L2 | 12 | 510 | 1562 | 2368 |

Table 3.8: Mean formant frequencies for vowels /ɑ/ through /ɛ/ (/AA/ through /EH/).

Since these were calculated with 30 degrees of freedom we can see (from the Fisher table given in Appendix A) that the two phonemes differed significantly with a 99% certainty in F1, no significant difference exists for F2 and a 95% level difference is found for F3. The difference score for this phoneme is therefore $2 + 0 + 1 = 3$. This gives an indication of how much two phonemes differ in a range from 0 (no significant difference) to 6 (significant difference in all three formant dimensions to a 99% certainty). The results for each of the vowels are discussed in more detail below.

| Vowel | Context | Group | n | F1 (Hz) | F2 (Hz) | F3 (Hz) |
|---|---|---|---|---|---|---|
| iy | africa_iy | L1 | 11 | 408 | 1739 | 2208 |
| | | L2 | 10 | 370 | 1689 | 2376 |
| | be_iy | L1 | 20 | 342 | 2065 | 2619 |
| | | L2 | 15 | 334 | 1886 | 2585 |
| | been_iy | L1 | 19 | 374 | 2029 | 2619 |
| | | L2 | 10 | 370 | 1855 | 2493 |
| | being_iy | L1 | 10 | 387 | 1980 | 2658 |
| | | L2 | 12 | 338 | 1873 | 2573 |
| | he_iy | L1 | 14 | 348 | 2113 | 2719 |
| | | L2 | 10 | 317 | 1910 | 2619 |
| | his_iy | L1 | 11 | 349 | 1892 | 2629 |
| | | L2 | 10 | 381 | 1773 | 2502 |
| | if_iy | L1 | 10 | 418 | 1893 | 2604 |
| | | L2 | 14 | 363 | 1903 | 2586 |
| | in_iy | L1 | 12 | 395 | 1905 | 2594 |
| | | L2 | 14 | 387 | 1900 | 2595 |
| | is_iy | L1 | 14 | 383 | 1783 | 2547 |
| | | L2 | 19 | 370 | 1902 | 2574 |
| | its_iy | L1 | 10 | 434 | 1862 | 2622 |
| | | L2 | 14 | 360 | 1860 | 2548 |
| | people_iy | L1 | 13 | 352 | 2010 | 2674 |
| | | L2 | 15 | 324 | 1898 | 2498 |
| | see_iy | L1 | 11 | 356 | 2055 | 2687 |
| | | L2 | 10 | 357 | 1806 | 2672 |
| | very_iy | L1 | 13 | 406 | 1902 | 2485 |
| | | L2 | 14 | 417 | 1661 | 2305 |
| | sabc_iy1 | L1 | 10 | 357 | 2036 | 2590 |
| | | L2 | 12 | 314 | 1890 | 2624 |
| | sabc_iy2 | L1 | 10 | 326 | 2118 | 2799 |
| | | L2 | 11 | 331 | 1916 | 2669 |

| Vowel | Context | Group | n | F1 (Hz) | F2 (Hz) | F3 (Hz) |
|---|---|---|---|---|---|---|
| el | people_el | L1 | 12 | 441 | 1124 | 2588 |
| | | L2 | 19 | 397 | 1096 | 2504 |
| er | first_er | L1 | 10 | 484 | 1438 | 2427 |
| | | L2 | 10 | 524 | 1438 | 2378 |
| ih | this_ih | L1 | 22 | 437 | 1475 | 2492 |
| | | L2 | 21 | 406 | 1733 | 2540 |
| ow | also_ow | L1 | 11 | 467 | 1373 | 2501 |
| | | L2 | 14 | 454 | 1186 | 2466 |
| | go_ow | L1 | 10 | 545 | 1478 | 2444 |
| | | L2 | 12 | 472 | 1049 | 2583 |
| | know_ow | L1 | 12 | 570 | 1472 | 2455 |
| | | L2 | 10 | 513 | 1076 | 2535 |
| | no_ow | L1 | 10 | 556 | 1359 | 2544 |
| | | L2 | 17 | 496 | 1070 | 2383 |
| | so_ow | L1 | 15 | 516 | 1324 | 2456 |
| | | L2 | 18 | 461 | 1012 | 2611 |
| | those_ow | L1 | 11 | 544 | 1498 | 2553 |
| | | L2 | 10 | 407 | 1060 | 2533 |
| uw | do_uw | L1 | 12 | 383 | 1775 | 2529 |
| | | L2 | 9 | 338 | 1237 | 2568 |
| | to_uw | L1 | 12 | 401 | 1678 | 2431 |
| | | L2 | 16 | 394 | 1295 | 2439 |
| | two_uw | L1 | 18 | 369 | 1663 | 2421 |
| | | L2 | 12 | 343 | 1087 | 2517 |

Table 3.9: Mean formant frequencies for vowels /i/ through /u/ (/IY/ through /UW/).

### /ɑ/ (/AA/)

The vowel /ɑ/ is defined as a low and back vowel [19, 15] and we therefore expect a large value for F1, but a small F2 [45]. If we compare the English vowel system with that of the African languages, we see that African languages do not explicitly use /ɑ/, but rather a slightly more front vowel /a/. Sotho uses an /a/ which is fairly close to that of English /a/ [21, 22], while the Nguni /a/ is not quite as far to the front and lies closer to the English /ɑ/ [24, 25].

We have obtained results for the vowel /ɑ/ within a single word context, that of the word *are*. From the first row of Tables 3.8 and 3.10 ("aa") we can see that:

- While both L1 and L2 speakers use large F1 values, F2 differs significantly be-

| Vowel | Context | DOF | F1 | F2 | F3 | Score |
|---|---|---|---|---|---|---|
| aa | are_aa | 31 | 0.02 | 22.55 | 2.04 | 2 |
| ae | africa_ae | 27 | 0.00 | 36.73 | 0.08 | 2 |
| | and_ae | 30 | 8.18 | 3.11 | 0.20 | 2 |
| | as_ae | 26 | 8.87 | 7.92 | 0.06 | 4 |
| | at_ae | 25 | 11.38 | 0.71 | 1.20 | 2 |
| | back_ae | 20 | 5.01 | 2.97 | 0.06 | 1 |
| | can_ae | 22 | 5.37 | 1.10 | 2.21 | 1 |
| | has_ae | 25 | 4.08 | 0.14 | 0.04 | 0 |
| | have_ae | 38 | 12.58 | 0.25 | 2.42 | 2 |
| | that_ae | 34 | 0.01 | 0.14 | 0.01 | 0 |
| ah | but_ah | 35 | 1.39 | 2.45 | 0.98 | 0 |
| | come_ah | 22 | 0.56 | 5.03 | 1.38 | 1 |
| | just_ah | 21 | 1.26 | 3.51 | 0.50 | 0 |
| | other_ah | 36 | 1.75 | 6.24 | 0.45 | 1 |
| | some_ah | 26 | 1.09 | 0.02 | 0.00 | 0 |
| ao | all_ao | 30 | 36.95 | 0.04 | 4.79 | 3 |
| | also_ao | 26 | 7.64 | 5.17 | 6.00 | 3 |
| | because_ao | 25 | 2.49 | 6.16 | 3.13 | 1 |
| | for_ao | 25 | 0.00 | 0.94 | 1.15 | 0 |
| | four_ao | 22 | 12.01 | 0.01 | 0.91 | 2 |
| | from_ao | 33 | 12.29 | 0.90 | 0.01 | 1 |
| | got_ao | 21 | 3.41 | 2.77 | 2.66 | 0 |
| | more_ao | 24 | 18.50 | 1.70 | 7.40 | 3 |
| | not_ao | 38 | 5.99 | 0.01 | 1.41 | 1 |
| | of_ao | 27 | 5.04 | 1.95 | 0.14 | 1 |
| | on_ao | 31 | 6.92 | 3.73 | 4.08 | 2 |
| ax | africa_ax | 19 | 0.58 | 6.85 | 1.22 | 1 |
| axr | other_axr | 34 | 1.73 | 8.73 | 0.29 | 2 |
| eh | get_eh | 27 | 16.81 | 8.64 | 2.98 | 4 |
| | sabc_eh | 21 | 19.44 | 17.29 | 0.04 | 4 |
| | said_eh | 20 | 3.08 | 3.01 | 0.56 | 0 |
| | very_eh | 21 | 5.41 | 3.18 | 0.02 | 1 |

| Vowel | Context | DOF | F1 | F2 | F3 | Score |
|---|---|---|---|---|---|---|
| el | people_el | 29 | 5.70 | 0.18 | 2.55 | 1 |
| er | first_er | 18 | 6.01 | 0.00 | 0.37 | 1 |
| ih | this_ih | 41 | 2.71 | 24.70 | 0.86 | 2 |
| iy | africa_iy | 19 | 3.82 | 0.35 | 4.89 | 1 |
| | be_iy | 33 | 0.30 | 13.36 | 0.21 | 2 |
| | been_iy | 27 | 0.04 | 8.99 | 4.91 | 3 |
| | being_iy | 20 | 4.62 | 2.93 | 1.72 | 1 |
| | he_iy | 22 | 2.83 | 9.27 | 3.38 | 2 |
| | his_iy | 19 | 1.45 | 1.97 | 3.85 | 0 |
| | if_iy | 22 | 6.03 | 0.01 | 0.04 | 1 |
| | in_iy | 24 | 0.20 | 0.01 | 0.00 | 0 |
| | is_iy | 31 | 0.66 | 2.12 | 0.16 | 0 |
| | its_iy | 22 | 10.20 | 0.00 | 0.89 | 2 |
| | people_iy | 26 | 3.67 | 3.19 | 9.30 | 2 |
| | sabc_iy1 | 20 | 0.00 | 6.90 | 0.03 | 0 |
| | sabc_iy2 | 19 | 0.19 | 16.27 | 6.64 | 2 |
| | see_iy | 19 | 2.79 | 4.14 | 0.47 | 1 |
| | very_iy | 25 | 0.04 | 6.65 | 4.88 | 3 |
| ow | also_ow | 23 | 0.41 | 12.40 | 0.20 | 2 |
| | go_ow | 20 | 5.25 | 64.15 | 1.77 | 3 |
| | know_ow | 20 | 4.18 | 88.47 | 0.46 | 2 |
| | no_ow | 25 | 4.37 | 25.47 | 5.13 | 4 |
| | so_ow | 31 | 7.07 | 46.10 | 7.15 | 4 |
| | those_ow | 19 | 193.16 | 103.61 | 0.08 | 4 |
| uw | do_uw | 19 | 3.01 | 25.49 | 0.35 | 2 |
| | to_uw | 26 | 0.08 | 20.78 | 0.01 | 2 |
| | two_uw | 28 | 1.94 | 63.30 | 2.17 | 2 |

Table 3.10: Analysis of variance results for vowel formant comparisons.

tween the two groups with the L1 mean at 1149 Hz and L2 much higher at 1355 Hz.

- This difference in F2 also translates to a second language /ɑ/ which is pronounced significantly more to the front, closer to the English /a/ or /ʌ/, as is evident from Figure 3.10. This seems to confirm the research of Flege [33] on the effect of equivalence classification which leads to vowel substitution as described by Arslan et. al [1]. In this case the second language speakers use the native /a/ when pronouncing *are*, instead of the English /ɑ/.

### /æ/ (/AE/)

The vowel /æ/ is a low, front vowel with large F1 and central F2 in English. This is not a valid phoneme in the native African languages of South Africa [21, 5, 7]. The closest vowel in these languages would be /ɛ/ which is a higher (more closed) vowel.

Our results for this vowel were obtained from a number of different word contexts, as can be seen from Table 3.10 and the graphs in Figure 3.10. The general trend seems to be the same for most contexts, although there are some exceptions as described below:

- In the contexts of *and*, *at*, *as*, *back* and *have* the first language /æ/ lies close to the Peterson and Barney reference vowel, while the second language /æ/ is consistently pronounced higher, in the region of /ɛ/. In all cases significant differences are found in F1. In the case of *as* some neutralisation seems to have affected the data, where an unstressed /ae/ tends more towards the neutral /ə/ and significant movement in F2 is also seen. Of note is the /æ/ in *and*, where the second language distribution has a much larger variance and actually contains the first language distribution. This may be expected, as some L2 speakers pronounce this vowel similar to the L1 phoneme, while others exhibit a more marked accent.

  These results seem to confirm the phonetic work of Wells [6] who states that the front vowels of SAE are pronounced more closed, specifically /æ/ is pronounced as [ɛ]. This would also seem to be a classic case of vowel substitution as suggested by Flege [33], where the unfamiliar phoneme /æ/ is replaced by the native /ɛ/.

  Schmied [3], however indicates that East African English (which also uses a five-vowel system, like the Nguni languages) uses the same /æ/ as in RP, while first language SAE uses the closer /ɛ/. Our results and other phonetic studies of the African languages [5, 7, 21, 25] seem to disagree with this, as /æ/ is not a valid phoneme in the Southern African native languages.

- A notable exception to the results above, is the pronunciation of /æ/ in the context of the word *africa*. In this case we find a drastic shift in F2, rather than

F1, as the L2 /æ/ is replaced with /ʌ/ or /a/. This may again be a indication of vowel substitution, but since this word is familiar with many native speakers in their own tongue or in Afrikaans [ɑ:frika], this may rather be a case of *word* substitution.

- In the case of /æ/ in *that* L1 and L2 speakers seem use the same phoneme, located somewhat higher and more central than the reference vowel. In both cases the distributions have a large variance which covers the range from /æ/ to /ɛ/ which suggests a wide range of different pronunciations in both language groups. In the context of *can* and *has* we also find no or little significant change from the ANOVA test. The graphs indicate that although the L2 means lie higher (smaller F1) in both cases, a large variance in F1 from the L1 speakers causes the distributions to overlap.

### /ʌ/ (/AH/)

Roach [15] defines /ʌ/ as a central vowel, lying somewhat to the back towards /ɔ/ and lower towards /ɑ/. Although /ʌ/ does not explicitly exist as a phoneme in the Southern African languages [5, 7], it lies in the proximity of native /a/, especially in the Nguni languages [25] where /a/ lies fairly to the back.

A number of words exists in the final data set, containing this vowel. In most contexts no appreciable difference was measured between L1 and L2 pronunciation.

- From the graphs in Figure 3.11 there seems to be two distinct variations for the phoneme /ʌ/, which is used by both first and second language speakers. In the contexts of *come*, *other* and *some* the vowel is pronounced close to the Peterson and Barney reference vowel, while it is pronounced higher (with smaller F1) and more to the front (larger F2) in the contexts of *but* and *just*.
- For the words *but*, *just* and *some*, no significant difference exists between the

L1 and L2 data. In the case of *come* and *other* a 95% significant difference is seen in F2, where L2 speakers use a smaller (more peripheral) F2 than their L1 counterparts. This seems to agree with Schmied's findings [3] where he states that central vowels (such as /ʌ/, /ə/ and /ɜ:/) are pronounced more peripherally, although the effects on /ʌ/ seem to be marginal.

### /ɔ/ (/AO/)

/ɔ/ is regarded as a back vowel, which is central in height and exists as a phoneme in both the Sotho and Nguni languages [24, 22]. From the analysis of variance results in Table 3.10 we see that most of the differences in pronunciation manifests as a change in the first formant. Looking at the graphs in Figure 3.12 on page 78 we see that:

- In the contexts of *all*, *also*, *four* and *more*, the L2 pronunciation is significantly lower with a larger F1 as also indicated by the analysis of variance results. In all cases we see a 99% level of certainty difference, except for *also* where the 95% level was reached in all three formant dimensions. The first and second language versions of /ɔ/ in *for* were statistically identical indicating no difference in pronunciation within this context.

- For the words *from*, *not*, *of* and *on* we also observe a significant change in F1, but in these cases the L2 distributions lie higher (with a smaller F1). The /ɔ/ in *got* displays the same movement, although it did not reach statistical significance.

- An exception was the /ɔ/ in *because*, where the first language speakers used a completely different phoneme, close to central /ə/, while the L2 phoneme lies more peripheral near /ɔ/. This manifests as a significant difference in F2 with a 99% certainty (from Table 3.10). We suspect this is due to centralisation by L1 speakers of the unstressed syllable containing the /ɔ/ in *because*, while the L2 speakers pronounce the word with equal stress on both syllables.

From these results it seems that first language speakers of SAE use two different phonemes for the vowel /ɔ/. The long form of the vowel (/ɔ:/) as in *all*, *more*, etc. is pronounced higher, in the direction of /ʊ/ (/UH/), while the short vowel (/ɔ/) as in *not*, *on*, etc. is pronounced lower and more central, towards /ʌ/ (/AH/). In contrast, second language speakers do not seem to discriminate between the two variants and pronounce both as /ɔ/ (/AO/).

These results agree with Lanham [26] who states that long vowels such as /ɔ:/, /a:/ and /ə:/ are pronounced unusually high in "White" South African English (WSAE) as compared to RP. The consistent use of native /ɔ/ by second language speakers agrees with the equivalence classification and vowel substitution theories of Flege [33] and Arslan et. al [1].

### /ə/ (/AX/)

The neutral /ə/ (/AX/) is a central vowel, pronounced with the vocal tract relaxed. It is often used as a substitute for another vowel in unstressed syllables (weak form of that vowel) for instance *because* pronounced as [bikə:z] instead of [bikɔ:z]. Although /ə/ is used phonemically in English [19] it is not a valid phoneme in the native African languages, which tend to avoid central vowels [14].

We only gathered data for /ə/ from a single word context as can be seen from the analysis of variance table on page 66. From the graph in Figure 3.11 on page 77 we see:

- In the context of *africa*, L2 /ə/ is a pronounced more to the back with a significantly smaller F2, in the direction of /ʌ/ (/AH/). This may well agree with Schmied [3] who states that central vowels tend to be pronounced more peripherally in general. On the other hand the familiarity of the word to speakers of native African languages may again influence their pronunciation of the word as

a whole ([ɑ:frikɑ] instead of [æfrikə]).

### /ɝ/ and /ɚ/ (/ER/ and /AXR/)

This is another central vowel (like /ə/), found in the context of words like *bird* or *sir*. It is transcribed with either the /AXR/ or /ER/ ARPABET symbols [17]. We can therefore also expect peripheralisation in the second language speech as this is not used as a phoneme in the African languages. From the graphs on pages 77 and 80 we find:

- As in the case of /ə/, we see from Figure 3.11 that in the context of *other*, the second language /ɚ/ (/AXR/) is positioned more peripherally with a significant shift to a larger F2 (to a certainty of 99%). The same peripheralisation is present for the /ɝ/ (/ER/) in *first* (Figure 3.14), where we have a 95% level significant movement to a larger F1. In these cases the word context is not familiar (as in the case of *africa*) and we can safely say that these results agree well with Schmied's findings [3] of peripheralisation and possibly also Flege's vowel substitution [33] where the unfamiliar phoneme /ɜ:/ or /ə/ is replaced by the native /a/.

### /ɛ/ (/EH/)

The /ɛ/ (/EH/) as in *get* is a front vowel and central in height. It exists as a valid phoneme in both English and the Southern African languages [15, 5]. From the analysis of variance results and the graphs on page 77 we see the following:

- In all cases we see the same trend, where the second language phoneme is pronounced markedly lower (larger F1 and smaller F2), closer to the Peterson and Barney reference vowel for /ɛ/ (/EH/), while first language speakers use a higher phoneme, closer to /ɪ/ (/IH/). In the contexts of the word *get* and the acronym for the *South African Broadcasting Corporation* (*SABC*, pronounced [ɛseɪbisi:])

the shift was significant (99%) in both F1 and F2, while for *very* only F1 reached significance (95%). In the case of *said* the data did not show statistically significant change, although the L2 mean moved in the same direction as in the other cases.

These results agree with previous studies, if we assume that the L2 speakers are substituting the L1 /ɛ/ with their lower native phoneme (Arslan et al. [1]), while the L1 speakers use the raised version of /ɛ/ (as compared to RP) described by Wells [6].

### /l̩/ (/EL/)

The /l̩/ (/EL/) as in *people* or *table* is also defined as a central vowel, but it exists only in combination with the subsequent pronunciation of the liquid /l/ ([teɪbl̩]). We only have data for this phoneme from the context of the word *people* and from the graphs on page 80 we see:

- The L2 pronunciation of /l̩/ lies higher and more peripheral towards /u/ (/UW/) ([pi:pul]), while the L1 phoneme is pronounced lower in the region between the Peterson and Barney /ə/ (/AX/) and /ʊ/ (/UH/) ([pi:pəl]). We also see from the analysis of variance results (page 66) that we have a statistically significant shift in F1 with a 95% level of certainty.

Schmied [3] states that central English vowels are often avoided and are pronounced more peripherally, which seems to be supported by these results.

### /ɪ/ (/IH/)

The /ɪ/ (/IH/) as in *pin* is defined as a high and front central vowel, between /ə/ and /i/ (/AX/ and /IY/). This phoneme is not valid in the native Southern African

languages [21, 25] and the closest vowel in their reference frames is the peripheral /i/. Looking at the graphs on page 80 we find:

- A large shift in F2 is present (99% level of certainty). First language speakers use a rather centralised version of /ɪ/ close to /ə/ (as was suggested as a hallmark of SAE by Lanham [26] and Wells [6]), while second language speakers use a phoneme much closer to the cardinal /i/ (/IY/). This tendency to replace /ɪ/ by /i/ is surely a case of vowel substitution as suggested by Flege [33] where the unknown (and central) /ɪ/ is replaced by the native /i/.

### /i/ (/IY/)

This vowel is defined as one of the cardinal vowels, as it lies on the very border of vowel space. It is pronounced with a high and front tongue position. This phoneme is present in both English and the African languages [14, 5].

From the analysis of variance results (Table 3.10), we see many words containing the vowel /i/. These results in combination with the graphs in Figure 3.13 on page 79 indicate the following:

- The general trend for the long vowel /i:/ is that L1 speakers use a more cardinal (front) /i:/ compared to the L2 pronunciation. The L2 phoneme is located to the back and slightly higher, which translates in significant changes in F2 and sometimes in F1. The same trend is seen in the case of *see*, but the translation was not statistically significant.
- In other cases (for the short form of the vowel), as in the contexts of *in*, *is* and *his* no significant movement is seen. In other contexts such as *if* and *its* we see L2 pronunciation to be significantly higher (smaller F1) than that of first language speakers.

### /o/ (/OW/)

Some texts consider /o/ as a vowel [15], while others regard it either as a short diphthong [46] or as a phoneme with dual status [17]. We have included this phoneme in both our vowel and diphthong experiments and give results for both cases. As a vowel, /o/ is defined as a back vowel, which is central in height. It does not explicitly exist as a phoneme in the Southern African languages [5] and the closest native phoneme is /ɔ/.

From the graphs on page 80 we see:

- A major shift in F2 is evident in all cases (99% certainty level), with the L2 speakers using a dramatically more peripheral (back) vowel near /ʊ/ (/UH/) compared to the first language speakers who pronounce /o/ close to neutral /ə/ (/AX/). The second language phoneme is generally also somewhat higher, resulting in significant movements in F1 in most cases. An extreme example is /o/ in the context of *those* where we have ANOVA $F$ ratios of 193.16 and 103.61 for F1 and F2 respectively (from Table 3.10 on page 66).

These results certainly agree with Schmied [3] on the issue of decentralisation, but seems to contradict his statement that short closing diphthongs (specifically /əʊ/) are monophthongised by L2 speakers because the *second* element is hardly heard. Our results indicate the opposite: The second element(/ʊ/) is dropped by *L1 speakers* (resulting in a central vowel near /ə/ (/AX/)). The opposite seems to be the case for L2 speakers, where the *first* element /ə/ (/AX/) is dropped resulting in a peripheral vowel near /ʊ/ (/UH/).

### /u/ (/UW/)

This vowel is regarded as a far back, high cardinal vowel, which lies more peripheral than /ʊ/ - therefore [bu:t] (*boot*) rather than [pʊl] (*pull*). The phoneme /u/ is used in both English and the native South African languages [15, 5].

