Making Accurate Formant Measurements: An Empirical Investigation of the Influence of the Measurement Tool, Analysis Settings and Speaker on Formant Measurements

Philip Thomas Harrison

PhD

University of York

Language and Linguistic Science

April 2013

Abstract

The aim of this thesis is to provide guidance and information that will assist forensic speech scientists, and phoneticians generally, in making more accurate formant measurements, using commonly available speech analysis tools. Formant measurements are an important speech feature that are often examined in forensic casework, and are used widely in many other areas within the field of phonetics. However, the performance of software currently used by analysts has not been subject to detailed investigation. This thesis reports on a series of experiments that examine the influence that the analysis tools, analysis settings and speakers have on formant measurements.

The influence of these three factors was assessed by examining formant measurement errors and their behaviour. This was done using both synthetic and real speech. The synthetic speech was generated with known formant values so that the measurement errors could be calculated precisely. To investigate the influence of different speakers on measurement performance, synthetic speakers were created with different third formant structures and with different glottal source signals. These speakers' synthetic vowels were analysed using Praat's normal formant measuring tool across a range of LPC orders.

The real speech was from a subset of 186 speakers from the TIMIT corpus. The measurements from these speakers were compared with a set of hand-corrected reference formant values to establish the performance of four measurement tools across a range of analysis parameters and measurement strategies.

The analysis of the measurement errors explored the relationships between the analysis tools, the analysis parameters and the speakers, and also examined how the errors varied over the vowel space. LPC order was found to have the greatest influence on the magnitude of the errors and their overall behaviour was closely associated with the underlying measurement process used by the tools. The performance of the formant trackers tended to be better than the simple Praat measuring tool, and allowing the LPC order to vary across tokens improved the performance for all tools. The performance was found to differ across speakers, and for each real speaker, the best performance was obtained when the measurements were made with a range of LPC orders, rather than being restricted to just one.

The most significant guidance that arises from the results is that analysts should have an understanding of the basis of LPC analysis and know how it is applied to obtain formant measurements in the software that they use. They should also understand the influence of LPC order and the other analysis parameters concerning formant tracking. This will enable them to select the most appropriate settings and avoid making unreliable measurements.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor Paul Foulkes for his support throughout.

Thank you to Peter French for giving me the opportunity to undertake the PhD and for his encouragement.

Thank you to my colleagues Louisa Stevens, Asfah Bhagdin, Richard Rhodes and Christin Kirchhübel for their support.

I am very grateful to Frantz Clermont for his enthusiasm, encouragement and for sharing his CAbS tracker.

I also would like to thank my parents for their continued support and encouragement.

Thank you, especially, to Virginie.

Declaration

This thesis has not previously been submitted for any degree other than Doctor of Philosophy of the University of York. This thesis is only my original work, except where otherwise stated. Other sources are acknowledged by explicit references.

Chapter 1 Introduction

The research presented in this thesis is motivated by the limited amount of practical guidance concerning the measurement of formants that is currently available within the field of forensic speech science. The goal of the thesis is to contribute to this guidance by providing empirically motivated advice and information to assist speech scientists, especially those working in forensic applications, when making and interpreting formant measurements. This guidance is derived from the results of a series of experiments that examine the influence of different software tools, analysis settings and speech material on formant measurement errors.

This chapter introduces what formants are, how they are measured and how they are used in forensic casework.

1.1 Formants

Before discussing what formants are and how they are measured, this chapter begins with a conceptual description of the source-filter model of speech production. This model is helpful in understanding the nature of formants and their measurement. It also forms the basis of the speech synthesis method employed in Chapter 4 and Chapter 5.

1.1.1 Source-Filter Model of Speech Production

A useful tool for the study of speech sounds and their production is the source-filter model of speech production. It can be encapsulated in the simple statement that a 'speech wave is the response of the vocal tract filter systems to one or more sound sources' (Fant 1960, p15). The implication is that speech sounds can be specified in terms of two components, a sound source and a filter response.

Vocalic sounds are often conceptualised in terms of a simple source-filter model, a representation of which is shown in Figure 1.1. In this model the sound source is the vocal folds, which produce sound by modulating airflow from the lungs as the folds open and close in a quasi-periodic manner. The sound produced by the vocal folds is often represented by a periodic amplitude waveform, shown in the top left of Figure 1.1, which has a pulse structure. An important property of this sound is its period (T_0), which is a measure of the time between pulses in the waveform. The inverse of the

period is the fundamental frequency (F_0) of the sound, i.e. the number of pulses within a second if measured in the unit Hertz.

A second important property of this source sound is that it is complex, meaning that it is composed of many frequencies. It is harmonic in nature, so the different frequencies, known as harmonics, are multiples of the fundamental frequency. The relative amplitudes of the fundamental and harmonics define the spectral characteristics of the source sound, which are influenced by factors such as vocal effort, fundamental frequency and the type of phonation. A spectral representation of a vocal fold sound is displayed in the bottom left of Figure 1.1 where the harmonics are the equally spaced vertical lines with decreasing amplitude.

The second part of the model, the filter, represents the resonant properties of the supralaryngeal vocal tract. The vocal tract is an acoustic space which shapes the frequency spectrum of the sound from the vocal folds as it passes through it. In the bottom centre of Figure 1.1 an example frequency response of the vocal tract is shown. The configuration of the vocal tract, in terms of tongue position, jaw height, lip rounding etc. alters the size and shape of the tract, which determines its resonance characteristics and therefore its frequency response. The frequency response is often characterised by the resonant frequencies which can be specified in terms of the centre frequency of the peak and the width of the peak, known as the bandwidth. The bandwidths of the peaks are governed by damping within the tract.

The resonant frequencies are those where the vocal tract allows sounds to interact more easily with the acoustic space and as a consequence the amplitude of these frequencies within the radiated speech sound is greater than others. The outcome of filtering the vocal fold sound source with the vocal tract response is represented in the spectrum of the radiated speech sound on the bottom right of Figure 1.1. The radiated amplitude waveform is shown above it. The radiated sound is still composed of the fundamental frequency and its harmonics, but their relative amplitudes, and therefore the spectral content, have been shaped by the filtering effect of the vocal tract. The influence of the resonant frequencies on the radiated speech sound is clearly visible as the global peaks within the radiated sound's frequency spectrum.

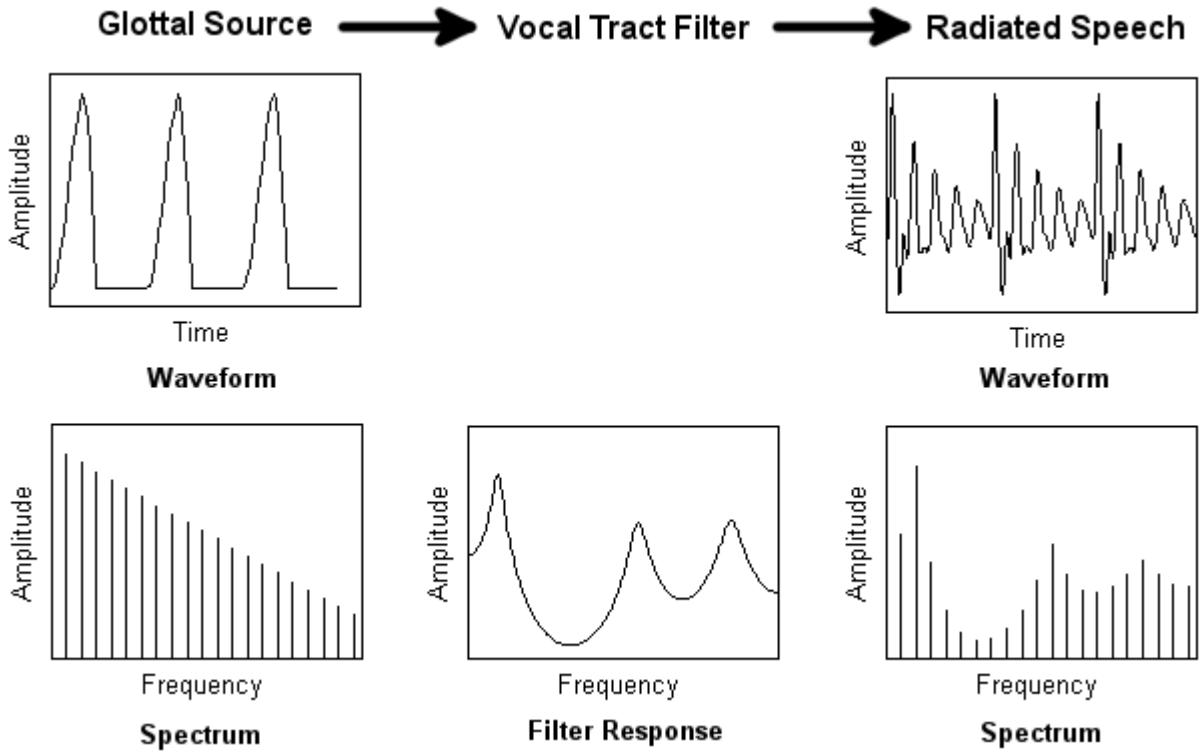

Figure 1.1 A conceptual representation of the source-filter model of speech production showing time-amplitude and frequency representations of the pulse-like glottal source on the left, the frequency response of the vocal tract in the centre and on the right the time-amplitude and frequency representations of the radiated speech sound resulting from the filtering of the glottal sound source by the vocal tract.

In this model the sound source and the filter are assumed to be independent. The resonant frequencies of the vocal tract and the fundamental frequency of the source with its associated harmonics are not related, but are free to vary independently of one another. It is possible for one of the harmonics to coincide with a resonant frequency, but such an occurrence would usually be by chance and is not a requirement or function of the speech production process.

The model of vocalic speech production described above is presented at a conceptual level rather than in mathematical terms. However, comprehensive descriptions of mathematical implementations of this model and comparisons with measurements from real speech data are presented in Fant (1960) and Stevens (1998), for example. These works also consider the modelling of consonantal sounds as well as more complex vocalic models such nasalised vowels where the nasal cavity is acoustically coupled to the oral cavity.

1.1.2 Definition of Formants

The preceding description of the source-filter model for the production of vocalic sounds introduced two important interrelated concepts. These are the resonant frequencies of the vocal tract and the spectral peaks in the radiated speech sound resulting from the filtering effect of the vocal tract. Whilst they are conceptually distinct, one being an acoustic property of the vocal tract and the other being a property of the radiated sound, the term 'formant' is often used to refer to both.

Whilst Fant (1960, p. 20) defines formants as the 'spectral peaks of the sound spectrum' he also notes that the two concepts 'should be held apart but in most instances resonance frequency and formant frequency may be used synonymously'. Conversely, Fry (1979, p. 76) states that 'formants are strictly the resonant frequencies of the driven system' (i.e. the vocal tract), 'but since a formant must give rise to a peak in the spectrum of the sound produced, the term formant is quite commonly applied to the frequency at which this peak occurs'. Johnson (1997, p. 84) is in agreement with Fry that 'the resonant frequencies of the vocal tract are also called formants', whereas Clark and Yallop (1995, p. 246) state that such a definition is 'technically imprecise' and that 'formants are a consequence of resonance, not resonance itself'.

These varying definitions show that there is no consensus on the precise use of the term formant. Perhaps what is most important is that where the distinction between the two definitions is relevant then it is made clear which meaning is intended by the use of the term. Alternatively, more verbose terminology or descriptions may be used where the use of the word formant could be confusing.

One aspect of formants where there is no disagreement is in the numbering convention used to describe the different resonances or spectral peaks. The resonance or peak with the lowest frequency is called the first formant (F1), the second lowest is the second formant (F2) and so on.

1.2 Measuring Formants

It is apparent from the source-filter model of speech production that determining the frequency of spectral peaks in a vocalic sound will yield information about the resonance characteristics and the configuration of the vocal tract that produced the sound. Measuring the frequency of the spectral peaks is inherently difficult due to the

spacing of the harmonics produced by the vocal fold source. Information about the shape of the spectrum only exists at the frequencies of the harmonics, and the overall shape, as well as the frequency of the peaks, must be inferred from this limited data. Given this inherent limitation the term 'formant estimate' might be more appropriate than 'formant measurement', but the latter will be used in this thesis for the sake of convention.

A number of measurement methods exist, and the necessary tools are nowadays readily available to analysts in free speech analysis software such as Praat (Boersma 2001) and WaveSurfer (Sjölander and Beskow 2006a). The following sections introduce three of the most common methods, including LPC analysis, which is the technique used in this thesis.

It is worth noting that methods also exist to derive or measure directly the resonant frequencies of the vocal tract rather than the spectral peaks of the speech signal. However, they either involve specialist medical imaging techniques, such as x-rays (Fant 1960) or MRI (Clément et al 2007), or specifically developed equipment such as that described in Epps et al (1997). These techniques are valuable research tools and have the potential to increase understanding of speech acoustics but they are not readily available to the majority of phoneticians and speech scientists, nor can they be employed for forensic casework.

1.2.1 Frequency Spectra

The simplest method of examining the frequency content of a sound is to generate a frequency spectrum. This is usually done by Fourier analysis, which deconstructs a time-amplitude representation of a signal into its constituent frequency components. Within computer software this is performed by the Fast Fourier Transform (FFT). The analysis can be performed on either a short section of the signal, a single analysis frame, or an average can be obtained over a longer time period. In order to accurately represent the frequency content of a signal and be able to measure precisely the frequency of features within it, a high resolution spectrum is required. Figure 1.2 shows such a spectrum for a 0.05 second section of the vowel /i:/ in the word 'he' spoken by the author.

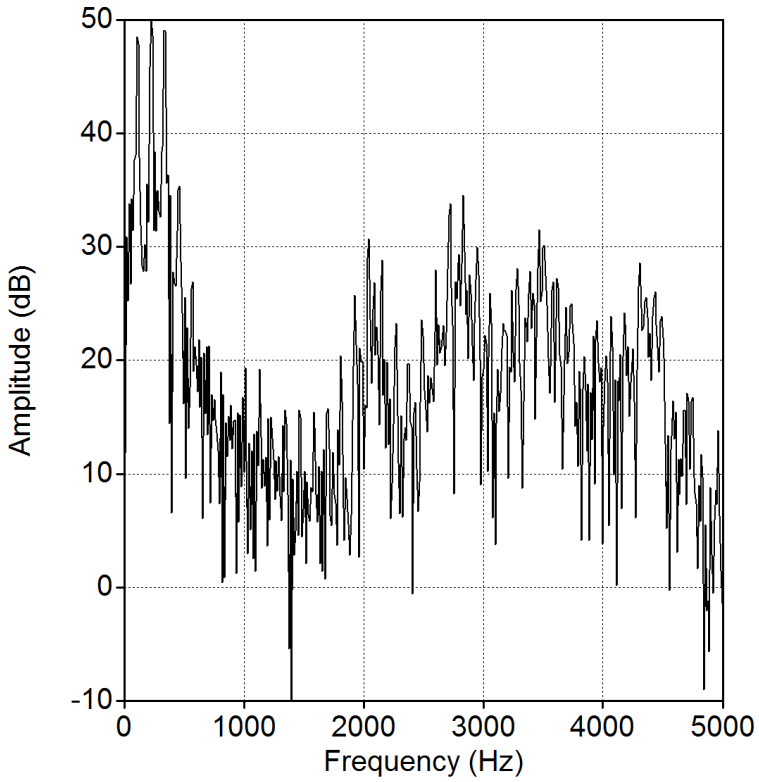

Figure 1.2 High resolution FFT spectrum of a 0.05 section of the vowel /i:/ in the word 'he' spoken by the author.

Figure 1.2 shows the harmonic structure of the vowel sound, where the left most peak in the spectrum represents the fundamental frequency, at approximately 100 Hz, and the other peaks are the harmonics, at approximately 100 Hz intervals. The variation in amplitude of the harmonics is also apparent. The normal way to measure the frequency of a feature in a spectral display is to place a cursor on the feature and read off its frequency value. This method is straightforward for determining the frequency of the fundamental and the harmonics since the peaks are reasonably well defined in the plot and a cursor can easily be aligned with them. However, the formants, i.e. the broader spectral peaks caused by the resonant frequencies of the vocal tract, are more problematic to measure since their location can only be inferred from the relative amplitudes of the harmonics.

One way of displaying the overall spectral shape rather than the constituent harmonics is to smooth the spectrum. This can be accomplished a number of ways including using a very short frame length of the order of 5 milliseconds. A smoothed spectrum of the same section of the /i:/ vowel spoken by the author is shown in Figure 1.3 with the peaks corresponding to the first five formants marked.

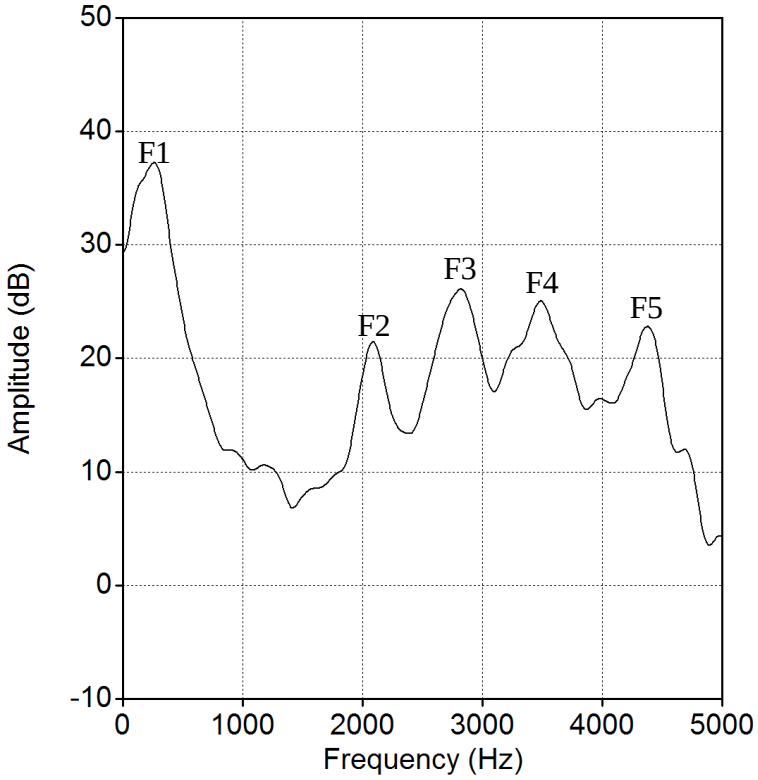

Figure 1.3 Smoothed FFT spectrum of the vowel /i:/ with the formants F1 to F5 marked.