The formant analysis results for this vowel are shown in Figure 3.14 on page 80 and as the final block in the analysis of variance results (Table 3.10 on page 66). From this we can see the following:

- A major shift in F2 is evident in all cases (99% certainty level), with the L2 speakers using a dramatically more back vowel near the Peterson and Barney /u/ (/UW/) compared to the first language speakers who surprisingly pronounce /u/ as a much closer (front) vowel. The L1 utterances for the three word contexts (*do*, *to* and *two*) all have tight distributions with small variances in F2, compared to the much wider L2 variances. This agrees with the hypothesis that some L2 speakers pronounce /u/ closer to the L1 phoneme, while others have a more marked accent.

A higher-level discussion and summary of the detailed results given here are presented in the concluding section of this experiment on page 81.

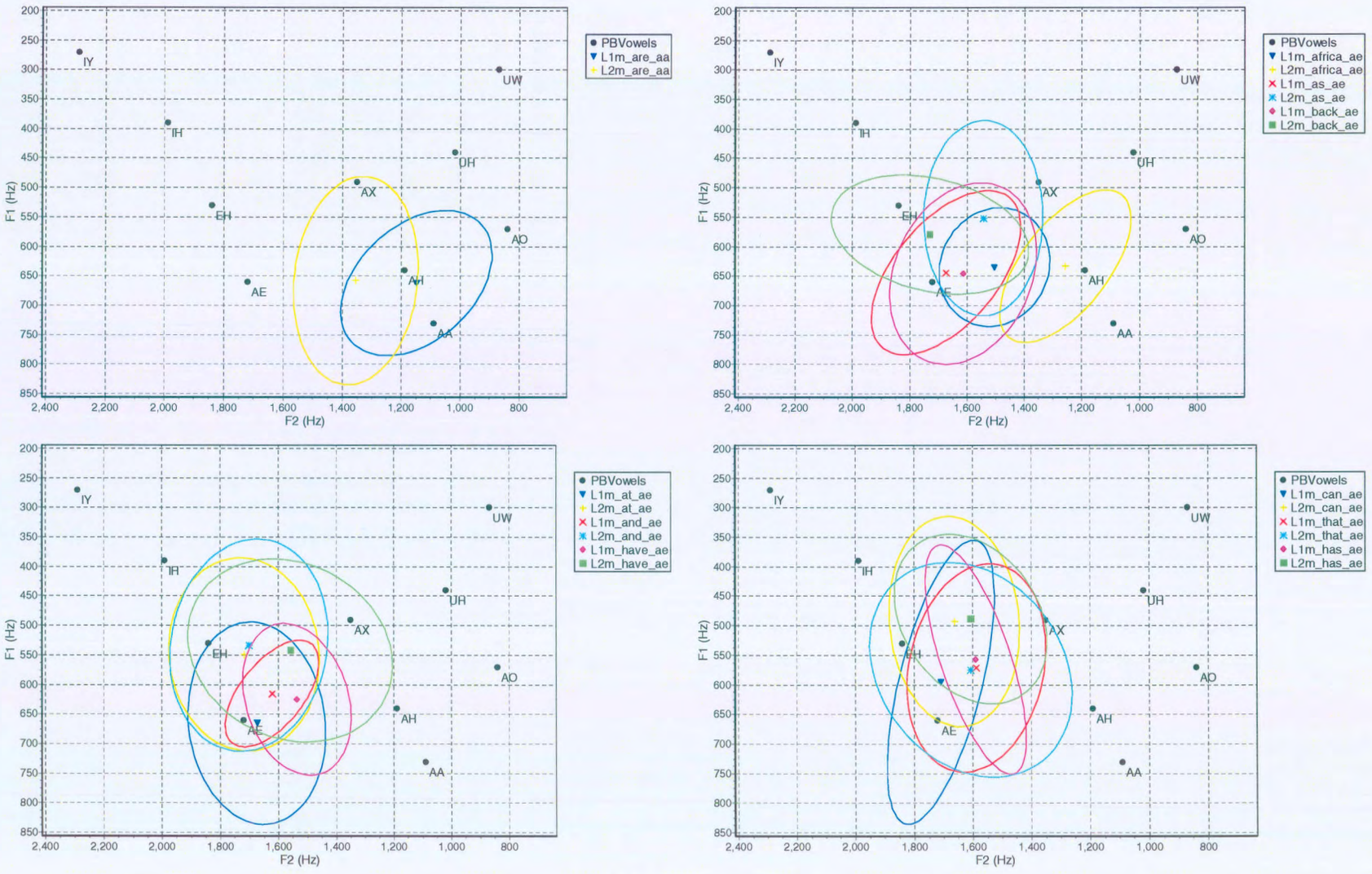


Figure 3.10: Formant results for the vowels /ɑ/ and /æ/ (/AA/ and /AE/).

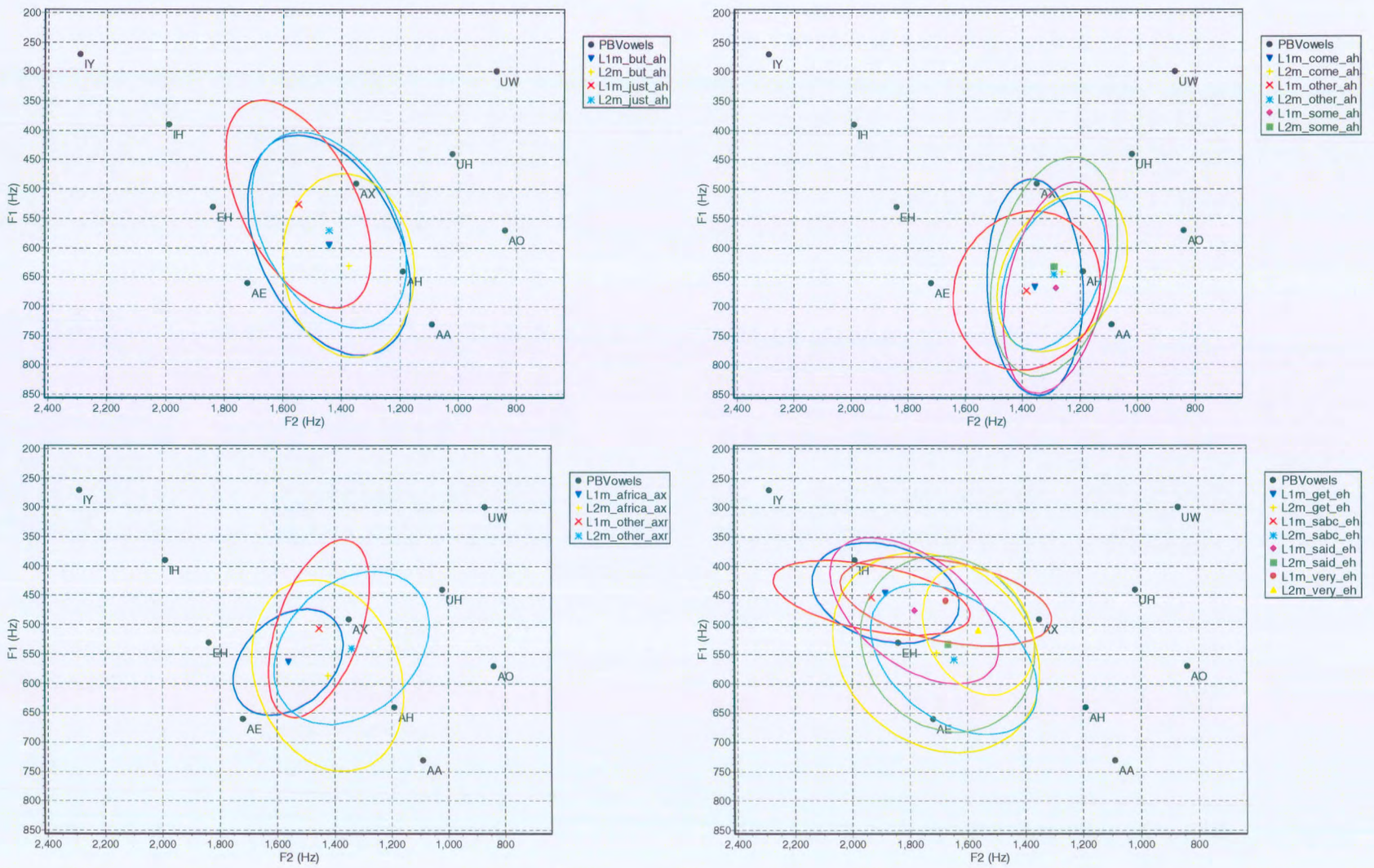


Figure 3.11: Formant results for the vowels /ʌ/, /ə/, /ɚ/ and /ɛ/ (/AH/, /AX/, /AXR/ and /EH/).

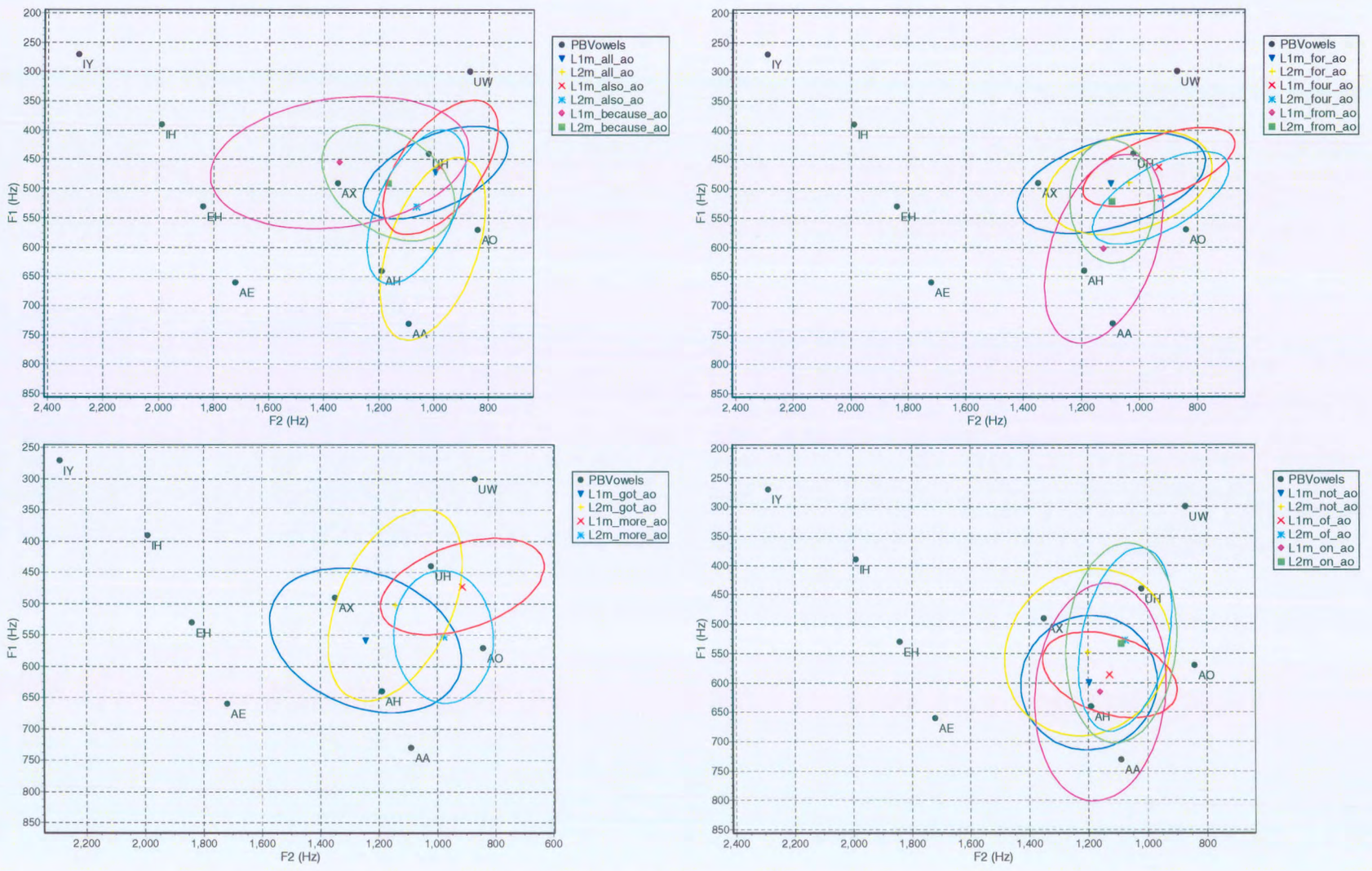


Figure 3.12: Formant results for the vowel /ɔ/ (/AO/).

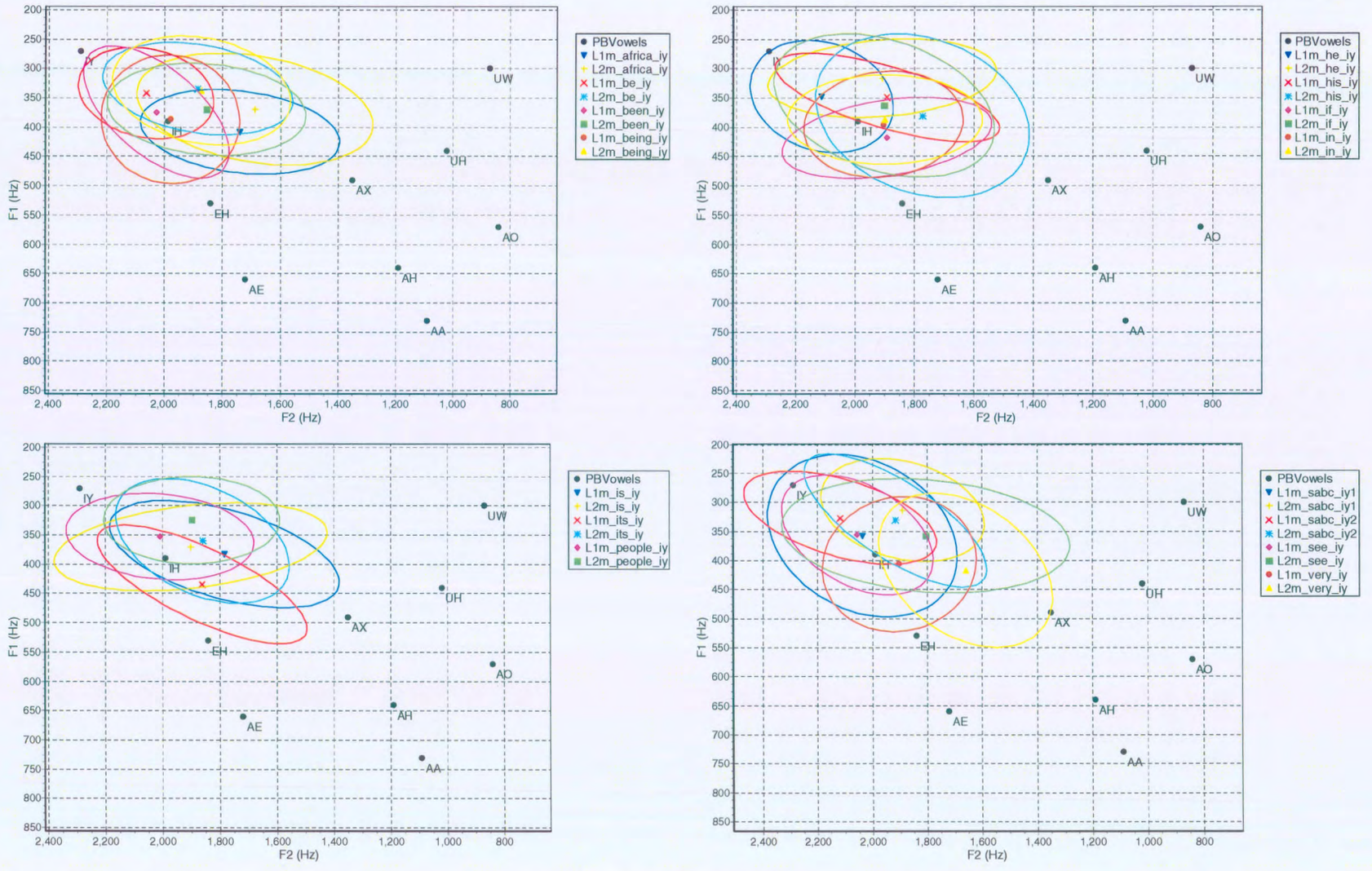


Figure 3.13: Formant results for the vowel /i/ (/IY/).

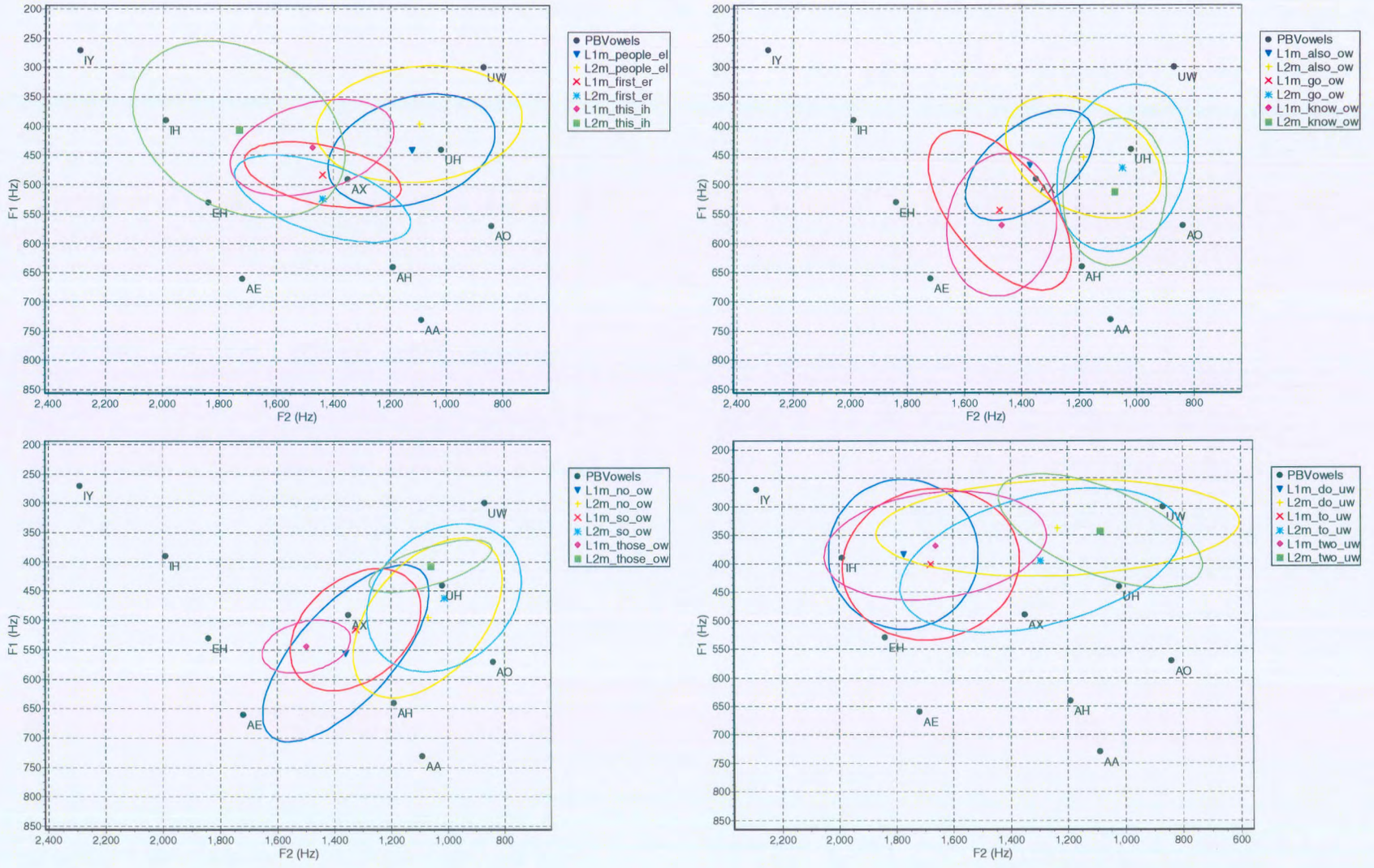


Figure 3.14: Formant results for the vowels /l̩/, /ɝ/, /ɪ/, /o/ and /u/ (/EL/, /ER/, /IH/, /OW/ and /UW/).

### 3.3.3 Conclusion

In this experiment the difference in formant distribution of 59 context dependent vowels were compared for first and second language South African English speakers. Significant differences were found in many cases, and a good agreement was found between our results and previous phonetic studies of the accents of English in Africa. It is clear from the analysis of variance results in Table 3.10 on page 66 that the first and second formants (F1 and F2) were the most sensitive to L2 accent. The lack of variation in F3 may indeed indicate that this frequency range is less strongly affected by African English accent, or it may be due to the fact that F3 is difficult to estimate accurately and that the Split Levinson algorithm chose a "default" value where no clear F3 could be determined. The F1 results are in disagreement with the findings of Arslan et al. [1] who state that the F2-F3 range is more sensitive to accent variations. We propose that in African English accents F1 is also significantly affected.

It is also clear that word context has a significant effect in second language pronunciation of the same (context independent) phoneme. An example is the L2 version of the word *africa* (L1 [æfrikə]) in which the L2 /æ/ moves in a completely different direction ([ɑ:frika]) as compared to the L2 /æ/ in other contexts such as *as*, *back* or *have* ([ɛz], [bɛk] or [hɛv]).

The vowels most seriously affected were those which do not exist as valid phonemes in the native South African languages (such as /ɑ/, /æ/, /ə/ and /ɪ/), where the mother tongue phoneme is frequently substituted as suggested by Flege [33]. These substitutions may well lead to L2 pronunciations where minimal pairs (such as *man→men*) are eroded as suggested by Jacobs [8].

In some cases first language South African English (described as WSAE by Lanham [26] and Wells [6]) differed in surprising ways from the anticipated phonetic transcription. For instance /ɛ/ is pronounced significantly higher by L1 speakers than in RP, L1 /u/ and the /ɔ/ in *because* are pronounced as a central or even front vowels. In these cases

L2 speakers were closer to the RP transcription, as the native language /ɛ/ and /u/ and /ɔ/ were again used instead.

Central vowels such as /ə/, /ɜ:/ and the lateral approximant /l̩/ were avoided in L2 speech and pronounced more peripherally, towards /ʌ/, /ɑ/ or /u/, which agrees with Schmied [3].

The long and short form of vowels, such as /ɔ:/ in *all* and /ɔ/ in *not*, are also often pronounced as different phones by L1 speakers, while no clear distinction is made in L2.

The phoneme /o/, when treated as a vowel, is located dramatically more to the back (smaller F2) in second language speech (near /ʊ/) as compared to the L1 pronunciation (near neutral /ə/).

Vowels such as /ʌ/ and /i/ were not affected as much (although significant movements are present) and it seems that these L1 vowels lie in close proximity to native African language phonemes /a/ and /i/ [15, 5].

In the next section an analysis of diphthongs in formant space is presented.

## 3.4 Comparative formant analysis of diphthongs

In this experiment the diphthongs of first and second language speakers are compared in formant space. A total of 27 context dependent diphthongs, as listed in ARBAPET format in Table 3.11, are examined. It may be noted that the context-dependent phonemes based on /o/ (/OW/) are included and treated as the diphthong /əʊ/ in this section. Of note is also the inclusion of a number of 'unconventional' diphthongs: /ɛɚ/ (/EH-AXR/), /ɪɚ/ (/IY-AXR/), /ʊa/ (/WAA/), /ʊæ/ (/WAE/), /ʊɛ/ (/WEH/), /ʊl̩/ (/WEL/), /ʊə/ (/WIH/), /ʊɪ/ (/WIY/) and /ju/ (/YUH/). These sounds are

included in the diphthong analysis from a purely acoustical and analytical perspective. The paired-vowel structure results in the pronunciation of a voiced sound that changes in vowel quality over time, resulting in a dynamic sound similar to a dipthong. Since our analysis is performed within word context, these supra-phonemic structures can therefore be analysed using diphthong models even though they are not linguistically regarded as such.

The same visualisation software tool as in the vowel analysis is used, this time to plot the formant trajectories of diphthongs. In the next section the specific formant analysis technique used here is described in more detail, followed by the results. We then give our interpretation of these results and some final concluding remarks.

| | | |
|---|---|---|
| now-aw | they-ey | was-waa |
| south-aw | here-iyaxr | what-waa |
| our-awr | also-ow | well-wae |
| by-ay | go-ow | where-weh |
| time-ay | know-ow | will-wel |
| their-ehaxr | no-ow | with-wih |
| there-ehaxr | so-ow | we-wiy |
| cape-ey | those-ow | which-wiy |
| sabc-ey | one-waa | you-yuh |

Table 3.11: Context dependent diphthongs used for formant analysis.

### 3.4.1 Experimental protocol

While vowels are static in nature and formant-space comparison could be made by comparing mean values and variances, diphthongs had to be analysed differently due to their dynamic nature. As the vowel quality of a diphthong changes with time (Figure 3.15) it forms a trajectory in formant space. These formant tracks were analysed with the aid of cubic splines which model each trajectory as a set of spline coefficients

(as described in Section 2.2.2 on page 30).

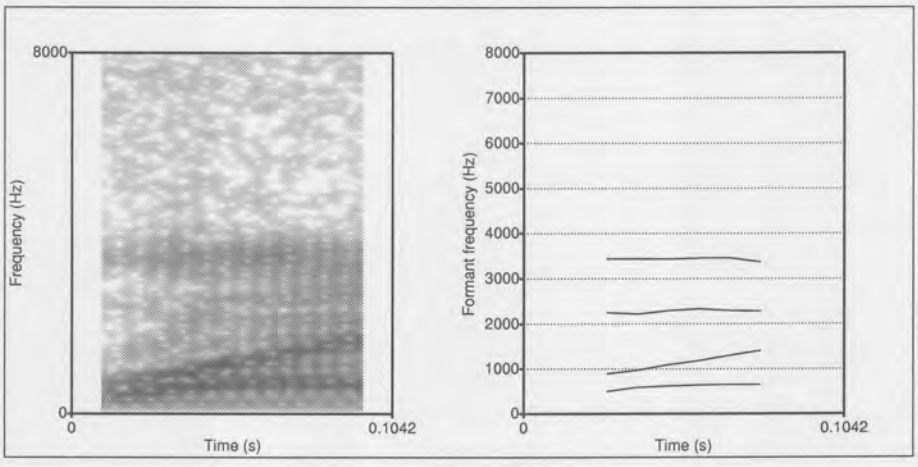


Figure 3.15: Varying formants for the diphthong /wa/ as in *one*.

In our data set, a diphthong's raw formant track typically contains around 5 to 15 data points in each formant dimension obtained at 10-millisecond intervals describing the trajectory. Each formant dimension is considered separately by interpolating the formant value over time using a cubic spline curve, which intersects each data point (Figure 3.16a). This time domain curve is then re-sampled (using the spline) to a fixed number of 128 samples - resulting in linear time scaling as all utterances now contain the same number of data points, irrespective of the original duration. The resulting curve is not very useful for direct comparison of formant trajectories, due to uncertainties in the extracted values of successive formant points of the utterance, which gives rise to a very complex spline representation. The curve is therefore simplified by dividing it into three sections of equal duration by calculating four points along the curve as shown in Figure 3.16b.

- Point 1 is set to the starting point of the trajectory.
- Point 2 is positioned along the time axis, one third from the origin. The formant value of this point is calculated as the mean value of the spline curve within a configurable region around this point.
- Point 3 is calculated like point 2, but at two-thirds from the origin.

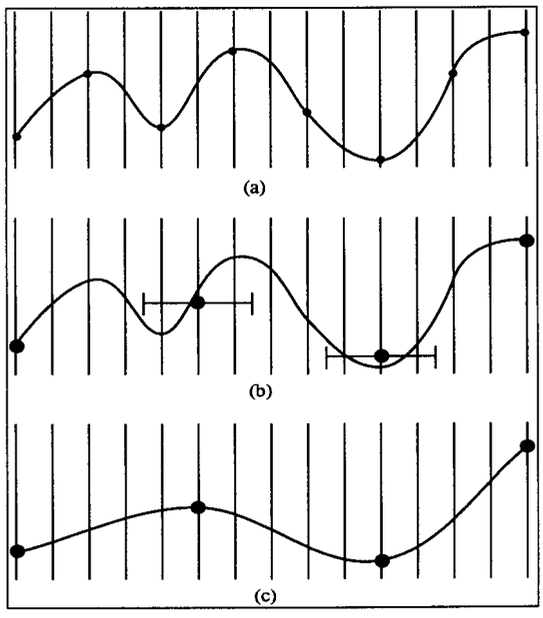


Figure 3.16: Fitting a cubic spline to arbitrary data.

- Point 4 is set to the final data point.

The original trajectory is now replaced by a simpler cubic spline intersecting these four points forming a *fitted spline* as in Figure 3.16c. Section $k$ of the fitted spline is defined by:

$$S_k(x) = s_{k,3}(x - x_k)^3 + s_{k,2}(x - x_k)^2 + s_{k,1}(x - x_k) + s_{k,0} \tag{3.3}$$

with $k = 0, 1, 2$ defining the three sections and $\{x_k\}$ specifying data points 1 through 3. Note that although all four points are used when the spline coefficients are calculated, the fourth point is not used thereafter (Eq. (3.3)), as each segment models the spline as an extrapolation from the data point at its origin. The complete fitted spline is defined by the constants $s_{k,m}$, with $m = 0, 1, \ldots, 3$ giving a total of 3 sections $\times$ 4 coefficients per section = 12 spline coefficients for each formant dimension. An example of such a fitted spline in F1-F2 space is shown in Figure 3.17.

By using this representation, formant trajectories are plotted implicitly over time in

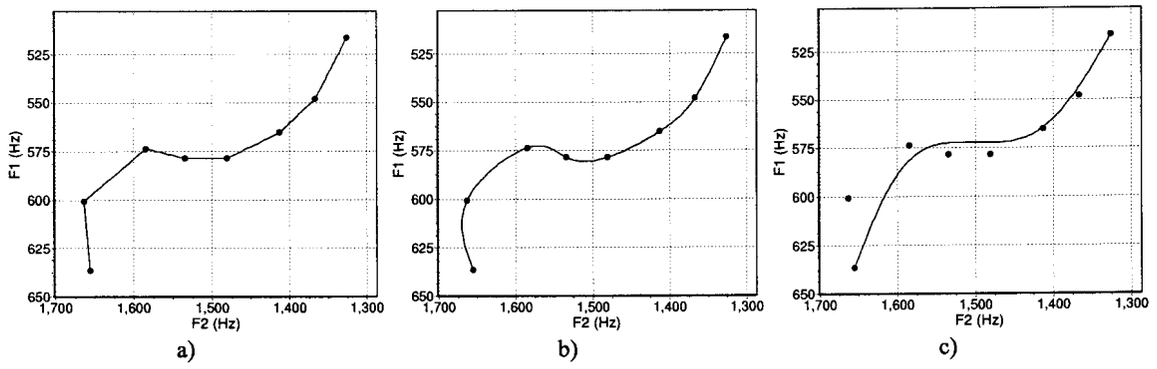


Figure 3.17: Example of splines derived from formant data: a) Linear interpolation, b) spline interpolation, c) fitted spline.

formant space, similar to previous work done on diphthongs by Holbrook and Fairbanks [47]. Two individual formant trajectories are therefore compared purely according to their shapes in formant (vowel) space (i.e. frequency domain) excluding all temporal features, which is what we wish to accomplish.

When the utterances of L1 speakers are to be compared against their L2 counterparts, the utterances of each language group need to be combined. This is accomplished by representing each of the 12 spline coefficients of the fitted spline as normally distributed random variables, with the coefficients of each individual utterance representing single instances of these variables. An example of this is shown in Figure 3.18, where we have the individual fitted splines of the utterances on the left, with the mean spline plotted on the right (the origins of the formant trajectories are circled). Note that in this representation only the mean is plotted and not the variance.

Each language group is therefore described by a 12-dimensional normal distribution for every formant dimension. The two groups can then be compared by applying the analysis of variance test to these distributions. These results are discussed in detail in the next section.

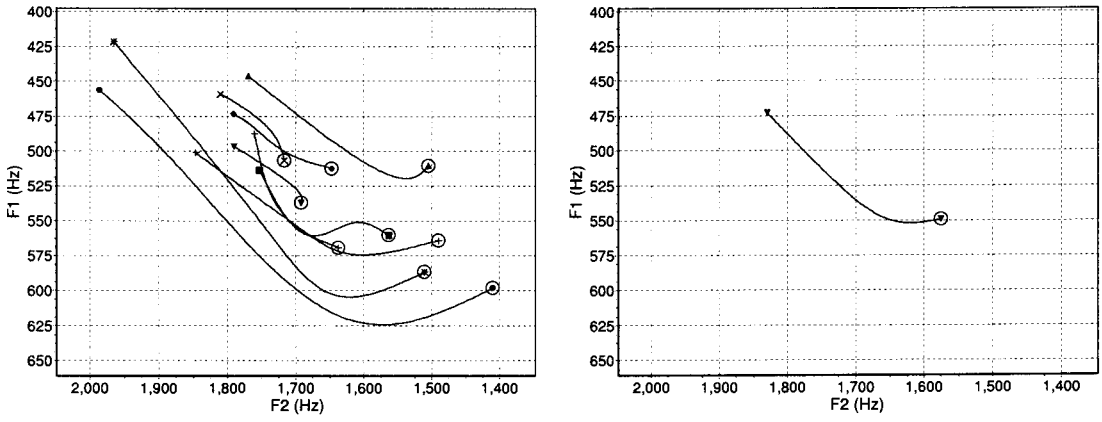


Figure 3.18: Example of how the mean formant trajectory for a language group is derived. *Left* indicates the fitted splines of the individual utterances, while on the *right* we have the mean fitted spline of the group as a whole.

### 3.4.2 Results and discussion

This section details the results found from the formant analysis of diphthongs described above and we describe the trends present in the data based on these results. The analysis of variance results are presented in Table 3.12 and is followed by figures where the mean formant trajectories used in the ANOVA tests are plotted to visualise the results. Throughout the tables and figures, phonemes are indicated using ARPABET symbols. These are augmented in the text by the I.P.A. symbols for ease of reference.

If we look at Table 3.12 the first column indicates the diphthong, while the second column indicates the context-dependent phoneme. The third column indicates the degrees of freedom ("DOF" = the total number of data points minus 2) of the calculated ANOVA $F$ ratio. Since diphthongs are compared in three sections, the following nine columns indicate where significant differences were found for the three formant dimensions.

If any of the four spline coefficients in a section differed significantly between language groups, the section was considered statistically different. A 95% certainty level is indicated by a light-grey block, while a 99% certainty is indicated by a dark-grey

| Diphthong | Context | DOF | F1 Section 1 | F1 2 | F1 3 | F2 1 | F2 2 | F2 3 | F3 1 | F3 2 | F3 3 | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| aw | now_aw | 26 | | | | | | | | | | 6 |
| | south_aw | 30 | | | | | | | | | | 5 |
| awr | our_awr | 24 | | | | | | | | | | 11 |
| ay | by_ay | 29 | | | | | | | | | | 12 |
| | time_ay | 22 | | | | | | | | | | 3 |
| ehaxr | their_ehaxr | 31 | | | | | | | | | | 16 |
| | there_ehaxr | 33 | | | | | | | | | | 0 |
| ey | cape_ey | 20 | | | | | | | | | | 10 |
| | sabc_ey | 20 | | | | | | | | | | 2 |
| | they_ey | 25 | | | | | | | | | | 12 |
| iyaxr | here_iyaxr | 23 | | | | | | | | | | 12 |
| ow | also_ow | 23 | | | | | | | | | | 8 |
| | go_ow | 20 | | | | | | | | | | 8 |
| | know_ow | 20 | | | | | | | | | | 10 |
| | no_ow | 25 | | | | | | | | | | 15 |
| | so_ow | 31 | | | | | | | | | | 16 |
| | those_ow | 19 | | | | | | | | | | 12 |
| waa | one_waa | 41 | | | | | | | | | | 1 |
| | was_waa | 27 | | | | | | | | | | 5 |
| | what_waa | 19 | | | | | | | | | | 5 |
| wae | well_wae | 21 | | | | | | | | | | 2 |
| weh | where_weh | 25 | | | | | | | | | | 10 |
| wel | will_wel | 27 | | | | | | | | | | 14 |
| wih | with_wih | 25 | | | | | | | | | | 10 |
| wiy | we_wiy | 32 | | | | | | | | | | 4 |
| | which_wiy | 24 | | | | | | | | | | 3 |
| yuh | you_yuh | 21 | | | | | | | | | | 9 |

Table 3.12: Analysis of variance results for diphthong formant comparisons.

block. The final column indicates the value of a difference score. This score is calculated similar to the one in Table 3.10 on page 66, where for every section a 95% level difference is counted as 1, while a 99% difference is counted as 2. This is then summed across all sections in all three formant dimensions. As an example, if we look at the first context (*now-aw*) for the diphthong /aʊ/ (/AW/), we see significant differences to a certainty level of 95% (light-grey blocks) in the second section for F1 and F2 and significant differences to a certainty of 99% in the third section for the same two formants. This

translates to a score of 0+1+2 for F1 and F2, and 0+0+0 for F3, giving a total score of 6. This score is therefore an indication of the relative difference between two diphthongs with 0 indicating no significant difference and 18 indicating 99% significant differences in all sections in all formant dimensions. The results for each of the diphthongs are discussed in more detail below.

### /aʊ/ and /aʊ˞/ (/AW/ and /AWR/)

/aʊ/ (/AW/) is a diphthong which starts low and back, near /ɑ/ and moves higher and further back, towards /ʊ/ [17]. The diphthong /aʊ˞/ (AWR) is similar in sound, but is found in the context of a subsequent /r/ as in *our*. We consider each context in our data set separately:

- From the first row of Table 3.12 we see that the L1 and L2 versions of the diphthong /aʊ/ in *now* start similar (no significant difference for section 1) and then progressively diverges from one another in both F1 and F2 (section 2 differs significantly to a certainty of 95% and section 3 to a certainty of 99%). If we compare this with the graph on page 99, we see that while both trajectories start in the same vicinity (origins are circled), the L2 diphthong (∗) quickly moves away to a more back and higher terminating point (smaller F1 and F2). The second and third sections of the L2 trajectory therefore differ progressively more from the L1 counterpart. L2 speakers therefore seem to pronounce the second element of /aʊ/ higher resulting in a more pronounced diphthong.

- In the context of *south* we see a different effect. From the graph we see definite monophthongisation where the trajectory of the L2 phoneme (+) displays much less movement in F2 than the L1 version (▼), with both origin and terminating point at almost the same position. This is indicated in the analysis of variance results by 99% certain significant differences in F2 for the first and second section. The third section shows a 95% certain significant difference, since here the L1

trajectory catches up with the L2 phoneme.

- In the case of /aʊˇ/ (/AWR/) in *our* we see strong changes in all sections of both F1 and F2 with 99% certain significant differences in all cases but one, and a score of 11. The reason for this strong difference can be seen in the graph on page 99, where the L2 trajectory (+) is located much higher and more peripheral (smaller F1 and F2) than the L1 trajectory (▼).

Wells [6] states that in the case of short, closing diphthongs the second element is often dropped or weakened. This monophthongisation is seen in the case of *south*, where the L2 phoneme is pronounced like the vowel /ʌ/ (/AH/) [sʌθ]. In the other cases we do not see this effect, as the /aʊ/ in *now* is retained and possibly even exaggerated, while the /aʊˇ/ in *our* is shifted to a more peripheral location in formant space.

### /aɪ/ (/AY/)

The diphthong /aɪ/ (/AY/) is defined as starting low and to the back and then moving higher and to the front of vowel space, in the direction of /ɛ/ and /ɪ/. From the graphs on page 99 and the analysis of variance table we see the following:

- In both the context of *by* and *time* the diphthong can clearly be seen to start in the region of the Peterson and Barney /ʌ/ (/AH/) and proceed to the front and higher, in the direction of /ɪ/ (IH). The diphthong in *by* (▼ and +) can also be seen to be more pronounced for both language groups.

- In both cases the L2 diphthong is pronounced significantly higher (smaller F1) as is also shown by the ANOVA results in Table 3.12. The /aɪ/ in *by* differs significantly to a certainty of 99% in all three sections for both L1 and L2, with a score of 12, while *time* shows a less prominent difference with a score of only 3.

These results seem to contradict the findings of Lanham [26] who states that the diphthongs most seriously monophthongised in SAE are /aɪ/ and /əʊ/. Our results show pronounced diphthongs for both language groups, with L2 speakers consistently using a higher pronunciation.

### /ɛɚ/ (/EH-AXR/)

The centering diphthong /ɛɚ/ is found when an /ɛ/ precedes the pronunciation of /r/ as in *bear* [bɛɚ]. From the graphs in Figure 3.20 on page 100 and the ANOVA results we can see that:

- In the context of *their* drastic diphthongisation takes place where the L2 diphthong starts somewhat higher than the Peterson and Barney /ɛ/ (/EH/) and moves in the direction of /ʌ/ (/AH/). The L1 phoneme remains near /ɛ/ and is pronounced as a vowel. These differences can also be seen from the ANOVA results, where 99% certain significant differences are seen in all sections of F1 and F2 and in two sections of F3. This is also one of the few diphthongs where a significant difference in F3 is seen.

  This agrees well with Schmied [3] who states that the central English vowels /ə/ and /ɜ:/ are avoided in African accents. He continues that this affects centering diphthongs, which then become opening diphthongs, therefore [ðɛa] instead of [ðɛɚ] for *their*. From Flege's theory of equivalence classification [33] it may also be that the unknown diphthong element /ə/ is replaced by the closest native African vowel /a/ by the L2 speaker, leading to an over-emphasised diphthong.

- In the case of *there* the behaviour is completely different, with both language groups treating the /ɛɚ/ as a vowel between the Peterson and Barney /ɛ/ (/EH/) and /ɪ/ (/IH/) (near cardinal /e/). This similarity is also clearly visible in Table 3.12 where no significant differences were found in any of the formant dimensions.

Since this phoneme is treated as a vowel in this context, Schmied's findings as above do not apply. Instead we see that this result agrees with Wells [6] and Lanham [26] who indicate that SAE differs from British English in that /ɛə/ is pronounced closer (higher) and monophthongal therefore [ðe:] *there*. This seems to apply to both L1 and L2 pronunciations of *there* in our analysis.

### /eɪ/ (/EY/)

/eɪ/ is defined as a closing diphthong starting from near cardinal /e/ and moving higher in the direction of /ɪ/ [17]. Our results are plotted in Figure 3.19 on page 99.

- In the contexts of *cape* and *they*, we clearly see from the graphs that the L1 phoneme is a pronounced diphthong starting between the Peterson and Barney /ɛ/ and /ə/ (/EH/ and /AX/) and moving higher in the direction of /ɪ/ (/IH/). The L2 phoneme on the other hand, is a much weaker diphthong where the first element (/ə/) has been dropped, resulting in a vowel-like phoneme near /ɪ/ (/IH/). This clear difference also translates to significant differences in the ANOVA results on page 88, with scores of 10 and 12 for *cape* and *they* respectively.
- The results for the /eɪ/ in the acronym *SABC* does not follow the same trend as above. In this case both language groups seem to use diphthongs of similar strength, with the L2 phoneme lying somewhat higher (smaller F1). This is also reflected as a much lower ANOVA score of 2, with significant change to a certainty of 95% in F1.

The results for *cape* and *they* agree with Schmied [3] who states that closing diphthongs of L2 English are monophthongised. However, he states that the *second* element (/ɪ/ in this case) is dropped, where our results clearly show that it is only this element which is retained, therefore [kep] or [kɪp] instead of [keɪp].

### /ɪɚ/ (/IY-AXR/)

This is another centering diphthong caused by /r/ following the diphthong /ɪə/ as in *here*. In this word context, our results show:

- From the graph on page 99, we see that the L1 diphthong (♦) starts at the Peterson and Barney /ɪ/ (/IH/) and proceeds lower to terminate in the central region between /ɛ/ (/EH/) and /ə/ (/AX/). The L2 phoneme (■) has the same origin, but is a weaker diphthong, which only extends to around /ɛ/ (/EH/) and remains more peripheral. This is also seen from the ANOVA results, where sections 2 and 3 show more significant differences than the initial section. This is another diphthong which differs strongly in F3 and reaches an ANOVA score of 12.

Again the observation by Schmied [3] that centering diphthongs are avoided is supported, therefore *here* is pronounced more peripherally as [hɪɛ] instead of [hɪɚ], resulting in some monophthongisation. This may also be an indication of vowel substitution as proposed by Flege [33], where the unfamiliar second element /ə/ is replaced by the closest native front vowel /ɛ/.

### /əʊ/ (/OW/)

In this experiment the vowel /o/ is treated as the diphthong /əʊ/, which starts near central /ə/ and moves in the direction of /ʊ/ or more cardinal /u/. Our results are plotted in Figure 3.20 on page 100 and the analysis of variance results are listed in Table 3.12.

- In all cases high ANOVA significance scores are reached in F1 and F2, and in some cases even F3. From the graphs we see that in the contexts of *also*, *go*

and *so* we have strong monophthongisation by L2 speakers where the phoneme is pronounced as the vowel [ʊ]. In the cases of *know*, *no* and *those* we find diphthongs of similar strength for both language groups.

- In all cases we find that the L2 phoneme lies much more peripheral, close to the Peterson and Barney /ʊ/ (/UH/), while the L1 phoneme is a pronounced diphthong starting central, near /ə/ (/AX/) and moving higher and more peripheral, towards /u/ (/UW/).

As was stated in Section 3.3.2, these results agree with Schmied [3] as far as monophthongisation of closing diphthongs are concerned, but our results show that L2 speakers do not drop the final element of /əʊ/ as suggested, but rather the initial element. Therefore we get [ɔlsɔ] or [ɔlsʊ] instead of [ɔlsəʊ]. This may be particular to African L2 English speakers, where the first element /ə/ is not a valid phoneme in their native language and is replaced by /ɔ/ or its raised version /ɔ'/.

### /ʊa/ (/WAA/)

/ʊa/ is a diphthong induced by /w/ preceding the vowel /ɑ/ as in *was* or *one*. Our results are shown in Figure 3.21 on page 101 and Table 3.12. From this we see the following:

- No drastic changes are apparent in the L2 pronunciation of /ʊa/ where both language groups use a pronounced diphthong moving from near the Peterson and Barney /ʊ/ (/UH/) in the direction of /ʌ/ or /ɑ/. In the cases of *what* and *was*, L1 speakers seem to use a more centering (weak form) diphthong [ʊə] instead of [ʊa]. This may be caused by L1 speakers using the weak form of the phoneme in unstressed cases, whereas L2 speakers use an overcorrected form more consistently. These results are reflected in the ANOVA scores, where *one*

only differed significantly to the 95% certainty level in one section of F1, while *was* and *what* showed more change in F1 and F2 respectively.

### /ʊæ/ (/WAE/)

The diphthong /ʊæ/ starts at /ʊ/ (/UH/) at the back of vowel space and moves across to the front and lower, towards /æ/ (/AE/). Our results for the word *well* is shown in Figure 3.21 on page 101:

- We see similar formant trajectories for both language groups, indicating a pronounced diphthong starting near /ʊ/ (/UH/) and moving in the direction of /æ/ (/AE/) and terminating fairly central. In the third section, the L1 phoneme (▼) curves down and terminates somewhat lower and more central than the L2 diphthong (+). This may be the result of L1 speakers using a weakened version of the diphthong in unstressed instances, where the second element is neutralised and tends towards /ə/ (/AX/). The similarity between these two diphthongs can also be seen in the analysis of variance results of Table 3.12, where only the last section of F2 shows significant difference.

### /ʊɛ/ (/WEH/)

The diphthong /ʊɛ/ is also caused by a /w/ preceding a vowel, in this case /ɛ/, where pronunciation starts at the back of vowel space and moves to the front, towards /ɛ/. This can be seen from the graphs on page 101 for the word *where*. Here we see that:

- A pronounced diphthong is present for both language groups, with the L2 phoneme (∗) originating more peripheral, closer to /ʊ/ (/UH/). Both diphthongs proceed to the front of vowel space (larger F2) with little vertical movement (change in F1), and both terminate near the Peterson and Barney /ɛ/ (/EH/).