In the smoothed spectrum the overall spectral shape and the locations of the peaks are now clearly visible. Measuring the frequency of the peaks, i.e. the formants, is now possible. However, this method has a number of limitations. Perhaps the most significant is that the extent of the smoothing can affect both the location and the appearance of peaks in the spectrum. If the smoothing is not sufficient then the harmonic peaks will still be visible, but if the smoothing is too great then definition will be lost and spectral peaks that are close together could merge. A further limitation of frequency spectra is that they only represent a point in time or an average over time and thus they cannot display the dynamic nature of speech. It is possible to plot a series of spectra across a number of analysis frames, in what is known as a cascade or waterfall plot, but these are not ideal for taking measurements from and they are often not available in commonly used software. Notwithstanding the limitations of spectra, they are often used in conjunction with spectrograms and LPC analysis to either check measurements made using one of the other methods or where the interpretation of the results from another method is problematic.

1.2.2 Spectrograms

Spectrograms are perhaps the most common method of visualising the frequency structure of speech. They are a series of frequency spectra over a period of time in which amplitude is represented by a varying colour scale, usually a greyscale, with higher amplitudes being darker shades of grey. Frequency is represented on the vertical axis, rather than the horizontal axis as is common for individual spectra, and time is represented on the horizontal axis. Since spectrograms are a series of spectra the same issues of resolution and smoothing occur. Figure 1.4 shows a spectrogram of the same /i:/ vowel in the word 'he' generated with a relatively long analysis frame (0.05 seconds), which produces what is known as a narrow-band spectrogram. Like the high resolution spectrum, the harmonic structure of the vowel is visible as the regularly spaced horizontal bars, especially at the lower frequencies. Again, the frequency of features within the spectrogram can be measured by placing a cursor at its location and reading off the frequency position of the cursor. Whilst this is easy to accomplish for harmonic features, narrow-band spectrograms are not suitable for measuring formants. The regions corresponding to the spectral peaks in the signal can be seen as the areas where several harmonics are darker but attempting to locate the frequency of the peak is problematic.

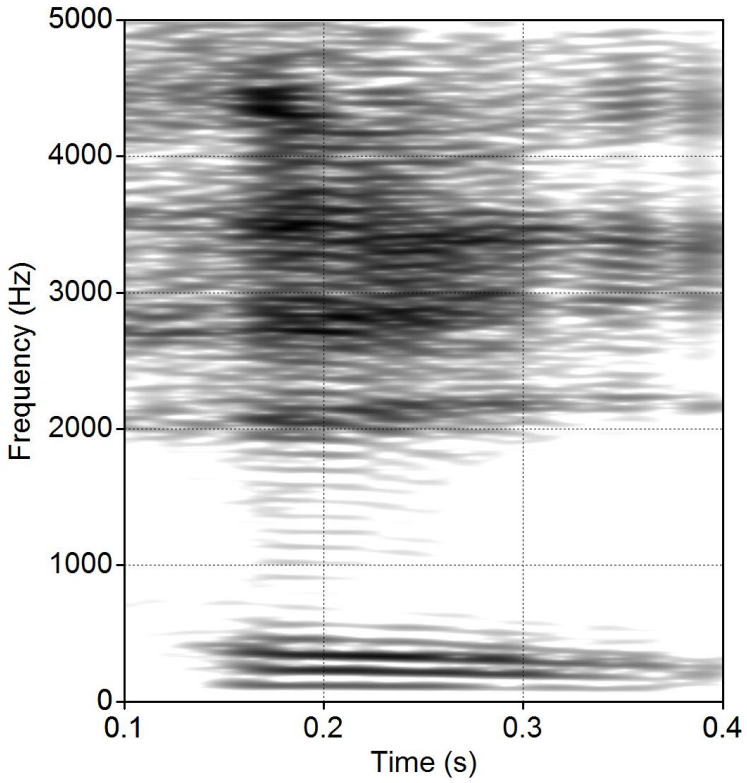

Figure 1.4 Narrow-band spectrogram of the vowel /i:/ in the word 'he'.

The solution to visualising the spectral peaks caused by the resonant frequencies of the vocal tract is again to use smoothing, which is normally achieved by using short duration analysis frames of the order of 5 milliseconds. The resulting spectrograms are often referred to as broad-band spectrograms and an example showing the same /i:/ vowel is in Figure 1.5 with the first five formants marked. The effective smoothing has resulted in the individual harmonics no longer being visible, and the spectral peaks can now be seen as the wider dark horizontal bars. The frequency of the peaks is measured by placing a cursor in the visible centre of each horizontal bar at a representative timing.

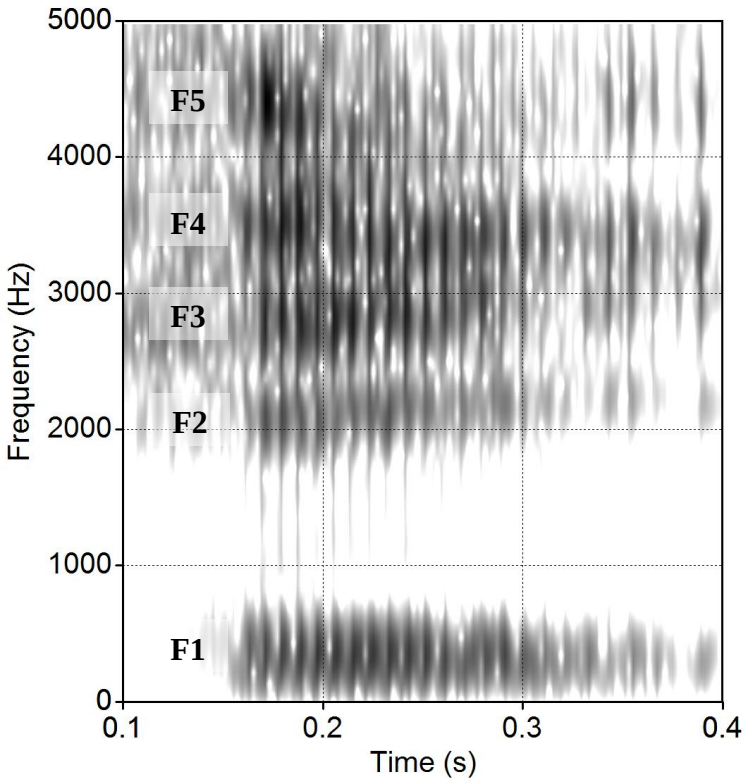

Figure 1.5 Broad-band spectrogram of the vowel /i:/ in the word 'he' with the formants F1 to F5 marked.

Spectrograms, like frequency spectra, also suffer from limitations caused by smoothing. The apparent location of peaks can be altered and closely space peaks can become merged as the degree of smoothing increases. Partial merging can be seen in Figure 1.5 where there is not clear separation between F3 and F4. Further discussion of the generation of spectrograms and other factors that can influence their appearance and interpretation can be found in Kent and Read (2002) and Howard (1998, 2002), among many others.

1.2.3 LPC

Frequency spectra and spectrograms are both convenient methods of visualising the frequency content of speech, and making measurements from them is relatively straightforward. However, taking measurements is essentially a manual process which is time consuming. Also, the smoothing required to make the spectral peaks visible introduces problems which are mentioned above. An alternative method of determining the frequency of spectral peaks which can overcome these issues is linear predictive coding (LPC) or linear prediction (LP) analysis.

LPC analysis of speech is fundamentally different from the spectral analysis methods described above, which produce frequency spectra from time-amplitude waveforms by means of the Fourier transform. LPC analysis considers the speech signal as the output of a source-filter speech production model and it determines the parameters for the model that result in the best estimate of the speech signal. Information about the speech signal, such as formant frequencies, is then derived from the model parameters. This method of analysing speech originates from the development of techniques to encode speech signals so they could be transmitted over low bit rate channels. Rather than transmitting the original speech signal, the parameter values of the model are sent, and the speech signal is reconstructed at the receiving end of the transmission channel (see Atal 2006 for a historical overview of linear predictive speech coding).

The basic principle of linear prediction is that the value of an individual sample of a digitised speech signal can be predicted from a weighted combination of previous sample values. Linear predictive coding takes advantage of the redundancy of the speech signal, i.e. within short time periods the signal is relatively stable, it repeats and is predictable. This means that the weighting values only need to be changed about every 10 milliseconds to produce intelligible speech. This results in a significant saving in terms of data as the weighting values are transmitted rather than the digitised speech signal. The analysis or encoding process involves finding a set of weights, normally referred to as coefficients, for each short segment of speech which minimises the difference, known as the error, between the original signal and the predicted signal (see Makhoul 1975 for the mathematical derivation of this approach).

The linear prediction coefficients also have an interpretation in the frequency domain, which is exploited for measuring formants. The coefficients define a digital filter which

represents the filtering effect of the vocal tract. By examining the frequency characteristics of this digital filter it is possible to obtain information about the frequency response of the vocal tract that produced the speech and, most importantly, derive formant frequency measurements.

There are two ways in which formant values can be obtained. The first method involves generating an LPC spectrum, which is essentially the frequency response of the filter defined by the coefficients. The peaks in the LPC spectrum, which correspond to the resonant frequencies of the modelled vocal tract, can either be measured by hand or a peak-picking algorithm can be employed to automatically locate the frequencies of the peaks. The second and most common method employed in readily available speech analysis software is the root solving approach. This method involves mathematically determining the frequency and bandwidth of the individual components, known as the poles, which contribute to the overall frequency response of the filter (see Atal and Hanauer 1971, Makhoul and Wolf 1972, and Markel and Gray 1976 for the theoretical and practical mathematical implementations of these approaches and discussions of the application of linear prediction to speech analysis and formant measurements).

Figure 1.6 shows an LPC spectrum generated for a 50 millisecond analysis frame of the same /i:/ vowel analysed in the previous sections overlaid on an FFT spectrum. The broad peaks in the LPC spectrum corresponding to the formants are clearly visible and their alignment with the less well defined peaks in the FFT spectrum can be seen. The frequency of the peaks can be measured easily by placing a cursor on such a display and reading off the corresponding frequency value.

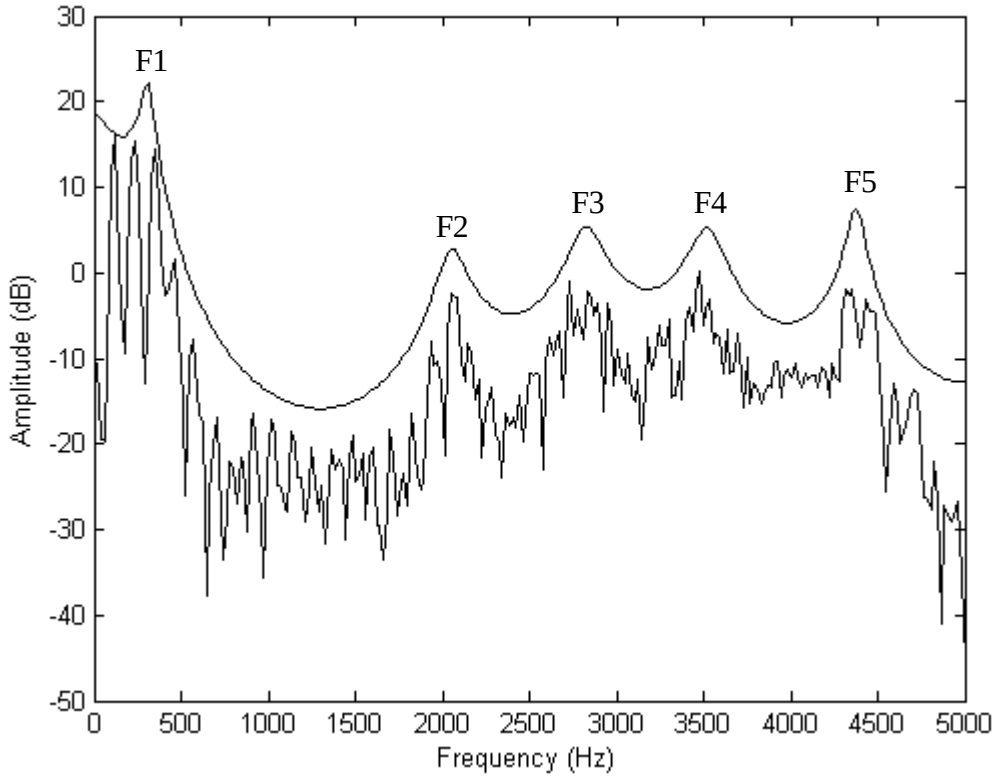

Figure 1.6 LPC spectrum of a 50 ms frame from the vowel /i:/ generated with an LPC order of 12, overlaid on an FFT spectrum with the formants F1 to F5 marked.

The shape of the LPC spectrum is the combined influence of the individual poles that define the digital filter obtained from the LPC analysis. In a typical configuration, most of the poles correspond to resonances in the vocal tract, whilst the remainder contribute to the overall slope and shape of the spectrum. Figure 1.7 shows the frequency responses of the individual poles that make up the overall LPC spectrum in Figure 1.6. The poles corresponding to the formants are labelled from F1 to F5 and are shown with solid lines, whilst the remaining pole, which contributes to the overall shape, is represented by the dashed line.

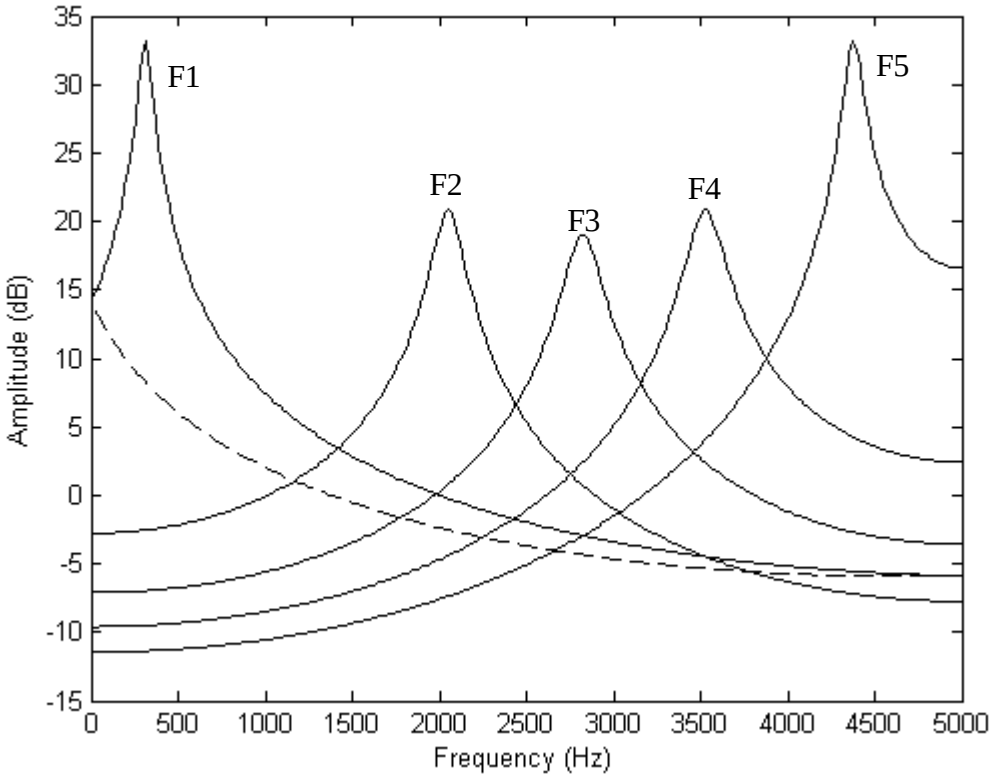

Figure 1.7 Frequency responses of the individual poles that contribute to the LPC spectrum in Figure 1.6. The responses of the poles that relate to the formants are labelled from F1 to F5 and the remaining pole that contributes to the overall spectral shape is shown as a dashed line.

The frequency values corresponding to each of the poles do not need to be measured manually via a cursor as they are automatically derived in the software from the LPC analysis by means of a root-solving algorithm. These values can then easily be logged or displayed. LPC analysis is often conducted over a series of analysis frames and the resulting pole frequencies can be overlaid on a spectrogram, as shown in Figure 1.8. This allows the alignment of the pole frequencies with the spectral peaks to be checked easily. In this particular example the pole frequencies align well with the centres of the formants visible in the spectrogram, suggesting that the values are relatively accurate.