These characteristics can also be seen from the ANOVA results in Table 3.12, where we have 99% certain significant differences in all formant dimensions for the first section (indicating the difference in origin). Significant differences are also seen for the remaining two sections of F2, as the L2 phoneme trails behind the L1 version (smaller F2) for all three sections.

### /ʊl̩/ (/WEL/)

The pronunciation of the vowel /l̩/ in the context of a preceding /w/ leads to the centering diphthong /ʊl̩/, which also starts from the back of vowel space. Our results for the word *will* are shown in Figure 3.21 on page 101:

- We see evidence of strong diphthongisation, where the L1 phoneme (▼) is pronounced as a vowel, near /ʊ/ (/UH/), while the L2 phoneme (+) is a marked diphthong located higher and more central and moving in the direction of /ɪ/ (/IH/) rather than /ə/ (/AX/). Therefore we get L2 [wɪ], rather than [wl̩].

Vowel substitution and decentralisation as proposed by Schmied [3] seem to affect this centering diphthong also, where the second element of /ʊl̩/ is replaced with the higher and front vowel /ɪ/ (/IH/), leading to an over-articulated diphthong /ʊɪ/.

### /ʊə/ (/WIH/)

In the case of /ʊə/ the preceding /w/ results in a relatively weak, centering diphthong moving from /ʊ/ to the central /ə/. Our results for the word *with* (Figure 3.21 on page 101) indicate the following:

- For the L1 phoneme case (♦) we see a fairly weak diphthong which originates (circled) between /ʊ/ and /ə/ (/UH/ and /AX/) and terminates slightly more

central.

- In the L2 case (■) we see a much more marked diphthong, starting more peripheral and higher, between /ʊ/ and /u/ (/UH/ and /UW/). It then moves across vowel space, heading towards the front vowel /ɪ/ (/IH/).

These results are also evident from the analysis of variance scores where we see significant differences for all sections of F1 and F2. It seems this centering diphthong is also avoided by L2 speakers and the second element (/ə/) is replaced with a more peripheral /ɪ/, in agreement with Schmied [3]. Therefore L2 [wɪθ] or [wiθ] in stead of L1 [wəθ]. Another reason for the marked difference in the second element may also be the fact that L1 South African English (SAE) speakers pronounce /ɪ/ as in *pit* as a neutral /ə/ [pət]. This is indicated as one of the hallmarks of L1 SAE by Lanham [26], in which case the L2 pronunciation in our results may in fact be closer to the British (RP) pronunciation [wɪθ].

### /ʊɪ/ (/WIY/)

In this case a preceding /w/ or /ʍ/ leads to a diphthong starting near /ʊ/ and moving in the direction of cardinal /i/. We present the results for the words *we* and *which* in the graphs on page 101:

- In the case of *which* we see that the L2 phoneme (■) tends to be a more pronounced diphthong, starting more peripheral near /u/ (/UW/) and moving forward towards /i/ (/IY/). The L1 diphthong (♦) originates more central and moves in the direction of /ɪ/ or /i/ (/IH/ or /IY/), consistently remaining below (larger F1) the L2 phoneme. This is also evident in the analysis of variance results in Table 3.12, where significant differences are present in the second and third sections of F1.

- For the word *we* the difference in diphthong strength seems to be less pronounced, although the L2 phoneme (∗) also starts more peripherally and remains higher that its L1 counterpart (smaller F1). The ANOVA results also indicate this difference in F1 as a significant difference for all thee sections of this formant dimension.

### /ju/ (/YUH/)

This diphthong is formed when the phoneme /j/ precedes the vowel /u/ or /ʊ/ and leads to a diphthong starting in the high, front corner of vowel space (near /i/) and moving across to the back near /ʊ/. Our results for the word *you* are indicated in Figure 3.19 on page 99.

- We can clearly see a marked diphthong from the graphs, which for both language groups start near /i/ or /ɪ/ (/IY/ or /IH/) and moves to the back (smaller F2), in the direction of /ʊ/ (/UH/). The L1 phoneme (×) lies more to the front with larger F2 throughout the trajectory, leading to the significant differences indicated by the analysis of variance results on page 88. This is also a diphthong which shows significant difference in F3 indicated by 99% certain significant change in all three F3 diphthong sections.

The detailed results described here are summarised on a higher-level in the following section.

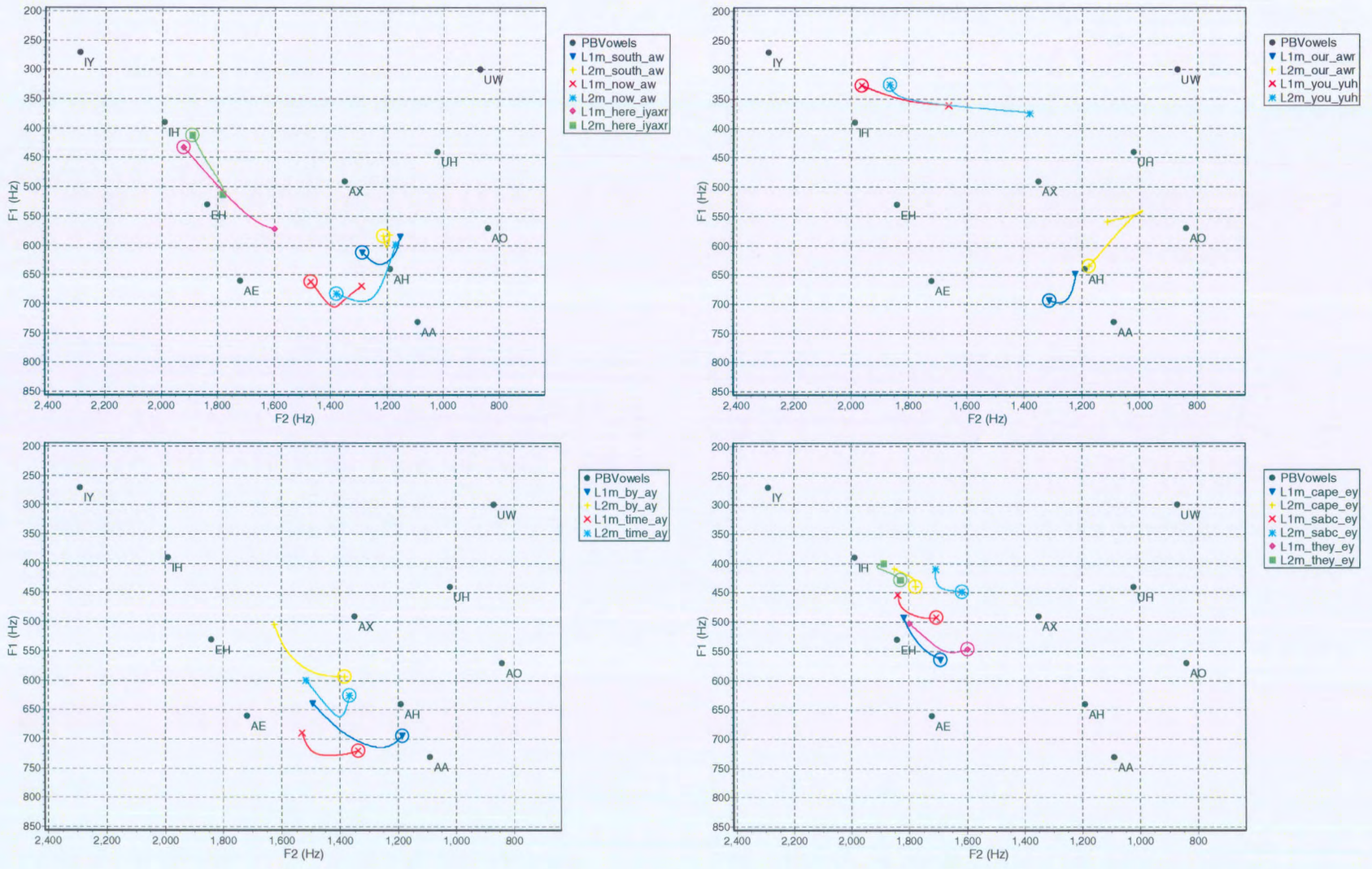


Figure 3.19: Formant results for the diphthongs /aʊ/, /aʊˠ/, /ɪəˠ/, /ju/, /aɪ/ and /eɪ/ (/AW/, /AWR/, /IY-AXR/, /YUH/, /AY/ and /EY/).

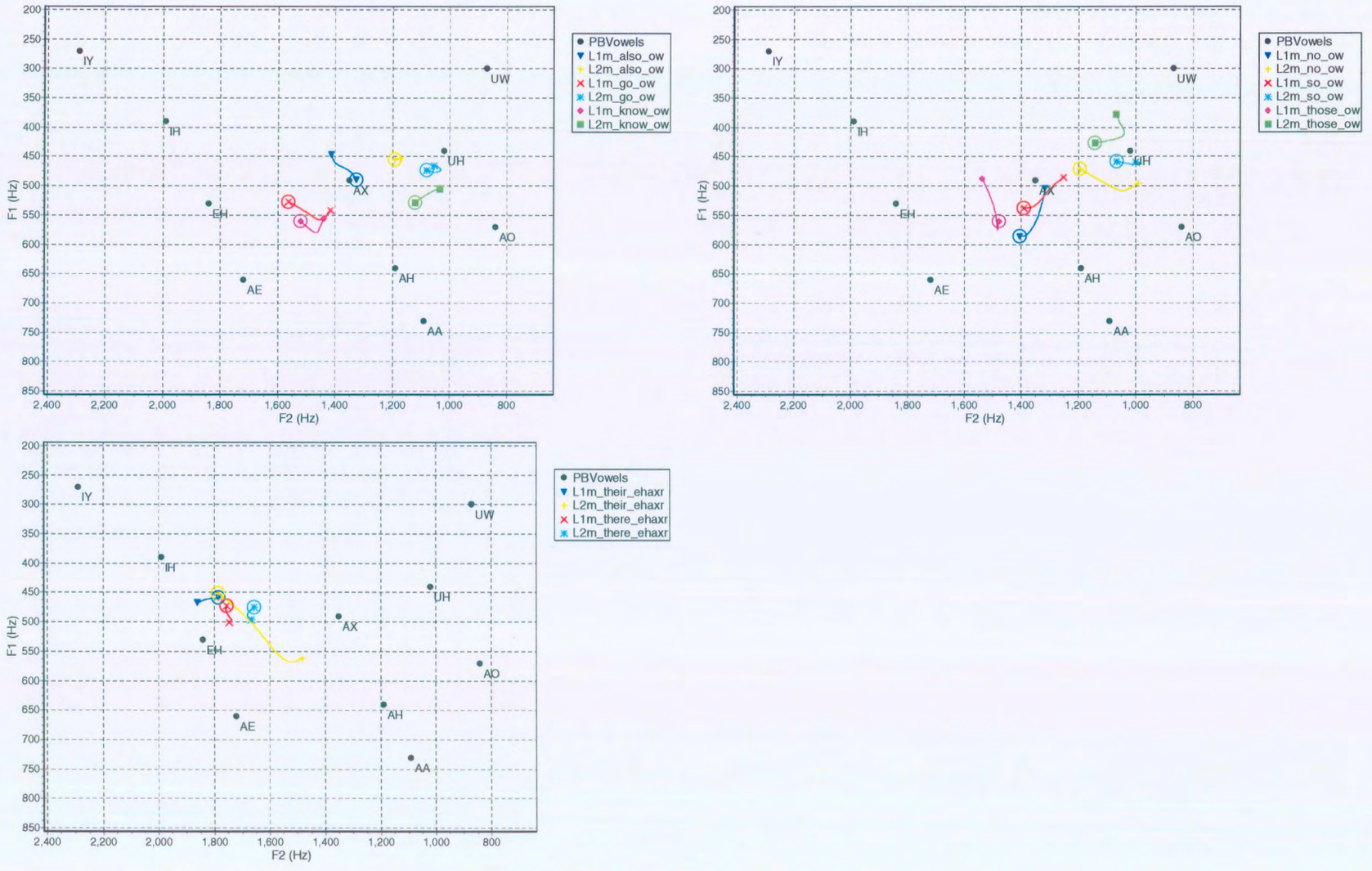


Figure 3.20: Formant results for the diphthongs /əʊ/ and /ɛɚ/ (/OW/ and /EH-AXR/).

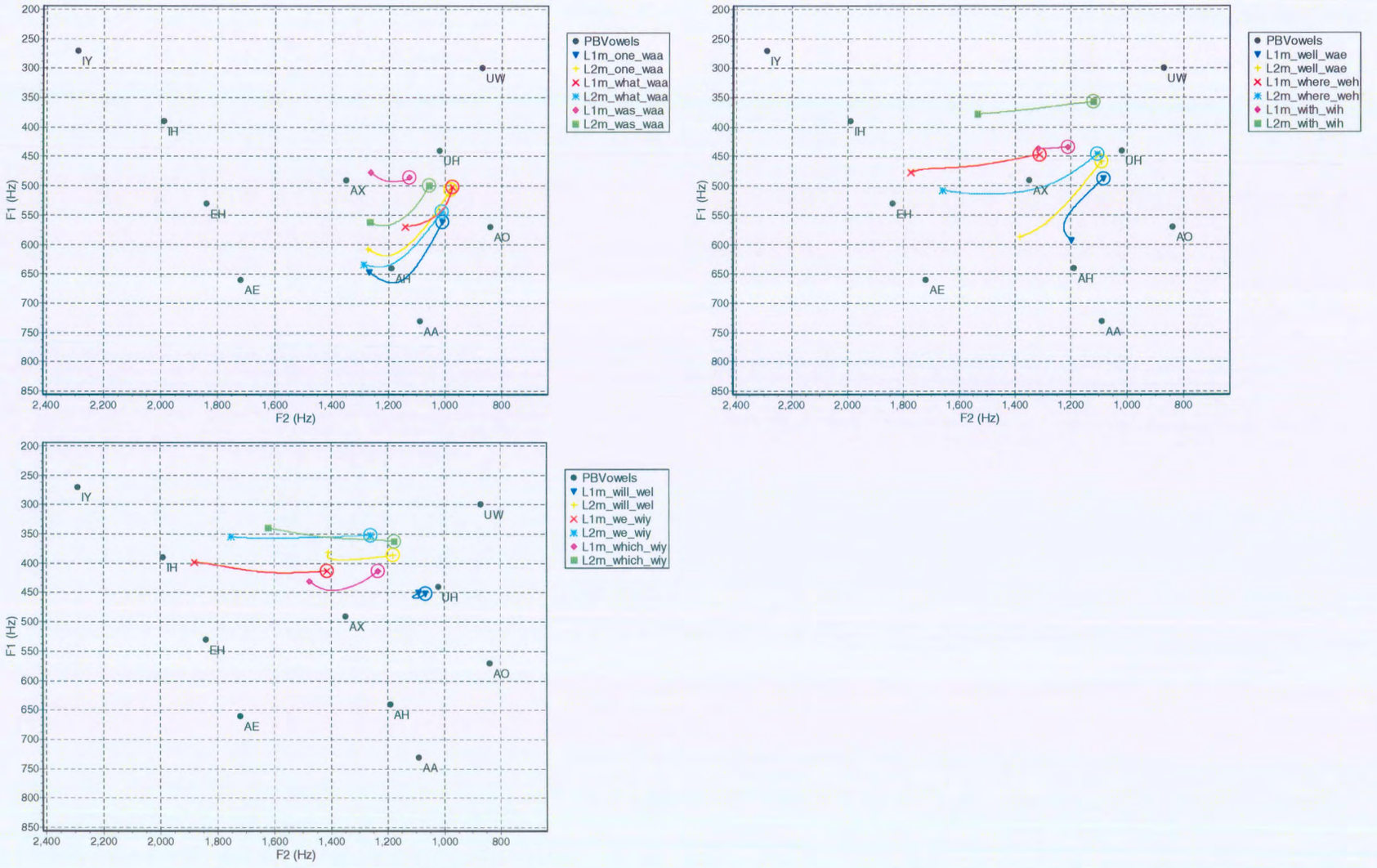


Figure 3.21: Formant results for the diphthongs /ʊa/, /ʊæ/, /ʊɛ/, /ʊə/, /ʊl̩/ and /ʊɪ/ (/WAA/, /WAE/, /WEH/, /WIH/, /WEL/ and /WIY/).

### 3.4.3 Conclusion

It is clear from the results of the previous section that diphthongs differ significantly between first and second language speakers of South African English. Although not present in their native language, L2 speakers seem to pronounce diphthongs authentically in many cases. The main cause for the differences in the remaining cases seem to be due to vowel substitution where an unfamiliar element of the diphthong is either replaced by the closest native vowel, or dropped. In the former case this often leads to over-emphasised diphthongs, such as [wiθ] for *with*.

In the latter case, the diphthong is weakened or monophthongised as in the case of the closing diphthong /əʊ/, where the unfamiliar first element /ə/ is dropped, resulting in [gɔ] or [gʊ] for *go*. The same applies to /aʊ/ in *south*, where we find [sʌθ] as L2 pronunciation. The diphthong /eɪ/ is similarly affected, where L2 pronunciation only retains the second element, therefore [kɛ:p] or [kɪ:p] for *cape*.

All centering diphthongs are also affected, where the second element (/ə/) is replaced with a more peripheral vowel, such as [ðɛa] for *their* or [hɪɛ] for *here*.

In other cases, such as the /aɪ/ in *by*, /ʊa/ in *one*, /ʊɛ/ in *where* and /ʊi/ in *we*, diphthong strength is not markedly affected, although the phonemes may still differ in location.

## 3.5 Mel-scaled cepstral analysis of vowels

Our formant analysis results indicated a number of significant differences between the pronunciation of vowels by first and second language English speakers in South Africa. In order to estimate what effects this may have on automatic speech recognition (ASR) systems, we use the same representation many of these systems employ to model phonemes (i.e. hidden Markov models based on Mel-scaled cepstral features) to determine if significant differences exist between phoneme models which are trained on either first or second language utterances.

In this experiment the vowels of first and second language speakers are compared in the Mel-scaled cepstral domain. The same 59 context dependent vowels as in the formant experiment are examined (Table 3.7, on page 58).

In the next section the specific analysis technique used here is described in more detail, followed by the results. We then give our interpretation of these results and some final concluding remarks.

### 3.5.1 Experimental protocol

As in the formant case, each static vowel is modeled by a normal distribution, this time using 13 Mel-scaled cepstral coefficients (MFCCs) as features. The MFCCs are determined using the following parameters:

- Hamming window size: 10-milliseconds.
- Step size: 5-milliseconds.
- Number of filters in filter-bank: 24 triangular filters, spaced at 132 mels with a bandwidth of 264 mels each.
- Cepstral order: 13 coefficients are used.

The normal distribution is represented by a single-state, single mixture hidden Markov model (HMM), which is trained on the MFCCs alone (no delta coefficients are used as some vowels are too short and included only a single data point). The reason for using such a trivial HMM instead of simply estimating a normal distribution directly, is that this experimental protocol can easily be extended to multi-state HMMs for modelling diphthongs in the next experiment, while using the same pre-processing techniques.

An HMM is therefore trained for each phoneme in each language group (using the utterances available in that group) resulting in a mean and variance pair for each state of each model. The two models are then compared using analysis of variance (ANOVA), similar to the comparison made for the formant distributions. In order to make the HMM comparisons extendable to multi-state HMMs where we do not know which data points (from each individual utterance) map to which HMM state, we do not use the original data points to calculate the variance $F$ ratio, but derive the required information directly from the state variables of the two HMM models with only the following information available:

- The mean ($\mu_j$) and variance ($\sigma_j^2$) of each language group.
- The number of utterances ($n_j$) in each group.

As discussed in Section 2.2.5 on page 37, the ANOVA $F$ ratio can be calculated from

$$\begin{aligned} F &= \frac{\hat{s}_b^2}{\hat{s}_w^2}, \text{ where} \\ \hat{s}_b^2 &= \frac{v_b}{a-1} \text{ and} \\ \hat{s}_w^2 &= \frac{v_w}{N-a}. \end{aligned} \tag{3.4}$$

We therefore require the following information in order to obtain the $F$ ratio:

- The number of data sets ($a$).

- The total number of utterances ($N$).
- The variance *between* data sets ($v_b$).
- The variance *within* data sets ($v_w$).

We know $a = 2$ for this two-class problem and the total number of utterances is the sum of the number of utterances in each group: $N = n_1 + n_2$. From Section 2.2.5 we remember that the variance *between* data sets can be defined as

$$v_b = \sum_{j=1}^{a} n_j(\mu_j - \mu_G)^2, \tag{3.5}$$

where $\mu_G$ is the grand mean of all the data points from all data sets:

$$\mu_G = \frac{1}{N}\sum_{jk} x_{jk}, \tag{3.6}$$

with $x_{jk}$ denoting data point $k$ of data set $j$. Also from Section 2.2.5 we saw that the variance *within* data sets can be defined as

$$v_w = \sum_{jk}(x_{jk} - \mu_j)^2. \tag{3.7}$$

However, from Eqs. (3.6) and (3.7) we see that the individual data points ($x_{jk}$) are indeed required. In order to find the grand mean from the available information, we rewrite Eq. (3.6) as

$$\mu_G = \frac{1}{N}\sum_{j=1}^{a}\sum_{k=1}^{n_j} x_{jk}. \tag{3.8}$$

Since $\mu_j = \frac{1}{n_j}\sum_{k=1}^{n_j} x_{jk}$, we can rewrite Eq. (3.8) as

$$\mu_G = \frac{1}{N}\sum_{j=1}^{a} n_j\mu_j \tag{3.9}$$

in which $N$, $n_j$ and $\mu_j$ are all defined.

With $\mu_G$ known, $v_b$ can be calculated directly from Eq. (3.5) and $v_w$ can be found by manipulating Eq. (3.7) as follows:

$$\begin{aligned} v_w &= \frac{1}{N}\sum_{jk}(x_{jk}-\mu_j)^2 \\ &= \frac{1}{N}\sum_{j=1}^{a}\sum_{k=1}^{n_j}(x_{jk}-\mu_j)^2 \\ &= \frac{1}{N}\sum_{j=1}^{a} n_j\sigma_j^2, \end{aligned} \tag{3.10}$$

where $\sigma_j^2 = \frac{1}{n_j}\sum_{k=1}^{n_j}(x_{jk}-\mu_j)^2$.

With both $v_b$ and $v_w$ known, $F$ can now be calculated to determine if this state of the two HMM models differ significantly. In the case of multi-state HMMs this is simply repeated for each state.

The results we obtained from using this method are discussed in more detail in the following section.

### 3.5.2 Results and discussion

This section details the results found using the experimental method described above to compare the vowels of first and second language South African English in the Mel-scaled cepstral domain. Table 3.13 summarises the analysis of variance results we obtained.