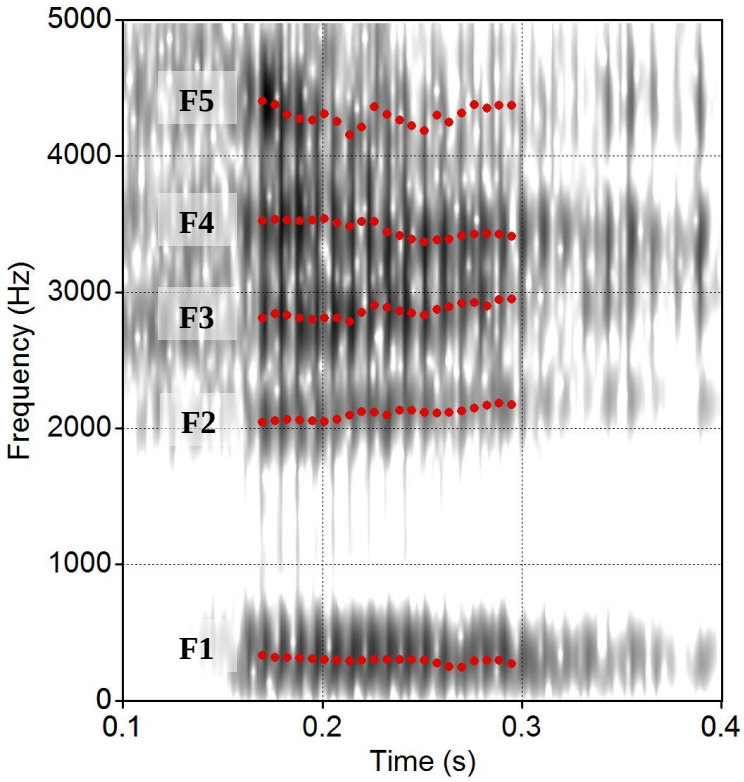

Figure 1.8 Broad-band spectrogram of the /i:/ vowel with overlaid LPC formant values from the Praat software with an LPC order of 10, every 6.25 ms.

Similarly to the spectral approaches discussed earlier, LPC analysis requires a number of parameters to be specified which can significantly alter the outcome of the analysis. The most important of these parameters is the LPC order, which specifies the number of coefficients in the linear prediction model. The number of poles obtained from the LPC coefficients is not fixed and can vary between analysis frames. The number of poles that contribute to the LPC spectrum is usually around half the LPC order. The influence of the LPC order on the LPC spectrum is demonstrated in Figure 1.9, which shows 13 LPC spectra generated for the same section of the /i:/ vowel with the LPC order increasing from 6 to 30 in steps of two. Odd numbered LPC orders can be specified but were not used in this example. The amplitude of each successive LPC spectrum has been reduced by 10 dB so that the structure of each one is visible.

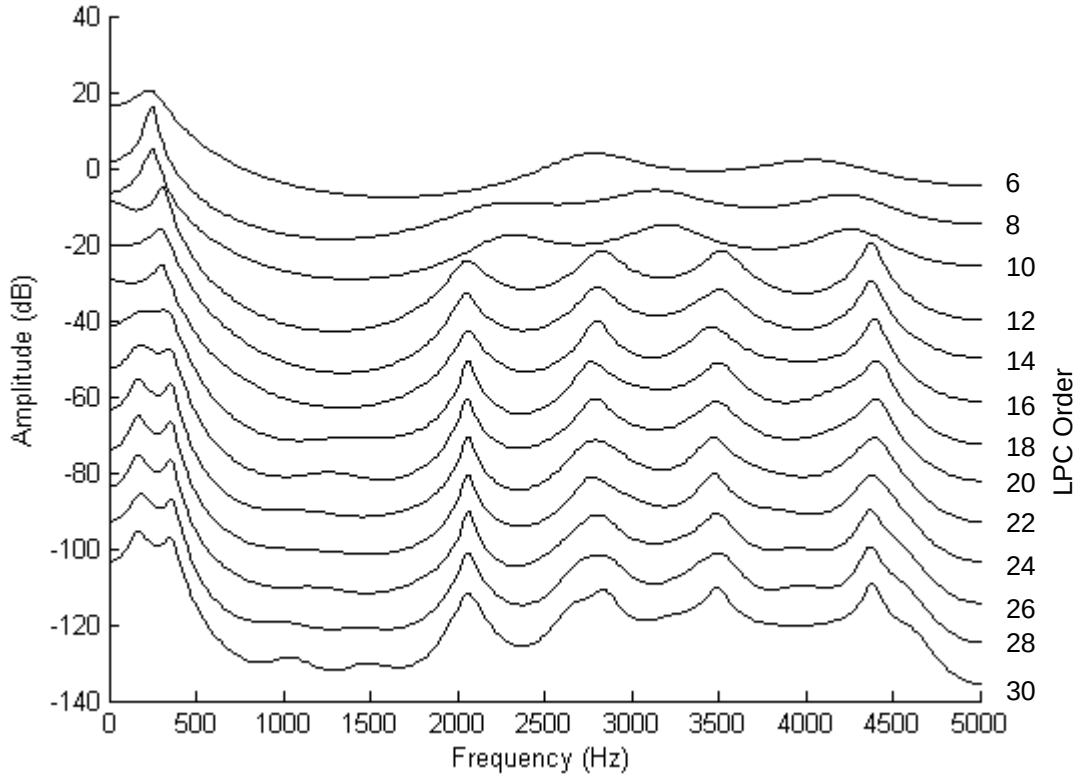

Figure 1.9 LPC spectra of /i:/ vowel with increasing LPC order from 6 to 30 in steps of two. The spectrum from each subsequent LPC order has its amplitude reduced by 10 dB so that the detail of each spectrum can be seen.

The change in the structure of the LPC spectra is apparent as the LPC order increases. At the lowest order of 6 there are only three peaks, whereas the FFT spectra in Figure 1.4 and Figure 1.5 show five peaks. At this low LPC order the model cannot adequately represent the spectrum of the signal. The same is true for LPC orders 8 and 10 which only have four peaks. At LPC order 12, which is also shown in Figure 1.6, the spectrum has five peaks which correspond well with those in the FFT spectra. As the LPC order increases beyond 12 and more poles are influencing the resulting spectra, a second peak appears in the LPC spectra around the region of the true F1 and the peak of F3 becomes increasingly broad. Smaller features and changes in the shape of peaks are also apparent as additional poles associated with the finer detail of the spectrum influence the LPC spectra. If the LPC order is increased enough then the peaks will eventually correspond to the harmonic frequencies of the vocal fold sound source. It is apparent from the figure above that there are a range of LPC orders that produce an acceptable estimate of the spectrum and pole frequencies that correspond to formants in the speech signal. The influence of different LPC orders on resulting formant frequency measurements is a central issue investigated in this thesis.

An important factor that is associated with LPC order is the frequency range over which the analysis is performed. This parameter is often specified in speech analysis software as a 'maximum analysis frequency' or 'maximum formant frequency'. It determines the sample rate of the speech signal that is subject to LPC analysis. In most speech analysis software the speech signal is resampled at a pre-processing stage of the LPC analysis, and the original signal will remain unchanged. The frequency range is important because as it increases one expects, up to a point, to observe more resonance peaks in the spectrum. Therefore the LPC order must be increased so that the model can represent the additional spectral peaks. Figure 1.10 shows the effect on the LPC spectrum of maintaining a constant LPC order of 12 whilst altering the maximum analysis frequency from 2 kHz to 8 kHz in steps of 1 kHz. Again, the amplitudes of successive spectra have been reduced by 10 dB so each one is visible and the maximum analysis frequency is also indicated by a label and a short vertical line.

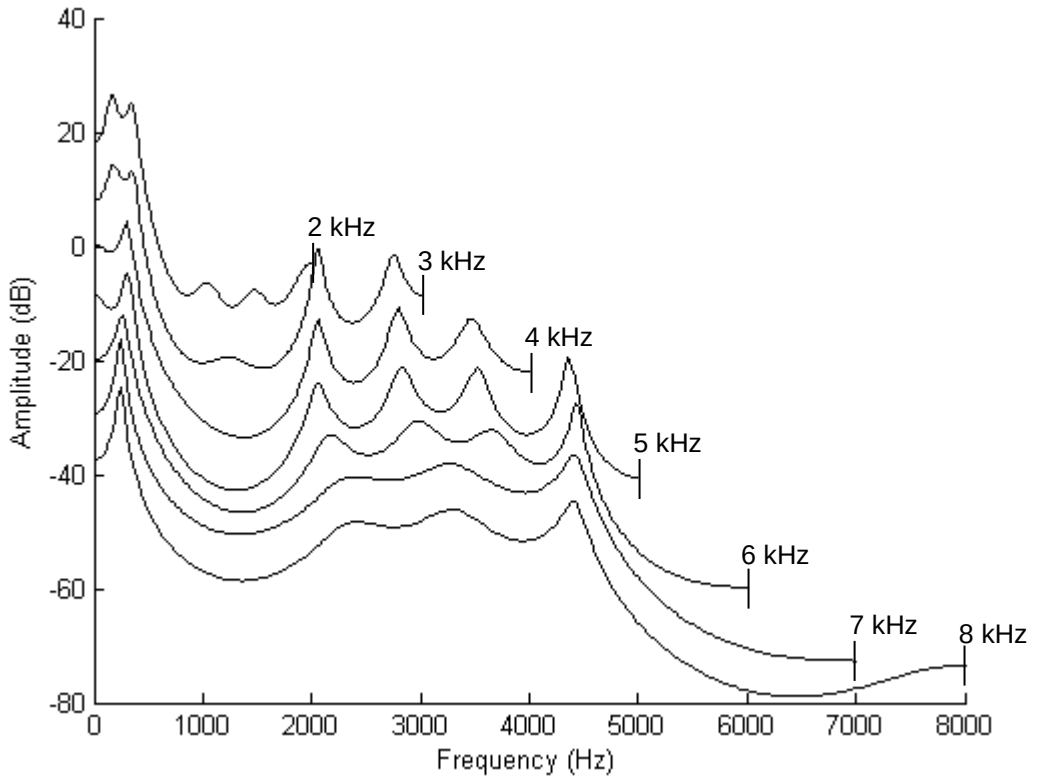

Figure 1.10 LPC spectra of /i:/ generated at LPC order 12 with increasing upper analysis frequency from 2 kHz to 8 kHz in 1 kHz steps. The spectrum from each subsequent upper analysis frequency has its amplitude reduced by 10 dB so that the detail of each spectrum can be seen.

In each spectrum in Figure 1.10 there are either four or five peaks since this is influenced by the LPC order, which remains constant at 12. However, as the maximum analysis frequency changes the location of the LPC spectral peaks and the features

within the spectrum of the speech signal that they are associated with changes. In the speech spectrum there is only one resonant peak corresponding to F1 which occurs below 2 kHz. When the maximum analysis frequency is 2 kHz two peaks are present in the region of F1 and two further peaks occur above 1 kHz where there are no broad peaks in the speech spectrum. At 3 kHz the peaks of the true F2 and F3 are now represented, but F1 still has a double peak and an extra peak lies between F1 and F2. The LPC spectra at the maximum analysis frequencies of 4 kHz and 5 kHz are reasonable estimates of the speech spectrum, but at higher frequencies the fit becomes worse and peaks corresponding to the formants are lost, as there are insufficient poles to model the greater frequency range of the speech signal. Figure 1.10 further demonstrates that in order to obtain a good estimate of the speech spectrum both the LPC order and maximum analysis frequency must be appropriate.

The essentially automated nature of measuring formants using LPC analysis makes it an attractive method, but it must be used with care. The apparent ease with which measurements can be obtained may lead analysts to be overly reliant and trusting of the results without checking or questioning them. However, the likelihood of unreliable measurements being accepted can be reduced through visual comparison of results with spectrograms, and through the application of knowledge about where formants should occur for a particular category of vowel. Many speech analysis programs also employ post-processing techniques to the LPC analysis results in an attempt to provide only the pole frequency values associated with formants, rather than those of spurious peaks or the global spectral shape. Such techniques use information about the nature of formants, such as their tendency to have narrow bandwidths and not vary significantly from one analysis frame to the next. These post processing techniques are normally employed in tools called ‘formant trackers’ as they track the formant values across a number of analysis frames. Again, the results from formant trackers are often overlaid on spectrograms so their accuracy can be assessed. The behaviour of three formant trackers is examined in Chapter 7 and further details of their operation and the analysis parameters available is given in Section 7.2.

LPC formant analysis is also limited by the fact that it is based on a simplified model of speech production which does not account for all aspects of the speech production process, e.g. interactions between the glottal source and the vocal tract or nasalisation of vowels. Whilst the approach may be sufficient for deriving formant measurements for

many vocalic speech sounds, it cannot be expected to perform well when the model cannot adequately represent the speech signal.

1.3 Use of Formants in Forensic Speech Science

Forensic speech science, also known as forensic phonetics, is concerned with answering questions related to speech, usually in recordings, for legal proceedings (an overview of the field is provided by Jessen 2008). The most common question concerns whether speakers on two different recordings could be the same person (forensic speaker comparison). Another common issue is to determine what was said in a recording. Formant measurements are often used to assist in answering both questions. How this is done is considered in detail in Sections 1.3.2 and 1.3.3. In many respects measuring formants for a forensic analysis is not dissimilar to undertaking the task in other areas of phonetics, however, the characteristics of the recordings that are encountered can make it more problematic. The recordings are often of poor quality with a restricted frequency range, they are often of limited duration, the analyst has no control over the circumstances in which they are made and there is normally no opportunity to make a better recording. The measurements and subsequent analysis is generally limited to the first three formants due to the restricted frequency range of the signals and in some instances only the second formant may be measured with any reliability. Some material is so poor that formants cannot be measured at all.

Before discussing how formants are used in forensic speech science, sources of variation in formants are introduced as some types of variation provide the basis for using formants in forensic speech science, whilst others highlight the need for caution when making and interpreting measurements.

1.3.1 Variation of Formants

The primary source of variation in formants is caused by changes in the vocal tract in order to produce different vowel sounds, which tend to result in broadly characteristic formant patterns. However, the relationships between articulatory configurations and formant frequencies are complex (see Fant 1960, Lindblom and Sundberg 1971, and Stevens 1998). The most common simplification relates an articulatory description in terms of the highest point of the tongue in the front-back and low-high dimensions in the oral cavity with the first two formants. The first formant correlates with the articulatory dimension of vowel height, with low or open vowels having high F1 values

and high or close vowels associated with low F1 values. The second formant is related to the front-back dimension of vowels with front vowels tending to have high F2 values and back vowels low F2 values. This approach leads to different vowels being represented by their characteristic F1 and F2 values and these measurements are often displayed on scatter plots with F1 on an inverted vertical axis and F2 on an inverted horizontal axis so that they align with the standard IPA vowel quadrilateral. The fact that characteristic F1 and F2 values are associated with different vowel categories makes them useful for differentiating and identifying different speech sounds and assisting in interpreting the words spoken in a recording.

It has been known for many years that the same vowel produced by different people will have different formant frequencies (Peterson and Barney 1952). This is in part due to physiology, with large differences caused by the variation in vocal tract lengths across men, women and children. However, studies of identical twins, who are assumed to have essentially identical physiology, show that there are differences between pairs of twins (Nolan and Oh 1996, Loakes 2006). Therefore, some of the differences between people are a consequence of learned individual behaviour. It is these differences, both physiological and learned, that makes formants an attractive parameter for helping to assess whether two speech samples could have originated from the same speaker.

However, there are many other sources of variation that can influence formant measurements and make their interpretation less than straightforward. Many forensic tasks, including those discussed below, involve comparing multiple instances of the same speech sound both within and across recordings. Aside from the fact that two productions of the same speech sound will never be identical, factors such as situational differences, speaking style, voice quality and health can all result in variations in formant frequencies in recordings from different occasions. Even instances of the same vowel in different phonetic environments will be influenced by the surrounding sounds resulting in co-articulatory effects that can alter formant frequencies (Hillenbrand et al 2001).

The sources of variation mentioned so far are associated with the speech production process, and they influence the radiated sound pressure waves from the vocal tract. For these speech sounds to be analysed they must be captured by a microphone, perhaps transmitted and then recorded. Each of these processes will to some extent modify,

shape or limit the frequency content of the speech signal that is ultimately stored and reproduced when played back and analysed. Section 2.1.3 discusses several research studies, mainly from the forensic perspective, which investigate the effects of these technical sources of variation.

The final source of variation is the measurement process itself. Section 1.2 highlighted the inherent difficulty in measuring formants, as the spectral peaks must be inferred either by smoothing the spectrum or modelling the speech signal. The differences between narrow-band and broad-band spectra and spectrograms, and the effect of altering the LPC order on the LPC spectra, illustrate how the measurement method and analysis settings can influence the appearance of formants and ultimately their measured values. Decisions must be made by the analyst including which measurement method to employ, what analysis settings to use, where in time the measurements are made, as well as whether to accept or reject the measurements. The measurement process is an interaction between the analysis tools and the analyst, who must use their knowledge and skills to obtain accurate measurements. These sources of variation are discussed in greater detail in Sections 2.1.1 and 2.1.2. Variation in formant measurements caused by altering analysis parameters and the analysis approach is a central issue which is examined in this thesis.

1.3.2 Speaker Comparison

The task of forensic speaker comparison usually involves comparing a recording of the known voice of a suspect with an unknown voice in a recording associated with a crime in order to assess the similarity of the voices and provide an opinion about the potential identity of the criminal. The sources of criminal recordings include CCTV footage of robberies, covert recordings of drug dealers, threatening voicemail messages and telephone calls to the emergency services, and the reference recordings, in the UK, are usually of police interviews. A number of different methodologies are employed around the world with no particular consensus amongst forensic practitioners (Gold and French 2011). The general types of analyses used can be considered as falling into one of three classes - auditory phonetic analysis, acoustic phonetic analysis and analysis by automatic systems. The usual approach of an analyst or laboratory could involve only applying one class of analysis, a combination of two analyses or all three. The most common approach adopted at present by those questioned in the survey reported by

Gold and French (2011) is the combined auditory and acoustic phonetic method (see further Foulkes and French 2012).

The combined auditory and acoustic phonetic approach utilises the componentiality of speech. Numerous aspects of the material, including voice quality, vowel and consonant realisations, pitch and rhythm are typically examined in a given case. Features are analysed in both the known and the criminal recordings and the results are compared. Similarities and differences will always be found across the recordings for the reasons discussed in Section 1.3.1, even when analysing two recordings of the same speaker. The analyst must interpret the results in light of their knowledge and experience in order to arrive at a conclusion.