The first column of Table 3.13 indicates the vowel, the second shows the context dependent phoneme, while the third column contains the degrees of freedom (DOF) of the ANOVA test. The ANOVA results for each Mel-scaled cepstral coefficient (MFCC)

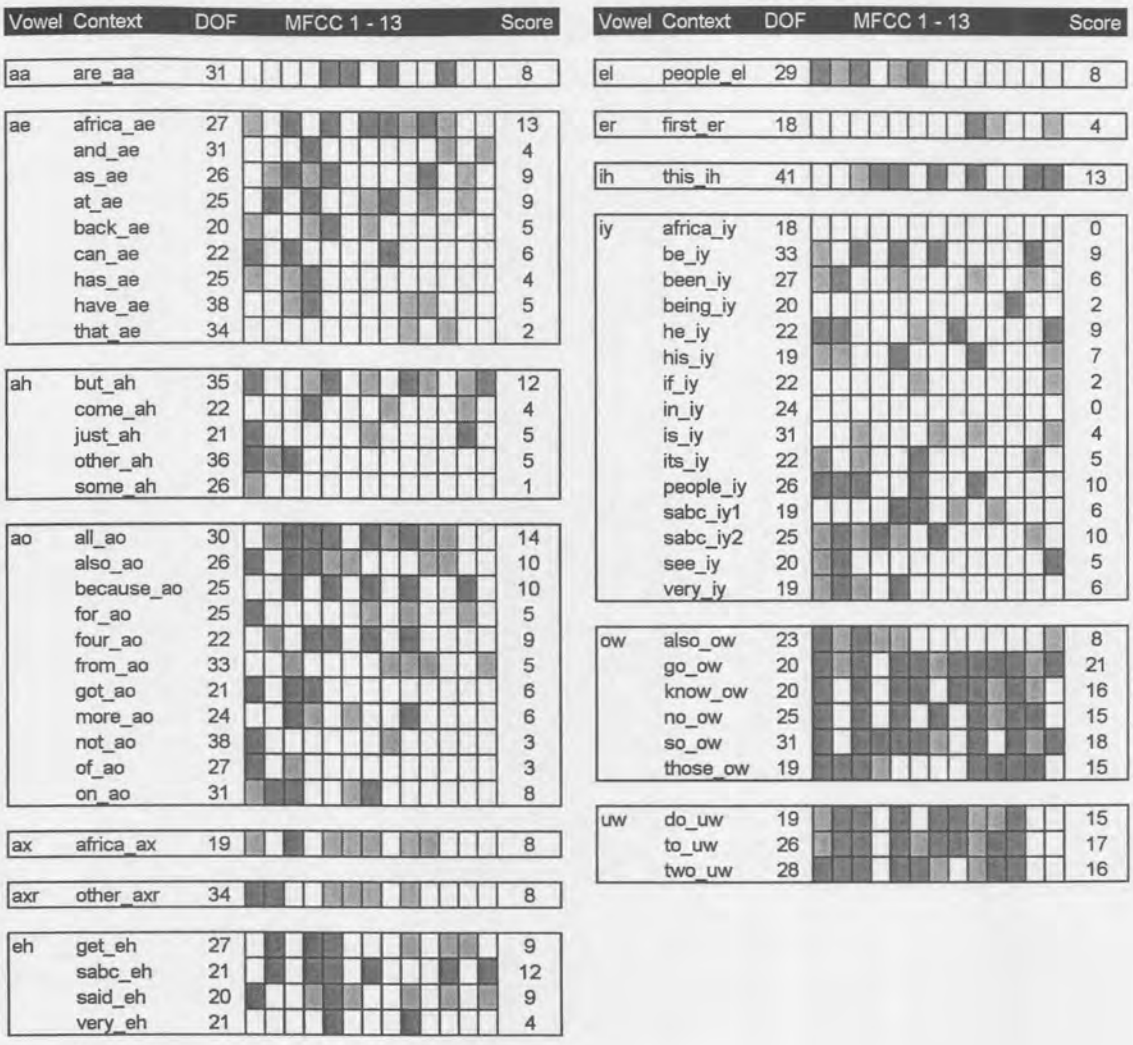


Table 3.13: Analysis of variance results for Mel-scaled cepstral vowel comparisons.

dimension is indicated in the next 13 columns, where a white block indicates no significant difference, light-grey indicates a significant difference with a certainty level of 95%, while dark-grey indicates a significant difference to 99% certainty. The final column contains a relative difference score, which was calculated as follows: For each 95% certain significant difference the score is incremented by one, for each 99% certain significant difference it is incremented by two. This is then summed across all 13 MFCCs to obtain a relative indication of how much the vowels of the two language groups differ with 0 indicating no significant difference in any cepstral dimension and 26 indicating significant differences to a 99% certainty for all the MFCC dimensions.

The results are discussed in more detail below, with reference to Table 3.13.

### /ɑ/ (/AA/)

For the vowel /ɑ/ (/AA/) a difference score of 8 was reached with four cepstral dimensions differing significantly to a certainty level of 99% between the L1 and L2 HMM models. It seems that the differences we saw in the formant results translate to an appreciable effect on the cepstral features.

### /æ/ (/AE/)

In this case the word context seems to play an important role. We can categorise the contexts into three broad categories: We see a fairly small difference between the models of *and*, *has* and *that* where the score was from 2 to 4. In other cases a significant, but not drastic, change is present with scores of 5 or 6 for *can*, *back* and *have*. A drastic change with a score of 13 is again seen for the /æ/ in *africa* as in the formant case, while *as* and *at* also showed significant change with a score of 9. Good correlation is found with the formant results (Table 3.10 on page 66), where *has* and *that* were also the least affected, while *as* and *africa* showed the most significant change.

### /ʌ/ (/AH/)

The general results for this vowel shows a significant, but not drastic change with scores of 4 or 5 for *come*, *just* and *other*. However, at the two extremes we find a definite difference for /ʌ/ in the context of *but*, while the pronunciation of *some* shows little change. The cepstral results therefore seem more sensitive to the changes of L2 /ʌ/ than in the formant case.

### /ɔ/ (/AO/)

Significant differences are seen in many cases, such as *all*, *also*, *because*, *four* and *on*, while in the other contexts like *from*, *not* and *of* very little change is seen. As was the case in the formant results, the long /ɔ:/ seems to behave differently from the short /ɔ/, where a more prominent difference between first and second language pronunciation exists in the former.

### /ə/, /ɚ/, /ɝ/ and /l̩/ (/AX/, /AXR/, /ER/ and /EL/)

In the cases of /ə/ in *africa*, /ɚ/ in *other* and /l̩/ in *people* a fair difference is seen with a score of 8. The /ɝ/ in *first* is the outlier in this case, showing less change with a score of 4. These results seem to agree with the formant results (Table 3.10) where a 95% certain significant change was found for one formant in most cases.

### /ɛ/ (/EH/)

For most contexts of the vowel /ɛ/, definite differences are seen, with *sabc* reaching the largest score of 12, followed by *get* and *said* at 9. The context of *very* shows less evidence of movement with a score of only 4. A surprise is the difference seen for *said*, since the formant results did not reach significance (Table 3.10), although some movement was still evident.

### /ɪ/ (/IH/)

The /ɪ/ in the word *this* showed definitive change with a score of 13 and significant changes to a 99% certainty in 6 cepstral dimensions. This agrees well with the formant results, where we found a significant change in F2.

### /i/ (/IY/)

The many contexts of /i/ shows a range of different results. Large differences are seen for *people*, the second /i/ in the acronym *SABC* [ɛseɪbisi:], *be* and *he* with scores of 9 or 10. More moderate differences are seen for *his*, *been*, the first /i/ in *SABC*, *very*,*its*, *see* and *is* with scores ranging from 4 to 7. In the cases of *being*, *if*, *africa* and *in* we see much less change with scores of 0 or 2. In general we see the same distinction between the long and short forms of the vowel /i/ as in the formant case, where the long /i:/ is more affected in L2 speech than the shorter /i/. Some exceptions to this generalisation are notably the long /i:/ in *being* which only scored 2, and the short /i/ in *his* which scored 7.

### /o/ (/OW/)

In the case of /o/ we find drastic changes in all contexts with scores from 15 to 21. The only exception (as in the formant case) is the /o/ in *also* which showed less difference with a score of 8. The different locations for the L1 and L2 phoneme /o/ we saw in the formant plots (Figure 3.14 on page 80) translates to significant differences in almost all of the cepstral dimensions.

### /u/ (/UW/)

The vowel /u/ showed a definite change between L1 and L2 HMM models with significant changes of 99% certainty in most cepstral dimensions and scores from 15 to 17. This agrees with the dramatic difference in L1 and L2 phoneme locations we saw in the formant plots on page 80 where L1 speakers use a much more central (forward) vowel.

This section presented the analysis of variance results for the vowels in detail. A

higher-level discussion and summary are given below.

### 3.5.3 Conclusion

In general most of the L1 and L2 HMM phoneme models differed significantly in a number of the Mel-scaled cepstral dimensions, which certainly implies that automatic speech recognition (ASR) systems based on British or first language South African English (SAE) would be adversely affected when exposed to second language SAE.

Although no direct relation seems exists for the degree of difference found here for each individual context dependent phoneme compared to that of the formant experiment, the more drastic trends do persist in the cepstral domain. For instance we find that the drastic changes in pronunciation of the vowels /ɪ/, /o/ and /u/ (/IH/, /OW/ and /UW/) also significantly affect the cepstral features, while the similarity between L1 and L2 /ʌ/ seen in the formant domain is also evident here in cepstral space.

In the next experiment the same comparison drawn here for the vowels is repeated for diphthongs.

## 3.6 Mel-scaled cepstral analysis of diphthongs

Our formant analysis results indicated a number of significant differences between the pronunciation of diphthongs by first and second language English speakers in South Africa. In order to estimate if this may have any significant impact on automatic speech recognition (ASR) systems, we again use hidden Markov models (HMMs) based on Mel-scaled cepstral features to model the phonemes as in the previous (vowel) experiment.

In this experiment the diphthongs of first and second language speakers are compared in the Mel-scaled cepstral domain. The same 27 context dependent diphthongs as in

the formant experiment are examined (Table 3.11, on page 83).

In the next section the specific analysis technique used here is described in more detail, followed by the results. We then give our interpretation of these results and some final concluding remarks.

### 3.6.1 Experimental protocol

As in the formant experiment, each diphthong is modelled as three sections. Where spline curves were used previously, we use a three-state left-to-right hidden Markov model in this case. No state skips are allowed, therefore forcing the HMM to traverse all three states. Each HMM state models that section of the diphthong as a single 13-dimensional normal distribution of the Mel-scaled cepstral coefficients (MFCCs). These coefficients are determined with the same parameters as in the previous experiment:

- Hamming window size: 10-milliseconds.
- Step size: 5-milliseconds.
- Number of filters in filter-bank: 24 triangular filters, spaced at 132 mels with a bandwidth of 264 mels each.
- Cepstral order: 13 coefficients are used.

The technique described previously in Section 3.5.1 is again used to perform the ANOVA test on each state of these HMMs directly from the model parameters ($\mu$ and $\sigma^2$). For a pair HMMs, the one modelling a first language phoneme and the other its second language counterpart, each state of the former is compared to its equivalent in the latter, resulting in a significance score for the difference between each state of the two HMMs. These results therefore indicate the difference between each of three sections of the L1 and L2 diphthong, much as in the formant analysis case. We discuss our findings in detail below.

### 3.6.2 Results and discussion

This section details the results found using the experimental method described above to compare the diphthongs of first and second language South African English in the Mel-scaled cepstral domain. Table 3.14 summarises the analysis of variance results we obtained.

| Diphthong | Context | DOF | 0.05 | | | 0.01 | | | Combined | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Section → | | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | |
| aw | now_aw | 26 | 6 | 5 | 5 | 3 | 3 | 3 | 9 | 8 | 8 | 25 |
| | south_aw | 30 | 6 | 7 | 3 | 4 | 5 | 2 | 10 | 12 | 5 | 27 |
| awr | our_awr | 24 | 6 | 6 | 6 | 5 | 5 | 3 | 11 | 11 | 9 | 31 |
| ay | by_ay | 29 | 6 | 8 | 6 | 5 | 5 | 3 | 11 | 13 | 9 | 33 |
| | time_ay | 22 | 4 | 6 | 6 | 3 | 3 | 4 | 7 | 9 | 10 | 26 |
| ehaxr | their_ehaxr | 31 | 4 | 9 | 7 | 4 | 7 | 6 | 8 | 16 | 13 | 37 |
| | there_ehaxr | 33 | 6 | 3 | 4 | 3 | 1 | 1 | 9 | 4 | 5 | 18 |
| ey | cape_ey | 20 | 6 | 7 | 2 | 4 | 5 | 0 | 10 | 12 | 2 | 24 |
| | sabc_ey | 20 | 4 | 6 | 2 | 2 | 6 | 2 | 6 | 12 | 4 | 22 |
| | they_ey | 25 | 8 | 11 | 8 | 5 | 8 | 8 | 13 | 19 | 16 | 48 |
| iyaxr | here_iyaxr | 23 | 1 | 6 | 4 | 1 | 2 | 2 | 2 | 8 | 6 | 16 |
| ow | also_ow | 23 | 6 | 9 | 6 | 2 | 4 | 3 | 8 | 13 | 9 | 30 |
| | go_ow | 20 | 10 | 8 | 9 | 10 | 7 | 6 | 20 | 15 | 15 | 50 |
| | know_ow | 20 | 11 | 9 | 8 | 7 | 9 | 4 | 18 | 18 | 12 | 48 |
| | no_ow | 25 | 7 | 8 | 7 | 6 | 5 | 7 | 13 | 13 | 14 | 40 |
| | so_ow | 31 | 10 | 10 | 8 | 9 | 9 | 8 | 19 | 19 | 16 | 54 |
| | those_ow | 19 | 7 | 9 | 11 | 6 | 7 | 8 | 13 | 16 | 19 | 48 |
| waa | one_waa | 41 | 4 | 3 | 1 | 3 | 3 | 1 | 7 | 6 | 2 | 15 |
| | was_waa | 27 | 3 | 5 | 6 | 0 | 3 | 4 | 3 | 8 | 10 | 21 |
| | what_waa | 19 | 0 | 4 | 3 | 0 | 2 | 3 | 0 | 6 | 6 | 12 |
| wae | well_wae | 21 | 2 | 2 | 5 | 1 | 0 | 3 | 3 | 2 | 8 | 13 |
| weh | where_weh | 25 | 6 | 4 | 3 | 2 | 2 | 2 | 8 | 6 | 5 | 19 |
| wel | will_wel | 27 | 3 | 4 | 9 | 2 | 3 | 7 | 5 | 7 | 16 | 28 |
| wih | with_wih | 25 | 4 | 5 | 8 | 3 | 4 | 4 | 7 | 9 | 12 | 28 |
| wiy | we_wiy | 32 | 4 | 3 | 6 | 4 | 2 | 4 | 8 | 5 | 10 | 23 |
| | which_wiy | 24 | 5 | 7 | 6 | 4 | 4 | 3 | 9 | 11 | 9 | 29 |
| yuh | you_yuh | 21 | 2 | 7 | 4 | 1 | 5 | 1 | 3 | 12 | 5 | 20 |

Table 3.14: Analysis of variance results for Mel-scaled cepstral diphthong comparisons.

The first three columns indicate the diphthong, context-dependent phoneme and degrees of freedom of the ANOVA test respectively. The next three columns show, for

each of the three diphthong sections, the number of cepstral dimensions for which significant differences with a certainty level of 95% was found. This is repeated for the 99% certainty level. The next three columns contains a combined score for each section where the 95% and 99% scores are summed. This indicates a relative score of difference for each segment with 0 indicating no significant difference and 26 significant differences in all 13 cepstral dimensions to a 99% level of certainty. Two arbitrary levels were chosen to aid in visualising these results: a score below 8 is indicated in a white block (small difference), from 9 to 14 as light-grey (medium difference) and 15 to 26 as dark-grey (large difference). The final column of Table 3.14 contains a difference score for the diphthong as a whole, which is simply the sum of the individual scores of the three sections, ranging from 0 (no significant difference) to 78 (all three sections differ significantly with a 99% certainty in all 13 cepstral dimensions). We examine these results in more detail for each individual diphthong below.

### /aʊ/ and /aʊ˞/ (/AW/ and /AWR/)

For the diphthong /aʊ/ in both *now* and *south* we see significant differences in many of the cepstral dimensions for all three sections. In the context of *now* we see analysis of variance scores of 8 or 9 in all three diphthong sections with three cepstral dimensions reaching significant difference to a level of 99% certainty. In the case of *south* also see high ANOVA scores for the first two diphthong sections, with less marked difference in the final section with a score of only 5. In the case of /aʊ˞/ in *our*, we also see a marked level of change, with ANOVA scores of 9 or 11 in all three sections and a total score of 31. These results seem to agree with the formant results indicated in Table 3.12 on page 88, where we also see significant differences for these phonemes.

### /aɪ/ (/AY/)

In the case of the diphthong /aɪ/, we find much the same level of difference as for /aʊ/, where we have a total ANOVA score of 33 and 26 for *by* and *time* respectively. Significant differences are present in 6 or more cepstral dimensions in most diphthong sections, although the first section of the /aɪ/ in *time* only differed significantly in 4 cepstral coefficients. These results closely match those of the formant case, where for *time* we also see a smaller score than for *by*, although the two language groups differ significantly in both instances.

### /ɛɚ/ (/EH-AXR/)

As in the formant case, the results for this phoneme are split according to context. In the case of *their* we see a marked difference where the ANOVA scores for the three sections range from 8 to 16 and a total score of 37. For *there* we see much less difference with a total score of 18, and only a single cepstral coefficient reaching the 99% certainty level in the second and third diphthong sections.

### /eɪ/ (/EY/)

For /eɪ/ in *cape* we see marked differences in the first and second diphthong sections, while for *they* all three sections show significant differences in at least 8 of the 13 cepstral dimensions. In the case of /eɪ/ in the acronym *SABC* a less pronounced difference is found.

### /ɪɚ/ (/IY-AXR/)

The analysis of variance results for /ɪɚ/ in *here* indicates very little change in the first diphthong section, with a score of 2, while the second and third sections show more

definite with scores of 8 and 6 respectively. This would indicate that while the L1 and L2 pronunciations of /ɪɚ/ start similarly, they deviate in the middle and final sections.

### /əʊ/ (/OW/)

Marked changes are visible in all contexts of /əʊ/ with the ANOVA scores of 13 or higher for most diphthong sections. The only exception seems to be *also*, where the highest score was 13 for the second section. However, total scores of 30 to 54 are reached in all cases, indicating a definite difference between the L1 and L2 HMMs.

### /ʊa/ (/WAA/)

The models for /ʊa/ in *one* and *what* seem very similar with ANOVA scores of 6 or less for all diphthong sections. In the case of *was* we see a more pronounced difference with scores of 8 and 10 in the second and third sections, indicating that diphthongs of the two language groups start similar and then diverge. This agrees well with the formant results, where we saw a weakened L1 diphthong with the second element of /ʊa/ being centralised by L1 speakers.

### /ʊæ/ (/WAE/)

In the case of /ʊæ/ in *well*, we see little difference, except for the final diphthong section where an ANOVA score of 8 was reached. This agrees with the formant ANOVA results (Table 3.12) where a significant change was also only found for the third section of F2.

### /ʊɛ/ (/WEH/)

For this diphthong in the context of *where* we see progressively less difference for the three sections with scores of 8, 6 and 5 respectively. It would therefore seem that the L1 and L2 phonemes start out differently and converge later on. These results agree well with the formant case, where we also see a significant difference in the first diphthong section of all three formants and the formant plots on page 101 also seem to indicate this convergence of the two phonemes.

### /ʊl̩/ (/WEL/)

Our results for /ʊl̩/ in *will* start with little significant difference in the first diphthong section with a score of 5, and proceeds to indicate larger differences in the following two sections with scores of 7 and 16 respectively. This would certainly indicate a divergence in pronunciation of the phoneme for the two language groups. If we compare these results with those obtained in formant space, we can clearly see the correlation, where we find significant differences for the second and third sections in all formant dimensions, while not for the first section. The formant plots show that the L1 phoneme is pronounced as a vowel, with the L2 version originating in the same region and then moving away, causing the progressively larger differences in the last to sections.

### /ʊə/ (/WIH/)

In the case of /ʊə/ in *with*, we also see progressively larger differences in the three diphthong sections of 7, 9 and 12 respectively. This supports the diphthongisation we see in formant space.

### /ʊɪ/ (/WIY/)

For /ʊɪ/ in *we* and *which* we find significant differences in most diphthong sections, although in the case of *we* the second section only reached a score of 5 with one cepstral coefficient showing significant difference to a certainty of 95% and two to a certainty of 99%. Although these results show the same general degree of difference as in the formant case, *we* showed more significant change in that instance.

### /ju/ (/YUH/)

Our results show only marginal difference for the diphthong /ju/ in *you*, with scores of 3, 12 and 5 for the three diphthong sections respectively.

These detailed results are summarised in the next section to give a higher-level overview of our findings.

## 3.6.3 Conclusion

A definitive parallel can be drawn between the Mel-scaled cepstral results and those obtained in formant space. It seems the changes seen in formant space can be directly translated to the Mel-scaled cepstral domain in most cases. For instance we again see the drastic difference between the L1 and L2 models of /əʊ/ as in *so* and /eɪ/ in *they* and the similarities for /ʊa/ as in *one* and /ɛɚ/ in *there*.

It seems that as we found for the vowels, automatic speech recognition (ASR) systems based on Mel-scaled cepstral features, will certainly be affected by the differences in L1 and L2 pronunciation of English in South Africa. Due to the similarities between the formant and cepstral space results, it may be possible to determine from acoustic and formant analysis which cepstral phoneme models need to be adapted in ASR systems

for better performance on L2 SAE.

The adaptation of existing SAE speech recognition systems could be the subject of further study, where alternative (L2) pronunciations of the significantly affected words are defined. New L2 phoneme models could be trained separately from the second language data, wherever these phonemes differ significantly from their L1 counterparts.

## 3.7 Diphthong- and Monophthongisation

In this experiment we develop a method to estimate the strength of a diphthong and then compare this metric between first and second language versions of the same diphthong. A diphthong's strength can be defined as the amount of change in vowel quality the phoneme undergoes as it is pronounced.

If the strength of an L2 diphthong is *less* that its L1 counterpart, it indicates that the diphthong is *monophthongised* (pronounced more like a vowel) by second language speakers. In the opposite case, where the L2 diphthong has a larger strength metric, *diphthongisation* is indicated, where the L2 version is spoken as a more pronounced diphthong.

The next section describes the specific analysis technique used here, and is followed by the results. We then give our interpretation of these results and some concluding remarks.

### 3.7.1 Experimental protocol

In this experiment we re-use the three-state hidden Markov models for the diphthongs obtained in the previous experiment and apply the analysis of variance (ANOVA) test in a different way to estimate the difference in diphthong strength between L1 and L2

speakers. In this case we do not use the ANOVA test to compare each L1 diphthong section with its L2 counterpart, but we rather estimate the overall strength of each diphthong separately. For each language group, an ANOVA test is performed on all three states of the HMM viewing the three state distributions as those of a three-class problem. A pronounced diphthong is then indicated by the three states differing significantly in many of the cepstral dimensions. The number of cepstral dimensions in which significant differences are found is therefore used as a relative measure of diphthong strength.

The technique to calculate the analysis of variance $F$ ratio from the HMM model parameters given in Section 3.5.1 can be extended for a three-class problem as follows:

The ANOVA $F$ ratio is again defined as

$$\begin{aligned} F &= \frac{\hat{s}_b^2}{\hat{s}_w^2}, \text{ where} \\ \hat{s}_b^2 &= \frac{v_b}{a-1} \text{ and} \\ \hat{s}_w^2 &= \frac{v_w}{N-a}. \end{aligned} \tag{3.11}$$

In this case $a = 3$ and the total number data elements in all data sets ($N$) is the sum of the number of utterances in each set. Since the three sets are derived from the three states of a single HMM, each set contains the same number of elements: $N = \sum_{j=1}^{a} n = a \cdot n$.

The grand mean of all data points ($\mu_G$) can still be found from

$$\mu_G = \frac{1}{N} \sum_{j=1}^{a} n_j \mu_j, \tag{3.12}$$

which simplifies for our case where $n_j = n$ and $N = a \cdot n$ as

$$\mu_G = \frac{1}{a} \sum_{j=1}^{a} \mu_j. \tag{3.13}$$

With $\mu_G$ known, $v_b$ can be calculated as before (Eq. (3.5) on page 105) and $v_w$ can be found by simplifying (Eq. (3.10) on page 106):

$$\begin{aligned} v_w &= \frac{1}{N}\sum_{j=1}^{a} n_j\sigma_j^2 \\ &= \frac{1}{a}\sum_{j=1}^{a}\sigma_j^2, \text{ since } n_j = n \text{ and } N = a \cdot n. \end{aligned} \tag{3.14}$$

With $v_b$ and $v_w$ known, $F$ can now be calculated as before to determine if the three states of the HMM model differ significantly, which would indicate a pronounced diphthong.

### 3.7.2 Results and discussion

In this section we detail the analysis of variance results obtained from the three-class comparisons made using the technique described above. The results are shown in Table 3.15.

The first column indicates the diphthong, while the second shows the context-dependent phoneme. This is followed by four columns indicating the following information for the first language phoneme (L1): The degrees of freedom for the ANOVA test (DOF), the number of cepstral dimensions for which a 95% certain significant differences was found (0.05), the same for the 99% certainty level (0.01) and finally the sum of these two values ($\Sigma$). This sum is used as the metric to indicate diphthong strength, where 0 indicates a vowel and 26 a very strong diphthong. These four columns are repeated for the second language phoneme (L2). In the final column (Ratio) we indicate the ratio of the L2 $\Sigma$ divided by the L1 $\Sigma$, which estimates the difference in diphthong strength between the two language groups. Diphthongisation by second language speakers is therefore indicated by a value larger than 1.00, 1.00 indicates equal diphthong strengths, while a value less than 1.00 indicates monophthongisation. Using arbitrarily chosen levels of 0.5 and 2.00, diphthongisation ($\geq$ 2.00) is indicated as a dark-grey block, while monophthongisation ($\leq$ 0.5) is indicated in light-grey. These results are discussed with

| Diphthong | Context | L 1 | | | | L 2 | | | | Ratio |
|---|---|---|---|---|---|---|---|---|---|---|
| | | DOF | 0.05 | 0.01 | Σ | DOF | 0.05 | 0.01 | Σ | |
| aw | now_aw | 45 | 6 | 3 | 9 | 33 | 4 | 2 | 6 | 0.67 |
| | south_aw | 42 | 6 | 5 | 11 | 48 | 2 | 0 | 2 | 0.18 |
| awr | our_awr | 30 | 3 | 3 | 6 | 42 | 3 | 2 | 5 | 0.83 |
| ay | by_ay | 42 | 7 | 5 | 12 | 45 | 7 | 4 | 11 | 0.92 |
| | time_ay | 30 | 3 | 1 | 4 | 36 | 1 | 0 | 1 | 0.25 |
| ehaxr | their_ehaxr | 45 | 3 | 1 | 4 | 48 | 4 | 4 | 8 | 2.00 |
| | there_ehaxr | 48 | 3 | 1 | 4 | 51 | 4 | 1 | 5 | 1.25 |
| ey | cape_ey | 33 | 5 | 3 | 8 | 27 | 1 | 0 | 1 | 0.13 |
| | sabc_ey | 27 | 4 | 3 | 7 | 33 | 2 | 1 | 3 | 0.43 |
| | they_ey | 45 | 6 | 6 | 12 | 30 | 3 | 0 | 3 | 0.25 |
| iyaxr | here_iyaxr | 42 | 5 | 4 | 9 | 27 | 3 | 2 | 5 | 0.56 |
| ow | also_ow | 30 | 2 | 1 | 3 | 39 | 1 | 0 | 1 | 0.33 |
| | go_ow | 27 | 6 | 0 | 6 | 33 | 0 | 0 | 0 | 0.00 |
| | know_ow | 33 | 1 | 0 | 1 | 27 | 2 | 0 | 2 | 2.00 |
| | no_ow | 27 | 3 | 2 | 5 | 48 | 7 | 4 | 11 | 2.20 |
| | so_ow | 42 | 3 | 3 | 6 | 51 | 4 | 3 | 7 | 1.17 |
| | those_ow | 30 | 1 | 0 | 1 | 27 | 2 | 0 | 2 | 2.00 |
| waa | one_waa | 57 | 10 | 7 | 17 | 66 | 10 | 8 | 18 | 1.06 |
| | was_waa | 33 | 7 | 6 | 13 | 48 | 6 | 4 | 10 | 0.77 |
| | what_waa | 27 | 5 | 4 | 9 | 33 | 8 | 6 | 14 | 1.56 |
| wae | well_wae | 36 | 4 | 4 | 8 | 27 | 6 | 2 | 8 | 1.00 |
| weh | where_weh | 45 | 9 | 8 | 17 | 30 | 5 | 4 | 9 | 0.53 |
| wel | will_wel | 39 | 2 | 1 | 3 | 42 | 4 | 4 | 8 | 2.67 |
| wih | with_wih | 33 | 5 | 3 | 8 | 42 | 6 | 4 | 10 | 1.25 |
| wiy | we_wiy | 48 | 7 | 6 | 13 | 48 | 9 | 8 | 17 | 1.31 |
| | which_wiy | 27 | 7 | 3 | 10 | 45 | 7 | 7 | 14 | 1.40 |
| yuh | you_yuh | 30 | 7 | 4 | 11 | 33 | 4 | 2 | 6 | 0.55 |

Table 3.15: Analysis of variance results indicating diphthongisation / monophthongisation in the Mel-scaled cepstral domain.

respect to each diphthong below.

### /aʊ/ and /aʊ˜/ (/AW/ and /AWR/)

The /aʊ/ in *south* shows definite monophthongisation with a diphthong strength ratio of 0.18. The L1 diphthong shows an ANOVA score for its three HMM states of 11, while the L2 diphthong strength score is only 2. In the context of *now* we see a less

pronounced effect with an L1 diphthong strength of 9 and L2 of 6.

In the case of /aʊɹ/ in *our*, we see no significant difference in the diphthong strength, with L1 and L2 diphthong strength scores of 6 and 5 respectively.

### /aɪ/ (/AY/)

For /aɪ/ in *by* we see no major difference in diphthong strengths between language groups. For both L1 and L2 speakers we see high diphthong strengths of 12 and 11 respectively. In the case of *time* monophthongisation is evident (a ratio of 0.25) while both L1 and L2 diphthong strength scores are much lower at 4 and 1 respectively, indicating a less pronounced diphthong. These results correlate well with the formant results, where we can see from Figure 3.19 on page 99 that *by* is pronounced as a stronger diphthong than *time*, although the strong monophthongisation of *time* is not evident in F1-F2 space.

### /ɛəɹ/ (/EH-AXR/)

Our results for /ɛəɹ/ in *their* show definitive diphthongisation (ratio = 2.00), with the L2 diphthong strength score at 8, compared to the L1 score of 4. In the context of *there* little difference is seen, with L1 and L2 scores of 4 and 5 respectively. This indicates similar diphthong strengths for L1 *their*, and both versions of *there*, while L2 *their* is much more pronounced. These results agree well with the formant results where we can clearly see the strong L2 diphthong for *their* (+) in Figure 3.20 on page 100.

### /eɪ/ (/EY/)

In all contexts of /eɪ/, we see definitive monophthongisation with diphthong strength ratios of 0.13, 0.43 and 0.25 for *cape*, *SABC* and *they* respectively. This is also visible

in the formant plots of Figure 3.19 on page 99, where the L2 diphthongs of *cape* and *they* (+ and ■) are much more peripheral and weakened.

### /ɪɚ/ (/IY-AXR/)

For /ɪɚ/ in *here* we see some monophthongisation with a ratio of 0.56 as the L1 diphthong strength reached a score of 9, compared to the L2 score of 5. This is also visible from Figure 3.19, where the L2 diphthong (■) travel is less.

### /əʊ/ (/OW/)

In the case of *so* no marked difference in L1 and L2 diphthong strength is seen with a ratio of 1.17, although *so* is a pronounced diphthong in both groups. Monophthongisation is present for *also* with a difference ratio of 0.33, while the /əʊ/ in *go* shows much more evidence of this, with a ratio of 0.00 (i.e. the L2 version is a vowel). In the cases of *know*, *no* and *those* the results indicate diphthongisation with L2/L1 ratios of 2.00, 2.20 and 2.00 respectively.

These results correlate well with the formant results (Figure 3.20), where the monophthongisation of *go* and *also* can be seen, although the apparent diphthongisations are not clearly visible in these F1-F2 plots.

### /ʊa/ (/WAA/)

Our results for /ʊa/ show little difference in the L1 versus L2 diphthong strengths with ratios of 1.06 and 0.77 for *one* and *was* respectively. In the case of *one* both L1 and L2 speakers use a pronounced diphthong with diphthong strengths of 17 and 18 respectively. In case of *what* the diphthong strength ratio of 1.56 does indicate some diphthongisation. These results agree well with the formant plots on page 101, where

we can see the strong diphthong used in the context of *one* and the more pronounced L2 version of *what* (+).

### /ʊæ/ (/WAE/)

No difference in diphthong strength is indicated for the /ʊæ/ in *well* with a ratio of 1.00, although both L1 and L2 speakers use a pronounced diphthong of strength 8. This can also be seen in the formant plots on page 101.

### /ʊɛ/ (/WEH/)

Although some monophthongisation is indicated by a diphthong strength ratio of 0.53 for the /ʊɛ/ in *where*, this is not evident from the F1-F2 plots on page 101.

### /ʊl̩/ (/WEL/)

Strong diphthongisation of the /ʊl̩/ in *will* is evident from a diphthong strength ratio of 2.67 where the L2 diphthong has a strength score of 8 compared the 3 of the L1 phoneme. This can clearly be seen in the formant plots of Figure 3.21 on page 101.

### /ʊə/ (/WIH/)

Both L1 and L2 models of /ʊə/ in *with* show large diphthong strength values of 8 and 10 respectively. This slight diphthongisation seems to be much more apparent in formant space if we refer to the graphs on page 101.

### /ʊɪ/ (/WIY/)

The models for /ʊɪ/ in *we* and *which* show marked diphthongs are used for this phoneme in both language groups with scores ranging from 10 to 17. This can also be seen in the formant plots on page 101, which also shows some diphthongisation in the case of *which* supported here with a diphthong strength ratio of 1.4.

### /ju/ (/YUH/)

Monophthongisation of /ju/ in *you* is indicated by a diphthong strength ratio of 0.55, where the L2 model reached a diphthong strength score of 6, compared to the L1 score of 11. This can also be seen in Figure 3.19 on page 99, where the L2 phoneme (■) shows much less travel.

The detailed results given here are summarised in the following section, giving a higher-level overview of our findings.

## 3.7.3 Conclusion

The use of a three-class ANOVA test to determine how significant the three states of a diphthong's HMM differs from one another seems to be a good indicator of diphthong strength. By comparing these metrics between the L1 and L2 models, a measure of diphthongisation can be found. This seems to have performed well as our results agree in most cases with the formant plots of the formant diphthong analysis experiment on page 82.

These results support the findings of the formant space experiment, where we see monphthongisation of the closing diphthongs such as /aʊ/ in *south*, /eɪ/ as in *cape* and the /əʊ/ in *go*. Diphthongisation is indicated for the centering diphthong /ɛɚ/ in *their*

and /ʊl̩/ in *will*, while little change is apparent for the /ɛɚ/ in the context of *there*, /aɪ/ in *by*, /ʊa/ as in *one* and /ʊæ/ in *well*.

# Chapter 4

# Summary and conclusion

This dissertation has presented the goals, background theory, technique and results on an acoustic analysis of the vowels and diphthongs of first and second language South African English (SAE). The second language speakers of interest were speakers of the native African languages found in Southern Africa, such as isi-Zulu, isi-Xhosa, Southern Sotho and Tswana.

Justification for this study can be found from both the speech processing and linguistic points of view:

- More detailed studies on the current state of second language South African English are required to lead the development and teaching of English in South Africa.
- Further study and acoustic analysis of the African languages for the development of automatic speech recognition (ASR) systems for these languages.
- In-depth knowledge of the phonetic differences between first and second language South African English can be used in the development of more robust English ASR systems in the South African context.

## 4.1 Summary of results

A large database of South African English speech was recorded containing both first (L1) and second language (L2) speakers. A total of 2214 usable phoneme utterances were extracted from the raw data, resulting in 80 unique context-dependent phonemes, which were acoustically compared between L1 and L2 speakers. This database is included on the accompanying media described in Appendix C and can be used in the training and further study of L2 SAE speech recognition systems.

Five experiments were conducted:

1. The L1 and L2 vowels were compared using formant analysis, thereby directly comparing phonemes from one language group to that of the other in a measurement space closely aligned with the linguistic representation, viz. two-dimensional formant-plots compare well with phonetic vowel charts.

2. Diphthongs were compared using the same formant approach, augmented by cubic splines used to model their dynamic nature.

3. The Mel-frequency scaled cepstral coefficients (MFCCs) of the vowels were used to train hidden Markov models (HMMs) for each context-dependent phoneme in each language group. The parameters of each pair of HMMs were then compared.

4. The same Mel-scaled cepstral analysis was repeated for the diphthongs, using multi-state HMMs.

5. Finally, diphthongisation of vowels and vice-versa (monophthongisation of diphthongs) were examined using a metric derived from the HMM state distributions.

Comparisons were performed using analysis of variance (ANOVA) and these results were discussed in detail in Sections 3.3 through 3.7. It was found that second language South African English differs significantly from the first language norm, both in the

formant and Mel-scaled cepstral feature spaces. The results give a detailed model of the differences between a number of context-dependent L1 and L2 phonemes.

The following general trends were observed for vowels:

- The vowels most seriously affected in L2 SAE are those which do not exist in the native African languages, where the closest vowel in the mother tongue is substituted. These include /ɑ/ as in *are*, /æ/ as in *can*, /ə/ as in *other* and /ɪ/ as in *this*.
- Central vowels such as /ə/, /ɜ:/ and /l̩/ are avoided in L2 speech and pronounced more peripherally, towards /ʌ/, /a/ or /u/.
- The long and short forms of certain vowels like /ɔ/ and /i/ (/AO/ and /IY/) seem to be affected differently, where the long form (for instance /ɔ:/ in *all* or /i:/ in *be*) is more affected by L2 accent than the short form (for example /ɔ/ in *not* or /i/ in the word *in*).
- The vowel /o/ is located dramatically more to the back of vowel space in second language speech (near /ʊ/) as compared to the neutral L1 pronunciation (near /ə/).

The detailed vowel results are summarised in Tables 4.1 and 4.2. The fist column indicates the ARPABET symbol of the vowel, while the second column indicates the word context. This is followed by the *degrees of freedom* of the ANOVA test (DOF = total number of utterances in both language groups - 2). The next three columns indicate the ANOVA F-ratio in each of the formant dimensions: a light-grey block indicates that the two distributions differed significantly to a 95% level of certainty, while a dark-grey block denotes a significant difference to a certainty of 99%. The formant result columns are concluded by a difference score ranging from 0 (no difference) to 6 (99% differences in all formant dimensions) calculated as follows: for each dimension where the ANOVA F-ratio reached the 95% level of certainty the score is incremented

by 1, while it is incremented by 2 for 99% certain significant differences. Similarly, the next 13 columns indicate the ANOVA results in each of the Mel-scaled cepstral dimensions, with 95% and 99% certain significant differences indicated by light- and dark-grey blocks respectively. The final column again shows a relative difference score similar to the formant case, calculated across the 13 cepstral dimensions (ranging from 0 to 26).

As an example, the first row of Table 4.1 indicates that the /ɑ/ (/AA/) in *are*, differs significantly to a certainty of 99% in the second formant dimension, giving a formant difference score of 2. The two cepstral models for this vowel differ to a 99% level of certainty in four of the cepstral dimensions, resulting in a difference score of 8.

In the case of diphthongs, the following trends were observed:

- The greatest factor influencing diphthongs is again vowel substitution, where an unfamiliar diphthong element is either replaced by the nearest native vowel, or dropped. In the former case over-articulation of diphthongs can appear, while in the latter the diphthong is reduced to a vowel. Examples include the /ʊə/ as in *with*, which becomes /ʊi/ and /əʊ/ as in *go*, which is replaced by the vowel /ʊ/.
- All centering diphthongs are affected, where the second element (/ə/) is replaced with a more peripheral vowel, such as [ðɛa] for *their* or [hɪɛ] for *here*.

The detailed diphthong results are summarised in Table 4.3. As in the vowel results tables, the first three columns indicate the phoneme, word context and degrees of freedom. This is followed by a set of three columns repeated for each of the three formant dimensions. Each set indicates the ANOVA results for the three spline sections of that formant dimension (light- and dark-grey blocks indicate a significant difference to a certainty of 95% and 99% respectively). The next column gives an aggregate formant score ranging between 0 and 18, calculated as before, indicating the degree of difference between two formant models. These formant results are followed by the cepstral

Summary of vowel results

| Vowel | Context | DOF | Formants F1 | F2 | F3 | Score | MFCC 1 - 13 Score |
|---|---|---|---|---|---|---|---|
| aa | are_aa | 31 | 0.02 | 22.55 | 2.04 | 2 | 8 |
| ae | africa_ae | 27 | 0.00 | 36.73 | 0.08 | 2 | 13 |
| | and_ae | 31 | 8.18 | 3.11 | 0.20 | 2 | 4 |
| | as_ae | 26 | 8.87 | 7.92 | 0.06 | 4 | 9 |
| | at_ae | 25 | 11.38 | 0.71 | 1.20 | 2 | 9 |
| | back_ae | 20 | 5.01 | 2.97 | 0.06 | 1 | 5 |
| | can_ae | 22 | 5.37 | 1.10 | 2.21 | 1 | 6 |
| | has_ae | 25 | 4.08 | 0.14 | 0.04 | 0 | 4 |
| | have_ae | 38 | 12.58 | 0.25 | 2.42 | 2 | 5 |
| | that_ae | 34 | 0.01 | 0.14 | 0.01 | 0 | 2 |
| ah | but_ah | 35 | 1.39 | 2.45 | 0.98 | 0 | 12 |
| | come_ah | 22 | 0.56 | 5.03 | 1.38 | 1 | 4 |
| | just_ah | 21 | 1.26 | 3.51 | 0.50 | 0 | 5 |
| | other_ah | 36 | 1.75 | 6.24 | 0.45 | 1 | 5 |
| | some_ah | 26 | 1.09 | 0.02 | 0.00 | 0 | 1 |
| ao | all_ao | 30 | 36.95 | 0.04 | 4.79 | 3 | 14 |
| | also_ao | 26 | 7.64 | 5.17 | 6.00 | 3 | 10 |
| | because_ao | 25 | 2.49 | 6.16 | 3.13 | 1 | 10 |
| | for_ao | 25 | 0.00 | 0.94 | 1.15 | 0 | 5 |
| | four_ao | 22 | 12.01 | 0.01 | 0.91 | 2 | 9 |
| | from_ao | 33 | 12.29 | 0.90 | 0.01 | 1 | 5 |
| | got_ao | 21 | 3.41 | 2.77 | 2.66 | 0 | 6 |
| | more_ao | 24 | 18.50 | 1.70 | 7.40 | 3 | 6 |
| | not_ao | 38 | 5.99 | 0.01 | 1.41 | 1 | 3 |
| | of_ao | 27 | 5.04 | 1.95 | 0.14 | 1 | 3 |
| | on_ao | 31 | 6.92 | 3.73 | 4.08 | 2 | 8 |
| ax | africa_ax | 19 | 0.58 | 6.85 | 1.22 | 1 | 8 |
| axr | other_axr | 34 | 1.73 | 8.73 | 0.29 | 2 | 8 |

Table 4.1: Summary of the analysis of variance results for vowels in both formant space and the Mel-scaled cepstral domain. Results shown for vowels /ɑ/ through /ɝ/ (/AA/ through /AXR/)

ANOVA results. Three columns indicate the L1/L2 difference scores (ranging from 0 to 26) of the three successive HMM states. The score is calculated over the 13 cepstral dimensions for each state. Two arbitrary levels were chosen to aid in visualising these results: a score below 8 is indicated in a white block (small difference), from 9 to 14

Summary of vowel results (continued)

| Vowel | Context | DOF | Formants F1 | F2 | F3 | Score | MFCC 1 - 13 Score |
|---|---|---|---|---|---|---|---|
| eh | get_eh | 27 | 16.81 | 8.64 | 2.98 | 4 | 9 |
| | sabc_eh | 21 | 19.44 | 17.29 | 0.04 | 4 | 12 |
| | said_eh | 20 | 3.08 | 3.01 | 0.56 | 0 | 9 |
| | very_eh | 21 | 5.41 | 3.18 | 0.02 | 1 | 4 |
| el | people_el | 29 | 5.70 | 0.18 | 2.55 | 1 | 8 |
| er | first_er | 18 | 6.01 | 0.00 | 0.37 | 1 | 4 |
| ih | this_ih | 41 | 2.71 | 24.70 | 0.86 | 2 | 13 |
| iy | africa_iy | 18 | 3.82 | 0.35 | 4.89 | 1 | 0 |
| | be_iy | 33 | 0.30 | 13.36 | 0.21 | 2 | 9 |
| | been_iy | 27 | 0.04 | 8.99 | 4.91 | 3 | 6 |
| | being_iy | 20 | 4.62 | 2.93 | 1.72 | 1 | 2 |
| | he_iy | 22 | 2.83 | 9.27 | 3.38 | 2 | 9 |
| | his_iy | 19 | 1.45 | 1.97 | 3.85 | 0 | 7 |
| | if_iy | 22 | 6.03 | 0.01 | 0.04 | 1 | 2 |
| | in_iy | 24 | 0.20 | 0.01 | 0.00 | 0 | 0 |
| | is_iy | 31 | 0.66 | 2.12 | 0.16 | 0 | 4 |
| | its_iy | 22 | 10.20 | 0.00 | 0.89 | 2 | 5 |
| | people_iy | 26 | 3.67 | 3.19 | 9.30 | 2 | 10 |
| | sabc_iy1 | 19 | 0.00 | 6.90 | 0.03 | 0 | 6 |
| | sabc_iy2 | 25 | 0.19 | 16.27 | 6.64 | 2 | 10 |
| | see_iy | 20 | 2.79 | 4.14 | 0.47 | 1 | 5 |
| | very_iy | 19 | 0.04 | 6.65 | 4.88 | 3 | 6 |
| ow | also_ow | 23 | 0.41 | 12.40 | 0.20 | 2 | 8 |
| | go_ow | 20 | 5.25 | 64.15 | 1.77 | 3 | 21 |
| | know_ow | 20 | 4.18 | 88.47 | 0.46 | 2 | 16 |
| | no_ow | 25 | 4.37 | 25.47 | 5.13 | 4 | 15 |
| | so_ow | 31 | 7.07 | 46.10 | 7.15 | 4 | 18 |
| | those_ow | 19 | 193.16 | 103.61 | 0.08 | 4 | 15 |
| uw | do_uw | 19 | 3.01 | 25.49 | 0.35 | 2 | 15 |
| | to_uw | 26 | 0.08 | 20.78 | 0.01 | 2 | 17 |
| | two_uw | 28 | 1.94 | 63.30 | 2.17 | 2 | 16 |

Table 4.2: Summary of the analysis of variance results for vowels in both formant space and the Mel-scaled cepstral domain. Results shown for vowels /ɛ/ through /u/ (/EH/ through /UW/)

as light-grey (medium difference) and 15 to 26 as dark-grey (large difference). This is followed by the sum of these three scores, resulting in a relative measure of the dif-

Summary of diphthong results

| Diphthong | Context | DOF | Formant F1 1 | F1 2 | F1 3 | F2 1 | F2 2 | F2 3 | F3 1 | F3 2 | F3 3 | Score | MFCC Score 1 | Score 2 | Score 3 | Total | Diph Strength L1 Σ | L2 Σ | Ratio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Section → | | | | | | | | | | | | | | | | | | |
| aw | now_aw | 26 | | | | | | | | | | 6 | 9 | 8 | 8 | 25 | 9 | 6 | 0.67 |
| | south_aw | 30 | | | | | | | | | | 5 | 10 | 12 | 5 | 27 | 11 | 2 | 0.18 |
| awr | our_awr | 24 | | | | | | | | | | 11 | 11 | 11 | 9 | 31 | 6 | 5 | 0.83 |
| ay | by_ay | 29 | | | | | | | | | | 12 | 11 | 13 | 9 | 33 | 12 | 11 | 0.92 |
| | time_ay | 22 | | | | | | | | | | 3 | 7 | 9 | 10 | 26 | 4 | 1 | 0.25 |
| ehaxr | their_ehaxr | 31 | | | | | | | | | | 16 | 8 | 16 | 13 | 37 | 4 | 8 | 2.00 |
| | there_ehaxr | 33 | | | | | | | | | | 0 | 9 | 4 | 5 | 18 | 4 | 5 | 1.25 |
| ey | cape_ey | 20 | | | | | | | | | | 10 | 10 | 12 | 2 | 24 | 8 | 1 | 0.13 |
| | sabc_ey | 20 | | | | | | | | | | 2 | 6 | 12 | 4 | 22 | 7 | 3 | 0.43 |
| | they_ey | 25 | | | | | | | | | | 12 | 13 | 19 | 16 | 48 | 12 | 3 | 0.25 |
| iyaxr | here_iyaxr | 23 | | | | | | | | | | 12 | 2 | 8 | 6 | 16 | 9 | 5 | 0.56 |
| ow | also_ow | 23 | | | | | | | | | | 8 | 8 | 13 | 9 | 30 | 3 | 1 | 0.33 |
| | go_ow | 20 | | | | | | | | | | 8 | 20 | 15 | 15 | 50 | 6 | 0 | 0.00 |
| | know_ow | 20 | | | | | | | | | | 10 | 18 | 18 | 12 | 48 | 1 | 2 | 2.00 |
| | no_ow | 25 | | | | | | | | | | 15 | 13 | 13 | 14 | 40 | 5 | 11 | 2.20 |
| | so_ow | 31 | | | | | | | | | | 16 | 19 | 19 | 16 | 54 | 6 | 7 | 1.17 |
| | those_ow | 19 | | | | | | | | | | 12 | 13 | 16 | 19 | 48 | 1 | 2 | 2.00 |
| waa | one_waa | 41 | | | | | | | | | | 1 | 7 | 6 | 2 | 15 | 17 | 18 | 1.06 |
| | was_waa | 27 | | | | | | | | | | 5 | 3 | 8 | 10 | 21 | 13 | 10 | 0.77 |
| | what_waa | 19 | | | | | | | | | | 5 | 0 | 6 | 6 | 12 | 9 | 14 | 1.56 |
| wae | well_wae | 21 | | | | | | | | | | 2 | 3 | 2 | 8 | 13 | 8 | 8 | 1.00 |
| weh | where_weh | 25 | | | | | | | | | | 10 | 8 | 6 | 5 | 19 | 17 | 9 | 0.53 |
| wel | will_wel | 27 | | | | | | | | | | 14 | 5 | 7 | 16 | 28 | 3 | 8 | 2.67 |
| wih | with_wih | 25 | | | | | | | | | | 10 | 7 | 9 | 12 | 28 | 8 | 10 | 1.25 |
| wiy | we_wiy | 32 | | | | | | | | | | 4 | 8 | 5 | 10 | 23 | 13 | 17 | 1.31 |
| | which_wiy | 24 | | | | | | | | | | 3 | 9 | 11 | 9 | 29 | 10 | 14 | 1.40 |
| yuh | you_yuh | 21 | | | | | | | | | | 9 | 3 | 12 | 5 | 20 | 11 | 6 | 0.55 |

Table 4.3: Summary of the analysis of variance results for diphthongs in both formant space and the Mel-scaled cepstral domain.

ference between the two diphthong models. The final three columns indicate changes in diphthong strength, with the first two columns indicating the diphthong strength measure for the L1 and L2 models (ranging from 0 to 26), followed by the ratio of

the two (L2/L1). Using arbitrarily chosen levels of 0.5 and 2, strong monophthongisation ($\leq$ 0.5) and diphthongisation ($\geq$ 2) are shown as light- and dark-grey blocks respectively.