Within this approach there are generally no rigidly prescribed methodologies that are followed. The features that are examined are often determined by their presence in the material and their relevance, which is determined on a case by case basis. However, the recent survey of forensic speech scientists reported by Gold and French (2011) showed that 97% of the 36 analysts questioned made some form of vowel formant measurements. Data was collected on what aspect of formants were examined. This revealed that 94% examined centre frequencies of monophthongs, 71% examined formant trajectories of diphthongs, 45% considered vowel consonant or consonant vowel transitions, 35% measured formant bandwidths and 13% examined formant densities. Unfortunately no data was gathered on how the measurements were made or how the values were then analysed and compared, but it is clear from these results that formants are used in casework on a very regular basis. Results from the survey showed that vowel formants were considered by practitioners to be the joint second most useful feature for discriminating speakers (along with dialect/accent variants, and after voice quality). It was noted, however, that one respondent did state that they found vowel formant analysis ‘rarely insightful’.

1.3.2.1 Monophthongs

The centre frequencies of front stressed monophthongs are perhaps the most commonly analysed vowels in speaker comparison analysis due to the relative ease with which they can be measured. Front vowels are often easier to measure as their formants are usually better spaced and suffer less from mergers. Stressed vowels are selected as they do not suffer from centralisation and their greater amplitude also makes spectrographic

interpretation easier. The measurements are often made or logged manually either from LPC derived values overlaid on a spectrogram or directly from the spectrogram. Measurements may also be made from smoothed FFT spectra or LPC spectra. Usually measurements will be taken from a single representative analysis frame around the temporal centre of the vowel or averages will be calculated over a number of central frames. The measured values will then be logged, either manually or automatically, and the time location and vowel category will also be recorded. The number of measurements made will often be governed by the quality and duration of the material, especially in the case of the questioned sample.

The subsequent analysis of the measurements may involve calculating mean values for each vowel category and plotting individual values or other representative measures on F1~F2 vowel space plots. There are no simple thresholds or metrics for assessing the results in order to reach a conclusion of identity. The same is true for the other parameters considered as part of a comparison, and analysts must assess the findings in light of their experience and knowledge about sources of variation of the individual parameters which are applicable to a specific case. Whilst mathematical approaches for evaluating the results are available, their wider adoption in the field has not yet occurred. These methods are introduced in Section 1.3.2.4.

1.3.2.2 Diphthongs & Formant Dynamics

The motivation for examining the dynamic behaviour of formants, usually for diphthongs, stems from the proposition that the movement of the articulators and changes of the vocal tract between phonetic targets will encode speaker specific information since a speaker has some freedom in choosing the route between them. Research studies have supported this position (McDougall 2005) and have shown improved identification rates using diphthong trajectories compared with monopthong centre frequencies (Greisbach et al 1995). The measurement process involves obtaining a series of formant values across the duration of a diphthong, which lends itself to the use of LPC analysis, particularly a tool that incorporates formant tracking. The subsequent analysis may simply involve plotting comparable formant contours over time or on an F1~F2 vowel space plot. The comparison methods used in research studies have tended to be objective in order to allow a decision to be made on identity or non-identity so that the performance of the approach can be evaluated and compared with other methods. One approach involves normalising for time and taking five equally

spaced measures along the trajectory, and averaging over multiple instances of the same vowel for each speaker (Greisbach et al 1995). Euclidian distances are then calculated and summed for the five measures between each test speaker and each speaker in the reference set. Identity is assigned to the reference speaker with the smallest summed distance from the test speaker. More complex approaches use polynomials to parameterise the formant trajectories, and the coefficients of the polynomial are then compared using discriminant analysis rather than the measured formant frequencies (McDougall 2005). If mathematical approaches to evaluating the measurements are applied in casework, the results still require interpretation by the analyst in light of the outcomes of other analyses in reaching a final conclusion.

1.3.2.3 Long Term Formant Distributions

Another approach that aims to capture more information from the speech signal than simply measuring centre frequencies of monopthongs is the generation and analysis of long term formant distributions (LTFDs). The approach was initially proposed by Nolan and Grigoras (2005) and involves measuring formants across all vocalic segments of a recording. Distributions are then obtained for each formant. The demarcation of vocalic segments and subsequent measuring of formants may be undertaken manually or the process can be entirely automated (French et al 2012). One advantage of the LTFD method is that it is not necessary to categorise each vocalic segment, and therefore the technique can in principle be used on languages which the analyst is not familiar with.

In Nolan and Grigoras (2005) the distributions from a number of samples were compared visually in order to reach a conclusion on speaker identity when the technique was applied to an old case. More recent developments of the method (Becker et al 2008) employ a GMM-UBM approach from the automatic speaker comparison field (Reynolds et al 2000) in order to compare samples and arrive at a conclusion. The distributions of each formant (F1, F2, F3) from each sample are modelled using Gaussian Mixture Models (GMMs). A Universal Background Model (UBM) is also created from the combined distributions of formant measurements from a reference population of speakers so that the degree of similarity between the criminal and suspect samples can be assessed against a reference population.

The GMM-UBM approach to LTFD analysis is more complex and is usually much harder to undertake than measuring and plotting the centre frequencies of

monophthongs. However, a software package, Vocalise, has recently become available that allows the entire process of formant measurement, model generation and comparison to be undertaken in a single piece of software (Alexander et al 2013). The software is targeted at forensic practitioners who may not have the necessary technical expertise to implement their own system, and it aims to bridge the gap between traditional phonetic methods and newer automatic speaker comparison techniques.

1.3.2.4 Bayesian Approaches

In recent years within the field of forensic speech science, and within forensic science more generally, the issue of how conclusions are expressed has received a significant amount of attention. Until recently, the conclusion of a speaker comparison analysis would be expressed as the likelihood that the suspect was the unknown speaker, e.g. 'The unknown speaker is very likely to be Mr Smith', where the degree of likelihood is selected from a predefined verbal scale. Such approaches are problematic for a number of reasons including the fact that they address the 'ultimate issue', i.e. whether the suspect is guilty or not, they are logically flawed and commit the prosecutor's fallacy, and make no overt acknowledgement of the number of other speakers who may be similar to the unknown speaker. See French and Harrison (2007), Rose and Morrison (2009) and French et al (2010) for further discussions of these issues and the potential solution discussed below, as well as a description of the current framework for expressing conclusions adopted by forensic speech practitioners in the UK.

A preferable way of expressing conclusions, which is both logically and legally correct, is using the Bayesian approach. This involves assessing the likelihood of obtaining the evidence, i.e. the findings, if the suspect was the unknown speaker versus the likelihood of obtaining the evidence if another person was the unknown speaker. The result is often expressed as a single value, known as the likelihood ratio, which is obtained by dividing the first likelihood value by the second (Evett 1998). Given the statistical nature of this approach, it requires numeric data in order to calculate the two probabilities. Consequently, it has been widely adopted in areas such as DNA analysis. It has also become commonplace in the automatic speaker comparison field which uses statistical models to represent speakers and speech samples derived from MFCC (Mel Frequency Cepstral Coefficients) features obtained across entire recordings.

The adoption of a Bayesian framework for expressing results obtained from traditional auditory and acoustic phonetic analyses is problematic because many of the tests are not numeric in nature. For those that are, such as fundamental frequency and formant analyses, one significant problem is a lack of reference data required for determining the typicality of features within the wider population in order to assess the likelihood of obtaining the evidence if a person other than the suspect was the unknown speaker. Satisfactory methods for combining the results from the analyses of the various components of speech, both impressionistic and numeric, in order to arrive at a final conclusion do not currently exist. Verbal likelihood ratios can be arrived at for subjective methods and used to express the final conclusion of an analysis but this practise has not yet been widely adopted.

Notwithstanding these issues, methods have been and continue to be developed to allow a numerical Bayesian analysis of phonetic features. One of the first of these involved the application of the Multi-Variate Kernel Density (MVKD) approach which was initially developed for the comparison of glass fragments by comparing their refractive indices (Aitken and Lucy 2004). This approach has been successfully applied to various types of formant data and fundamental frequency measurements (e.g. Rose 2002 and Morrison 2011). The GMM-UBM approach, mentioned above, has also been used with long term formant distributions and formant dynamics (Morrison 2011). It is likely that in the future these numerical approaches will become more commonly used by forensic practitioners as the use of simpler tools becomes more widespread, more reference data becomes available and as methods need to be validated (see Section 1.3.5). Formant measurements lend themselves to these kinds of approach as they are relatively easy to measure, a large amount of data can be collected from a speech sample and they can be interpreted in terms of their articulatory and phonetic origins.

1.3.3 Content Determination – Transcription & Disputed Content

The second most common type of work that is undertaken by forensic speech scientists is determining spoken content in recordings. Most frequently this involves preparing a verbatim transcript of what can be heard. Often the recordings that are submitted are of poor quality otherwise they would be transcribed by typists or police officers. The material is usually replayed from waveform editing software to allow precise control over the replay and repetition of material during the preparation of the transcript. Occasional use may be made of acoustic analysis tools, particularly spectrograms, but

the transcriber will generally rely on their own skill, and knowledge of speech and language processes in order to determine what was said.