It was found that in most cases the differences evident in formant space directly translates to significant changes in the Mel-scaled cepstral domain, which would adversely affect automatic speech recognition systems. Formant space and other acoustic analysis techniques can therefore be effectively used to determine which phoneme models need to be updated for improved L2 speech recognition performance.

| *Vowel* | *ARPABET* | *Average MFCC difference score* | *Diphthong* | *ARPABET* | *Average MFCC difference score* |
|---|---|---|---|---|---|
| u | UW | 16 | əʊ | OW | 45 |
| o | OW | 15.5 | aʊ˞ | AWR | 31 |
| ɪ | IH | 13 | eɪ | EY | 31 |
| ɛ | EH | 8.5 | aɪ | AY | 30 |
| ɑ | AA | 8 | ɛə˞ | EH-AXR | 28 |
| ə | AX | 8 | ʊl̩ | WEL | 28 |
| ə˞ | AXR | 8 | ʊə | WIH | 28 |
| l̩ | EL | 8 | aʊ | AW | 26 |
| ɔ | AO | 7.2 | ʊɪ | WIY | 26 |
| æ | AE | 6.3 | ju | YUH | 20 |
| ʌ | AH | 5.4 | ʊɛ | WEH | 19 |
| i | IY | 5.4 | ɪə˞ | IY-AXR | 16 |
| ɝ | ER | 4 | ʊa | WAA | 16 |
| | | | ʊæ | WAE | 13 |

Table 4.4: Simplified results for the vowels and diphthongs, ignoring word-context.

In an attempt at simplifying the results as a guide to adapting ASR systems, the average Mel-scaled cepstral difference scores across all contexts for each vowel/diphthong is shown in Table 4.4. These lists are sorted from the phoneme models most seriously affected in L2 SAE to those least affected. However, this is certainly an over-

simplification of the problem, since word-context plays an important role and needs to be taken into account.

The results tabled in this section serve as a guideline for the adaptation of English ASR systems in South Africa. Further studies should therefore investigate how these results can be applied to improve recognition accuracy in SAE speech interfaces. Ways to adapt such systems would include updating their phoneme and word models where the L2 pronunciations differ significantly from the L1 norm. Higher-level adaptations may also be required where minimal pairs are not retained, such as updated language and grammar models.

# Appendix A

# Fisher significance table

This appendix lists the $95^{th}$ and $99^{th}$ percentile significance levels (Table A.1) for the $F$ ratio of a two-class analysis of variance (ANOVA) problem. The first column indicates the degrees of freedom (DOF) of the ANOVA test, the second column indicates the value the $F$ ratio must exceed to constitute a significant difference between the two classes with a certainty of 95%. The third column indicates the level for a certainty of 99%.

As an example, if we have two groups of data with 17 and 18 data elements each, the degrees of freedom (DOF) are $17 + 18 - 2 = 33$. The relevant row in the table is the one for which $x_i < DOF \leq x_{i+1}$ where $x_i$ and $x_{i+1}$ are the listed degrees of freedom in two consecutive rows of the table. From the row with $DOF = 40$ we therefore see that the 95% and 99% levels are 4.08 and 7.31 respectively.

| *DOF* | *95% level* | *99% level* | *DOF* | *95% level* | *99% level* |
|---|---|---|---|---|---|
| 1 | 161 | 4052 | 18 | 4.41 | 8.29 |
| 2 | 18.5 | 98.5 | 19 | 4.38 | 8.18 |
| 3 | 10.1 | 34.1 | 20 | 4.35 | 8.10 |
| 4 | 7.71 | 21.1 | 21 | 4.32 | 8.02 |
| 5 | 6.61 | 16.3 | 22 | 4.30 | 7.95 |
| 6 | 5.99 | 13.7 | 23 | 4.28 | 7.88 |
| 7 | 5.59 | 12.2 | 24 | 4.26 | 7.82 |
| 8 | 5.32 | 11.3 | 25 | 4.24 | 7.77 |
| 9 | 5.12 | 10.6 | 26 | 4.23 | 7.72 |
| 10 | 4.96 | 10.0 | 27 | 4.21 | 7.68 |
| 11 | 4.84 | 9.65 | 28 | 4.20 | 7.64 |
| 12 | 4.75 | 9.33 | 29 | 4.18 | 7.60 |
| 13 | 4.67 | 9.07 | 30 | 4.17 | 7.56 |
| 14 | 4.60 | 8.86 | 40 | 4.08 | 7.31 |
| 15 | 4.54 | 8.68 | 60 | 4.00 | 7.08 |
| 16 | 4.49 | 8.53 | 120 | 3.92 | 6.85 |
| 17 | 4.45 | 8.40 | $\infty$ | 3.84 | 6.63 |

Table A.1: $95^{th}$ and $99^{th}$ percentile significance levels for a two-class problem (from Fisher and Yates [48]).

# Appendix B

# Extended formant plots

In this appendix the detailed formant plots are shown in colour, from which the graphs given in Chapter 3 were derived. The vowel graphs given here include the actual data points from which the distributions used in the ANOVA tests were calculated.

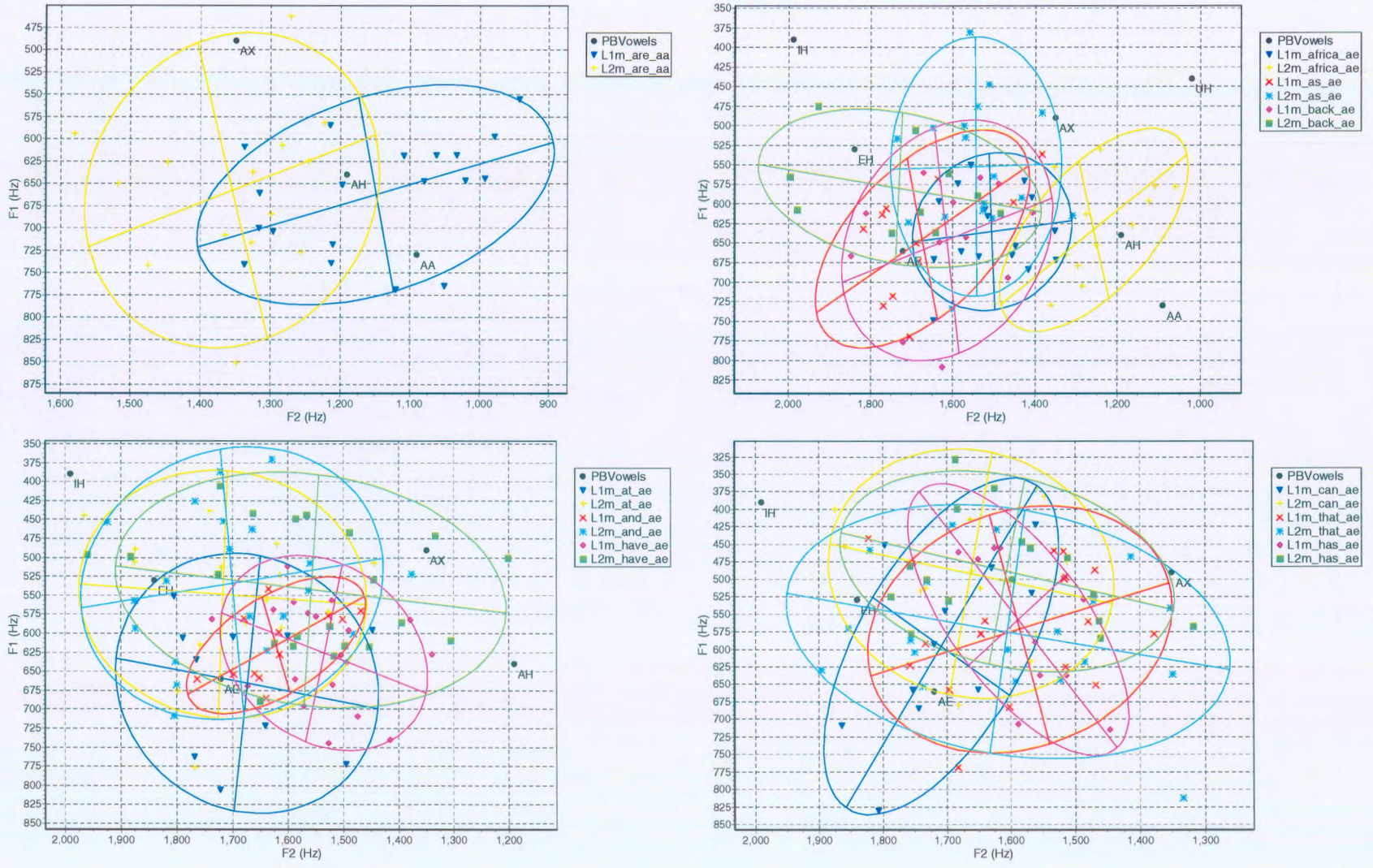


Figure B.1: Formant results for the vowels /ɑ/ and /æ/ (/AA/ and /AE/).

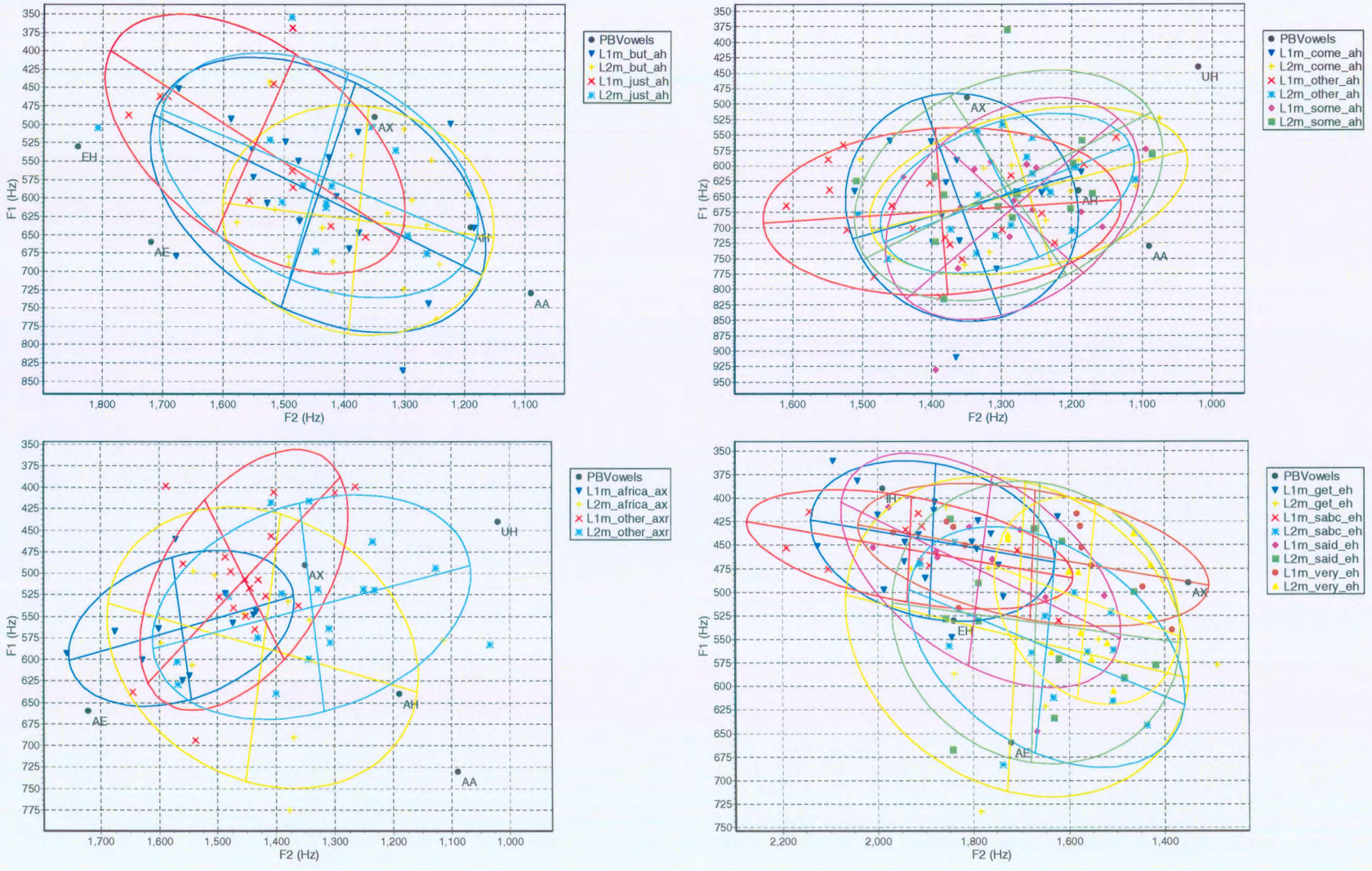


Figure B.2: Formant results for the vowels /ʌ/, /ə/, /ɚ/ and /ɛ/ (/AH/, /AX/, /AXR/ and /EH/).

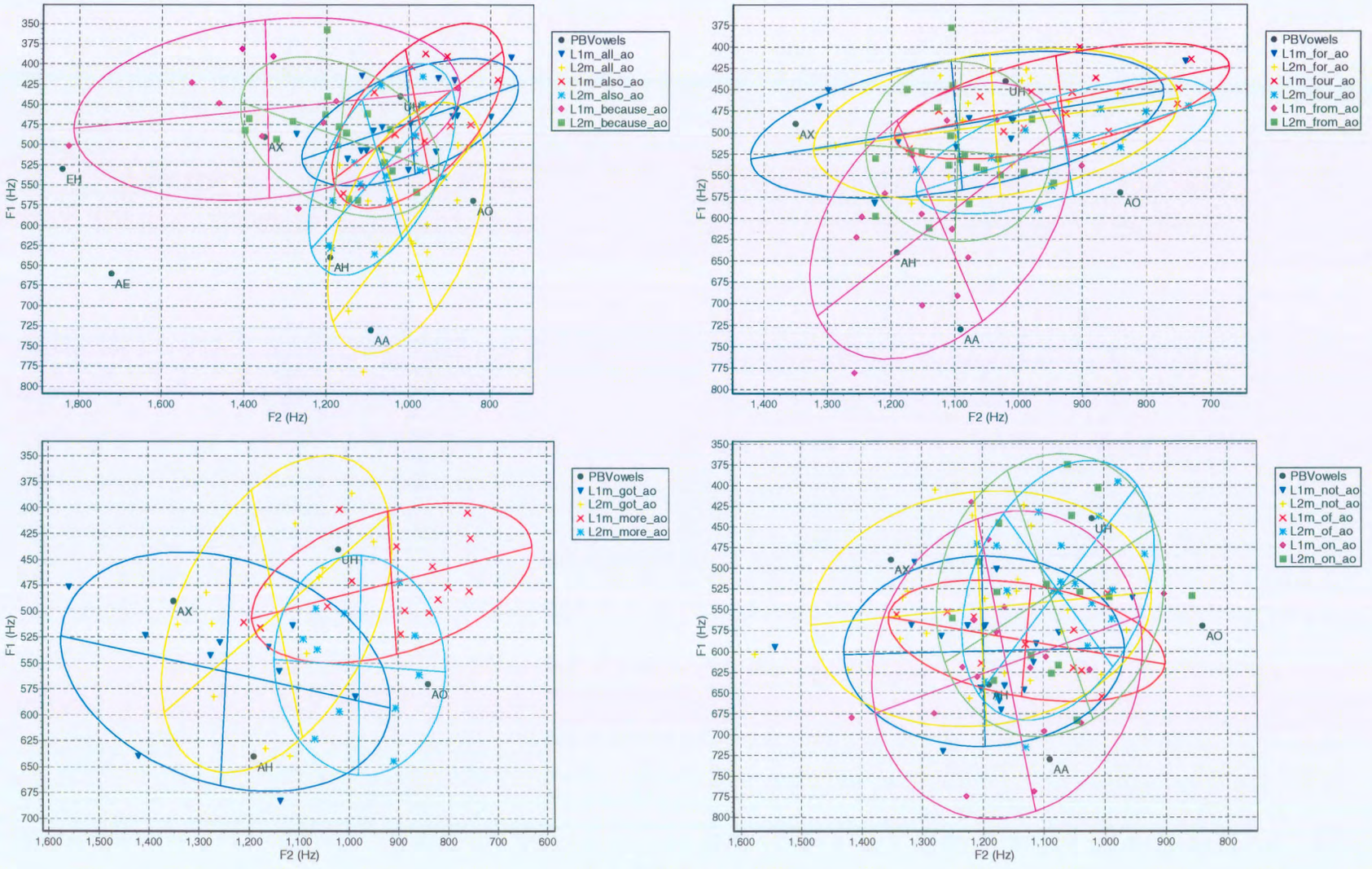


Figure B.3: Formant results for the vowel /ɔ/ (/AO/).

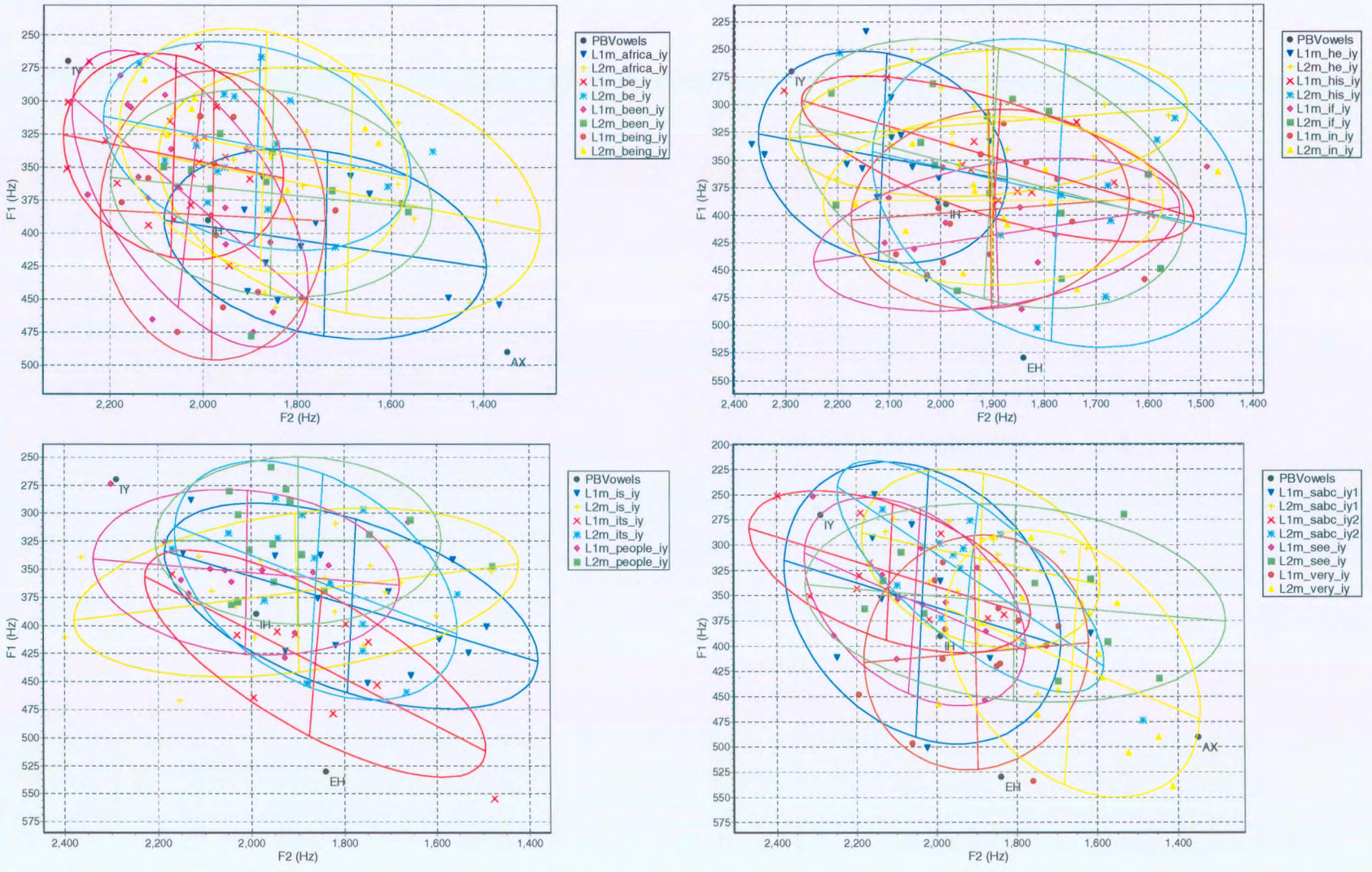


Figure B.4: Formant results for the vowel /i/ (/IY/).

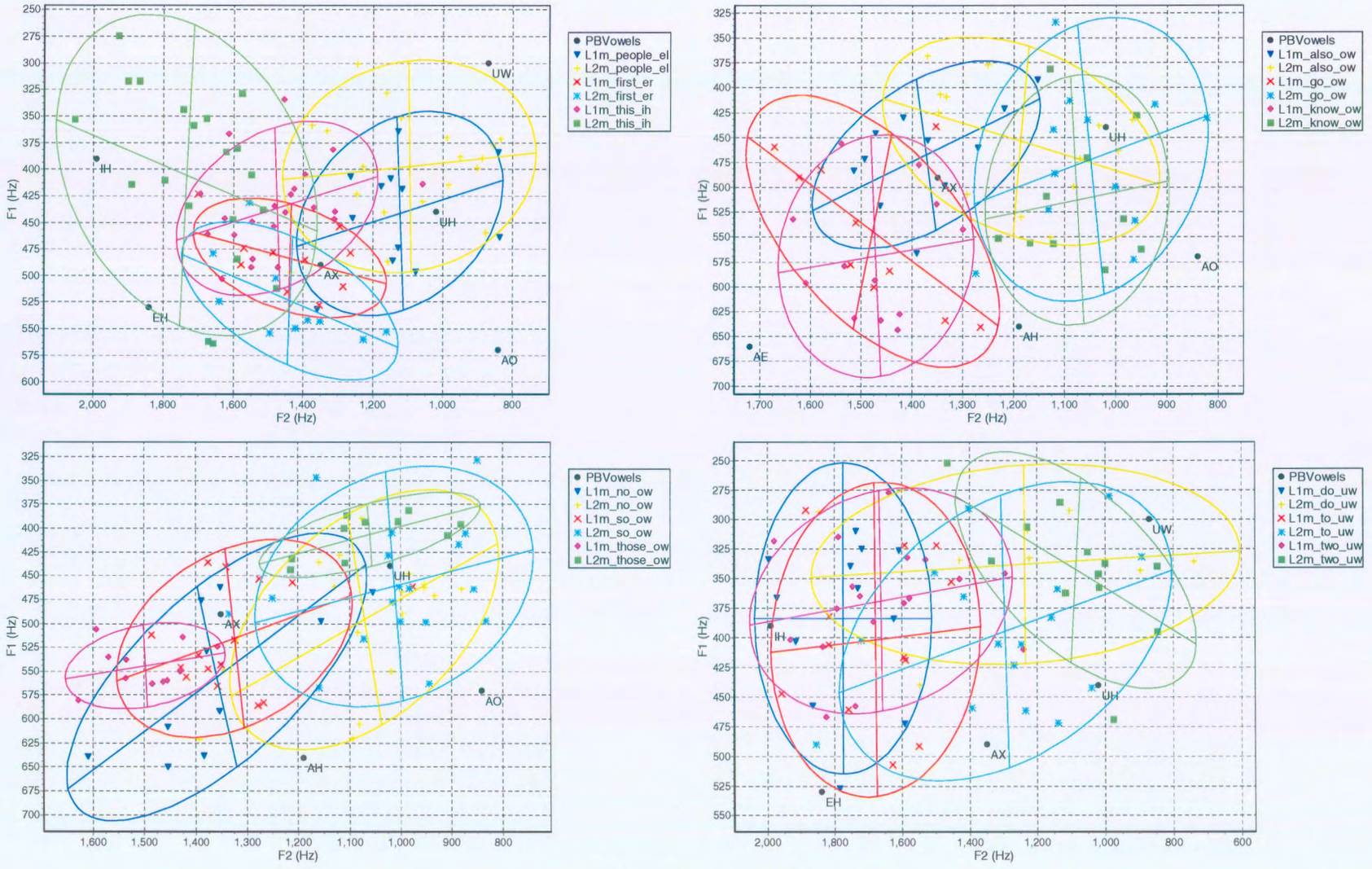


Figure B.5: Formant results for the vowels /l̩/, /ɝ/, /ɪ/, /o/ and /u/ (/EL/, /ER/, /IH/, /OW/ and /UW/).

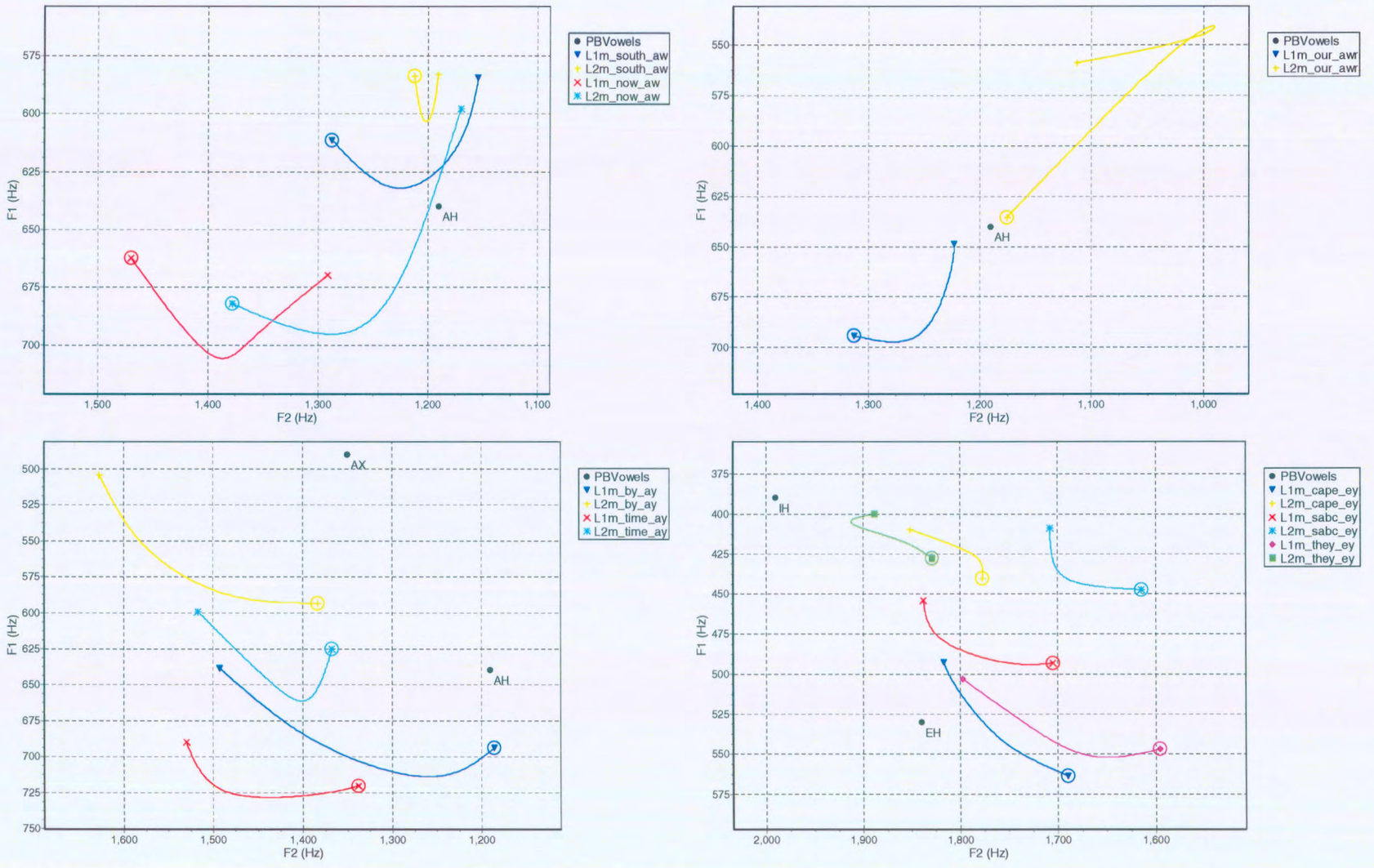


Figure B.6: Formant results for the diphthongs /aʊ/, /aʊˠ/, /aɪ/ and /əɪ/ (/AW/, /AWR/, /AY/ and /EY/).

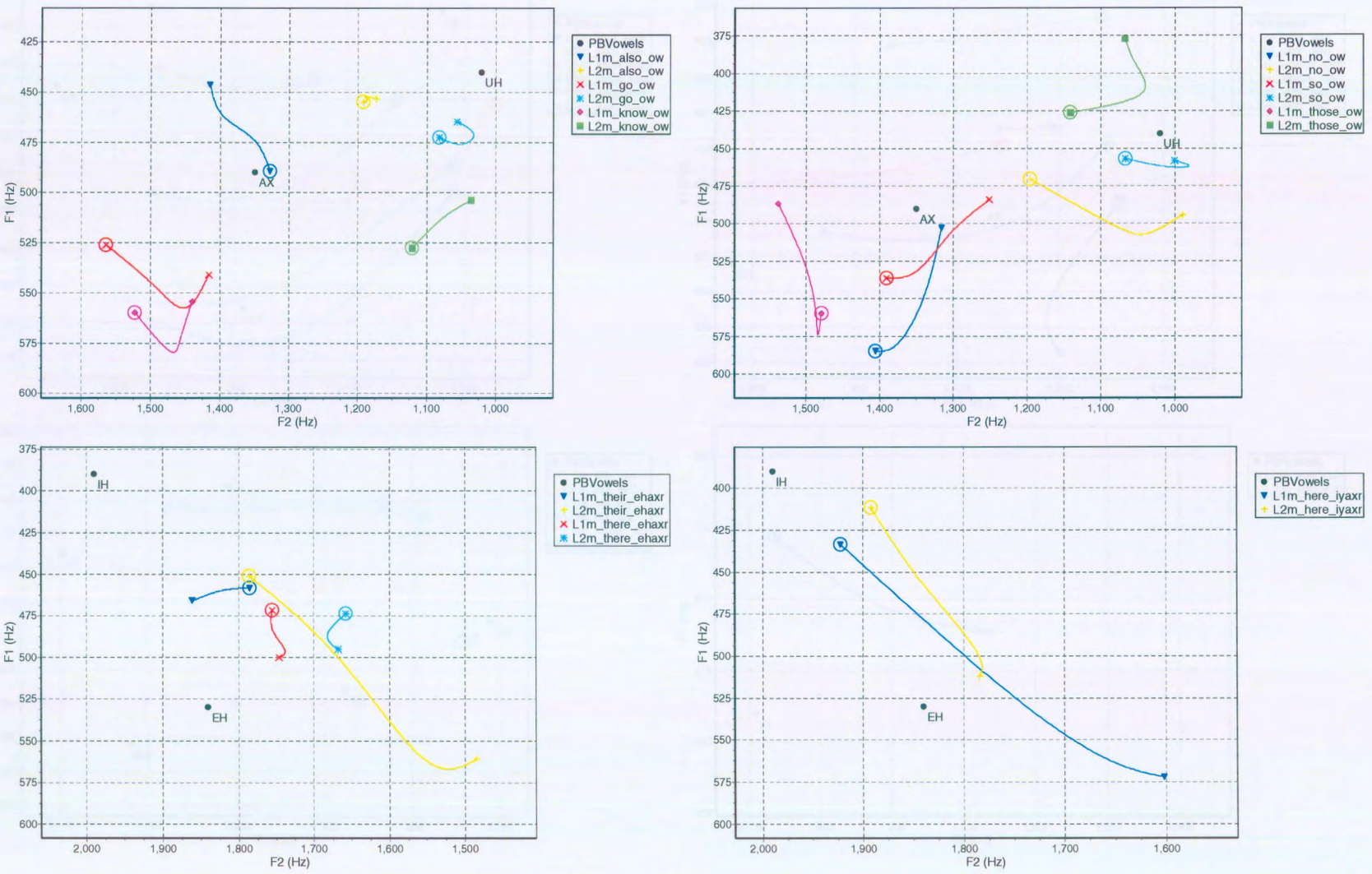


Figure B.7: Formant results for the diphthongs /əʊ/, /ɛɚ/ and /ɪɚ/ (/OW/, /EH-AXR/ and /IY-AXR/).

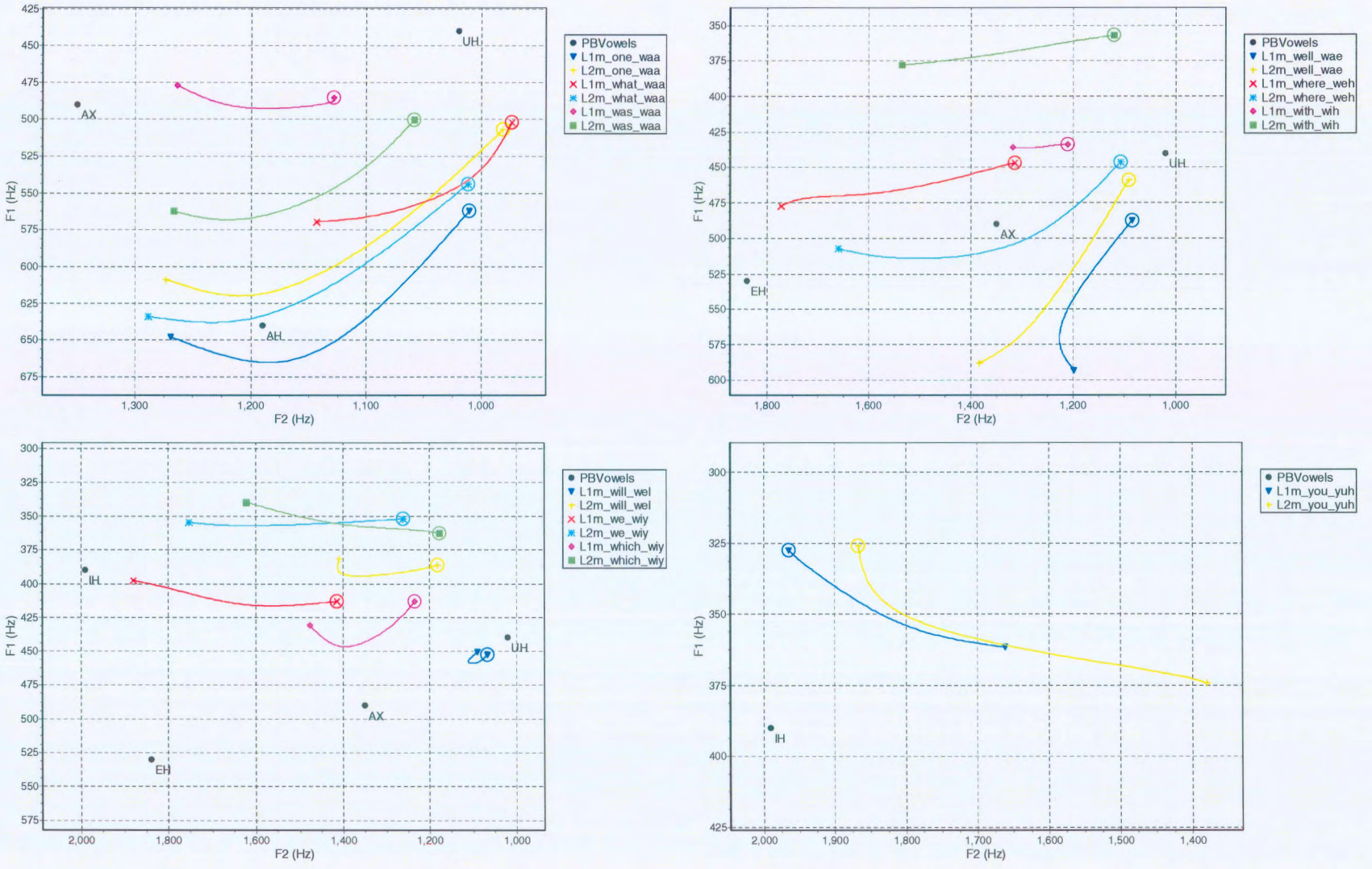


Figure B.8: Formant results for the diphthongs /ʊa/, /ʊæ/, /ʊɛ/, /ʊə/, /ʊl̩/, /ʊɪ/ and /ju/ (/WAA/, /WAE/, /WEH/, /WIH/, /WEL/, /WIY/ and /YUH/).

# Appendix C

# Accompanying software and data

The accompanying compact disks contain the data and some of the software tools used in this study as listed below.

The data is located on both compact disks as indicated below. In all cases first and second language data are stored in `L1` and `L2` sub-folders respectively:

- `raw_data` folder (both compact disks): Contains the raw speech data in 16-bit signed PCM WAV format, sampled at 22050 Hz. Each speech file has an associated text transcription (`.txn`) file. Label (`.xref`) files are also provided, where available, containing word-level labels of the speech files. A label file is present when the speech file contains words included in the final data set, as detailed in Section 3.2.4.

- `processed_data\words` folder (compact disk 2): Contains the audio files for the extracted words using the labels defined for the raw speech data. The same audio file format is used as for the raw data, and the associated label files indicate phoneme labels, where available. Phoneme labels are present for the phonemes contained in the final data set, as listed in Section 3.2.4.

- `processed_data\phonemes` folder (compact disk 2): Contains the audio files for the extracted phonemes.
- `processed_data\formants` folder (compact disk 2): Contains the formant tracks for each phoneme as a text file description of an $n \times 5$ matrix (5 formant values for each of $n$ frames).

The software tools are located on the second compact disk, in the `tools` folder[1]:

- *Txscribe* - The software tool used to create/view the text transcriptions of the raw audio files.
- *Wyre* - A software tool for labelling speech files under Microsoft Windows using both the time signal and spectrogram.
- *GPlot* - A software tool to display vowel distributions and diphthong tracks in formant space.

[1]Please note, these tools are provided as-is, with no guarantees as to functionality or compliance in any way and the author(s) do not accept any liability for loss or damage due to the use of this software.

# References

[1] L.M. Arslan and J.H.L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 28–40, 1997.

[2] J.E. Flege, "Factors affecting the degree of perceived foreign accent in English sentences," *Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 70–79, July 1988.

[3] J.J. Schmied, *English in Africa*, Longman, London, 1991.

[4] B. van Rooy and G.B. van Huyssteen, "The vowels of BSAE: current knowledge and future prospects," *South African Journal of Linguistics, Supplement 38*, pp. 15–33, 2000.

[5] C.M. Doke, *The Southern Bantu Languages*, Oxford University Press, London, 1954.

[6] J.C. Wells, *Accents of English, Beyond the British Isles*, Cambridge University Press, London, 1982.

[7] D. Ziervogel, *Handbook of the speech and sound changes of the Bantu languages of South Africa*, University of South Africa, Pretoria, 1967.

[8] M. Jacobs, "Consonantal variation in Zulu English mesolect," *South African Journal of Linguistics*, vol. 12, no. 1, pp. 16–25, 1994.

[9] J.D. Brink and E.C. Botha, "Towards an acoustic comparison of first and second language South African English," in *Proceedings of the* $10^{th}$ *Annual Symposium of the Pattern Recognition Association of South Africa*, Stellenbosch, South Africa, Nov. 1999, vol. CD-ROM: paper number 0009.

[10] J.D. Brink and E.C. Botha, "An acoustic comparison of first and second language South African English," in *Proceedings of the* $11^{th}$ *Annual Symposium of the Pattern Recognition Association of South Africa*, Johannesburg, South Africa, Nov. 2000, pp. 1–6.

[11] J.D. Brink and E.C. Botha, "An acoustic comparison of the vowels and diphthongs of first and second language South African English," in *Proceedings of the* $12^{th}$ *Annual Symposium of the Pattern Recognition Association of South Africa*, Franschoek, South Africa, Nov. 2001, pp. 103–108.

[12] K.L. Pike, *Phonemics: A Technique for reducing Languages to Writing*, University of Michigan Press, Michigan, 1947.

[13] K.L. Pike, *Phonetics: A Critical Analysis of Phonetic theory and a Technique for the Practical Description of Sounds*, Oxford University Press, London, 1943.

[14] G. Knowles, *Patterns of spoken English (An introduction to English phonetics)*, Longman, London, New York, 1987.

[15] P. Roach, *English Phonetics and Phonology*, Cambridge University Press, London, 1983.

[16] T. Chiba, *The Vowel: Its Nature and Structure*, Phonetic Society of Japan, Tokyo, 1958.

[17] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey 07632, 1993.

[18] I.C. Ward, *The Phonetics of English*, Heffer, Cambridge, 1958.

[19] D. Jones, *The pronunciation of English*, Cambridge University Press, London, 1919.

[20] E.B. van Wyk, *Die Bantoetale, 'n Beknopte Oorsig*, J.L. van Schaik, Pretoria, 1966.

[21] L.J. Louwrens, *Northern Sotho*, Lincom Europa, Muenchen, 1995.

[22] D.T. Cole, *An Introduction to Tswana Grammar*, Longmans, Green and Co., Cape Town, 1955.

[23] A. Werner, *Introductory Sketch of the Bantu Languages*, Kegan Paul, Trench, Trubner and Co., London, 1919.

[24] C.M. Doke, *Text-Book of Zulu Grammar*, Longmans, Green and Co., Cape Town, 1941.

[25] J. McLaren, *A Xhosa Grammar*, Longmans, Green and Co., Cape Town, 1936.

[26] W.L. Lanham, *The way we speak*, Van Schaik, Pretoria, 1967.

[27] J.H.L. Hansen and L.M. Arslan, "Foreign accent classification using source generated prosodic features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, MI, USA, 1995, IEEE, vol. 1, pp. 836–839.

[28] L.M. Arslan and J.H.L. Hansen, "Language accent classification in American English," *Speech Communication*, vol. 18, pp. 353–367, 1996.

[29] L.W. Lanham, *The Pronunciation of South African English*, Balkema, Cape Town, 1967.

[30] L.W. Lanham and K.P. Prinsloo, *Language and communications studies in South Africa. Current issues and directions in research and enquiry*, Oxford University Press, Cape Town, 1978.

[31] B. van Rooy, D. Wissing, and C. van den Heever, "The perception of the vowels of Tswana-English," *South African Journal of Linguistics, Supplement 38*, pp. 89–122, 2000.

[32] R. Adendorff and M. Savini-Beck, "The teaching of English vowels and consonants in the new South Africa," *South African Journal for Language Teaching*, vol. 27, no. 3, pp. 232–248, 1993.

[33] J.E. Flege, "The production of 'new' and 'similar' phones in a foreign language: evidence for the effect of equivalence classification," *Journal of Phonetics*, vol. 15, pp. 47–65, 1987.

[34] F.M. Chreist, *Foreign accent*, Prentice Hall, Eaglewood Cliffs, N.J., 1964.

[35] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–253, 1961.

[36] L.F. Willems, "Robust formant analysis," *Institute for Perception Research Report 529, Eindhoven, The Netherlands*, pp. 1–25, 1986.

[37] N. Levinson, "The wiener RMS (root mean square) error criterion in filter design and prediction," *Journal of Mathematics and Physics*, vol. 25, pp. 261–278, 1946.

[38] J.H. Mathews, *Numerical Methods for Mathematics, Science and Engineering*, Prentice-Hall, London, 1992.

[39] S.B. Davis and P. Mermelstein, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, 1980.

[40] R Dugad and U.B. Desai, "A tutorial on hidden Markov Models," Tech. Rep. SPANN-96.1, Signal Processing and Artificial Neural Networks Laboratory, Department of Electrical Engineering, Indian Institute of Technology, Mumbai, India, May 1996.

[41] B.H. Juang and L.R. Rabiner, "An introduction to hidden Markov models," *IEEE Acoustics, Speech and Signal Processing Magazine*, pp. 4–16, June 1986.

[42] G.D. Forney Jr., "The Viterbi algorithm," *IEEE Proceedings*, vol. 61, no. 3, pp. 263–278, March 1973.

[43] B.H. Juang and L.R. Rabiner, "The segmental k-means algorithm for estimating the parameters of hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-38, no. 9, pp. 1639–1641, 1990.

[44] M.R. Spiegel, *Probability and Statistics*, McGraw-Hill, New York, 1980.

[45] G. Peterson and H. Barney, "Control methods used in a study of vowels," *Journal of the Acoustical Society of America*, vol. 42, pp. 175–184, 1952.

[46] C.P. Prinsloo, "A comparative acoustic analysis of the long vowels and diphthongs of Afrikaans and South African English," Master's dissertation, University of Pretoria, Pretoria, South Africa, 2000.

[47] A. Holbrook and G. Fairbanks, "Diphthong formants and their movements," *Journal of Speech and Hearing Research*, vol. 5, no. 1, pp. 38–58, 1962.

[48] R.A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Longman Group Ltd., London, 1964.