A subtask of content determination is termed questioned or disputed content analysis. This type of analysis is requested when the content of an utterance has the potential to have significant evidential value. Either two competing interpretations will have been offered for an utterance, often an incriminating one by the prosecution and a benign one by the defence, or a single interpretation exists and it requires confirmation due to its potential significance. For the analysis to be requested there is usually some ambiguity in the interpretation of the utterance. It is not sufficient for a forensic analyst to simply state that they prefer or agree with a certain interpretation without providing a justification based on a comprehensive examination of the material. This will involve a detailed analysis of the utterance in question, both auditorily and acoustically, and a comparison with non-disputed speech from the same speaker, either within the same recording or in other reference recordings. In the ideal circumstances unambiguous instances of the competing interpretations would exist for analysis and comparison. If these are not available then other realisations of the same vowels or consonant segments and transitions from the two interpretations are examined.

Formant measurements from vowels, vowel-consonant and consonant-vowel transitions are some of the most common features considered in questioned content analysis. The determination of vowel phonemes relies on the comparison of measured vowel centre frequencies and diphthong trajectories, whereas the interpretation of consonants is assisted by visual examination of formant transitions into and out of vowels. Since the number of tokens to be measured is often small, careful measurements can be made using multiple methods if necessary. The measurements from the utterance in question will then be compared either numerically or visually via a plot with the non-disputed reference tokens. Further discussion of the approach together with examples from cases in which formant measurements were used to resolve issues of disputed content are presented in French (1990) and French and Harrison (2006 p. 259-260).

The conclusion concerning the interpretation of the disputed utterance will normally be expressed verbally. In cases where the measurements from the reference material for two interpretations form two distinct distributions and the measurements from the questioned utterance falls within one distribution and not the other, the conclusion may

be expressed in categorical terms. Where there is sufficient doubt in the interpretation of the measurements and the auditory analysis the analyst may decide to offer no opinion. However, the measurements made to assist in reaching a conclusion do lend themselves to evaluation within a Bayesian framework which reduces reliance on the subjective interpretation by the analyst. Morrison et al (2014) describe a case in which formant values and voice onset time (VOT) were used in the resolution of a disputed utterance using the normal methods described above. The article also presents an analysis from a Bayesian perspective in which different statistical approaches are adopted in order to address the question using an entirely numerical and statistical approach.

1.3.4 Voice Line-ups

A further area of work in which forensic analysts are occasionally asked to assist is in the creation or assessment of voice line-ups, which are the auditory equivalent of a visual identification parade. A victim or witness may hear a criminal but not see them, and claim that they could identify the voice of the criminal if they heard them speak again. A set of guidelines were drawn up in the UK (Home Office 2003) to assist in the construction of voice parades to ensure that they are carried out in a fair and appropriate way. One of the requirements is that a forensic phonetician assesses the samples of the foils' speech against that of the suspect to ensure that they are sufficiently similar. However, no formal method is suggested for comparing and assessing the samples. Methods under development involve assessing the similarity judgements of lay listeners using multidimensional scaling techniques (McDougall 2013) and comparing those results with acoustic measurements of the voices, including formants (McDougall 2011). It is envisaged that a quantitative method will be developed for assessing the similarity of the voices of foils to ensure that they are perceptually similar enough to the voice of the suspect to ensure a fair line-up. The outcomes of this work will also have relevance to speaker comparison analysis since the fundamental issue that the research addresses is quantifying the similarity of voices.

1.3.5 The Increasing Use of Formants

There is little doubt that over the past 15 years the use of formant measurements within the forensic field has increased. This can probably be attributed to a number of factors including the greater availability of software analysis tools, an increase in forensic research concerning formants and the broad acceptance of the combined auditory-acoustic method for speaker comparison. Their use in the UK has also been influenced

by external judicial factors and will continue to be effected by regulatory developments within forensic science more generally, both of which are discussed below.

In 2002 an appeal was heard in the Northern Ireland Criminal Court of Appeal concerning the conviction of Anthony O'Doherty in 1997 (R v O'Doherty [2002]). The conclusion from a speaker comparison analysis was an important piece of prosecution evidence at the original trial but it was based on limited acoustic analysis and no formant analysis. The appeal heard from several experts and the general view was that acoustic analysis was an important component of speaker comparison examinations. The conviction was quashed and in their ruling the appeal judges stated that:

> 'in the present state of scientific knowledge no prosecution should be brought in Northern Ireland in which one of the planks is voice identification given by an expert which is solely confined to auditory analysis. There should also be expert evidence of acoustic analysis … which includes formant analysis.'

The ruling is binding on the criminal courts within Northern Ireland but not in England and Wales. However, the position adopted in the ruling would be seen as persuasive in the courts in England and Wales and would be difficult to argue against. This ruling has resulted in a marked increase in the use of formants in speaker comparison cases across the UK.

More recently, in 2008, the England and Wales Court of Appeal (Criminal Division) heard an appeal in the case of Flynn and St John which also involved speaker comparison evidence (R v Flynn & Anor [2008]). Whilst the case primarily concerned the identification of voices by police officers, rather than by forensic analysts, the appeal judges stated:

> 'we think it neither possible nor desirable to go as far as the Northern Ireland Court of Criminal Appeal in O'Doherty which ruled that auditory analysis evidence given by experts in this field was inadmissible unless supported by expert evidence of acoustic analysis'.

This ruling makes the situation in England and Wales less prescriptive than in Northern Ireland, but given the extensive use of formants and acoustic analysis by experts demonstrated in the survey reported by Gold and French (2011) it still remains difficult for a practitioner to argue against their use.

Recently, forensic science as a whole has come under greater scrutiny as the validity and reliability of many of its disciplines are questioned and compared with the gold standard of DNA analysis (National Research Council 2009, Law Commission 2011). Unlike DNA analysis, many branches of forensic science have developed around expert opinion and interpretation without the competence of individual analysts or the methods having been rigorously tested under forensically realistic conditions. This position is changing both in the UK and abroad with the introduction of regulation and accreditation to international standards. The standard that is being applied in the UK and elsewhere is ISO 17025 titled 'General requirements for the competence of testing and calibration laboratories' (International Organisation for Standards, 2005). The recently appointed Forensic Science Regulator for England and Wales has also produced a Codes of Practice document (Forensic Science Regulator, 2011) which sets out how ISO 17025 should be applied to forensic science in England and Wales. One of the key requirements is that methods that are regularly used for casework are validated, i.e. it must be demonstrated that a method, technique or process is capable of achieving what it claims to. In the case of interpretive methods, such as speaker comparison analysis, a critical aspect of the validation will involve the competency testing of experts since, the interpretation of results and formulation of conclusions is inextricably linked to their individual training, skills and knowledge. Notwithstanding this, the measurement tools and individual analysis methods that provide the results on which a conclusion is based will also need to be validated. This means that formant measurement tools and the measurement and analysis approaches discussed above will be subject to validation. Although this has not yet taken place it is likely that the validation of formant measurement methods will require their accuracy, behaviour and limitations to be determined under different circumstances. The accreditation also requires the creation of standard operating procedures for the different methods and tools. Information within these procedures will be based on the outcome of the validation testing and will no doubt contain advice and guidance on the use of specific measurement tools and methods. It is envisaged that the experiments and results described in this thesis and the resulting guidance will be beneficial to analysts designing and implementing validation testing and writing standard operating procedures. Whilst the experiments have not been specifically designed as validation tests, they could be modified and extended relatively easily to fulfil the requirements for validation testing.

The use of formant based methods is likely to continue to increase in the future for a number of reasons. Their numeric nature means they can be tested and their performance can be assessed more easily than subjective methods that often require a considerable expenditure of time. They are also less subjective than many of the other techniques employed by forensic speech scientists which should make the results from different practitioners more consistent. The ability to automate analyses allows greater amounts of reference data to be obtained which will facilitate the presentation of results in a Bayesian framework. These reasons are also cited as some of the benefits of using automatic speaker comparison systems. One advantage that formants based methods have over automatic systems is that formant measurements can be readily understood and interpreted from a phonetic perspective, unlike the MFCCs used in automatic systems which are opaque.

1.4 Summary

This chapter has introduced formants by means of the source-filter model of speech production and presented three commonly used methods for measuring them. The final section discussed how formants are used in the field of forensic speech science. The following chapter builds on this introduction by examining the literature concerning formant measurement errors before the research aims and questions for this thesis are presented.

Chapter 2 Literature Review

This chapter presents an overview of research on formant measurement accuracy and the limited advice concerning their measurement. Together with the introductory material in the previous chapter, this provides the background to the research described in this thesis. Following this, the motivation and goals for the present research are presented with a formal statement of the research questions.

2.1 Formant Measurement Accuracy

2.1.1 Measurement Method

The following section considers a number of studies where the main focus is on the underlying performance of the measurement method.

One of the earliest studies concerning formant measurement accuracy examined the errors in formant measurements made from wide-band spectrograms and wide-band spectral sections (Lindblom, 1960). The study involved five subjects determining the frequency of the first three formants for six synthetic vowels created with six different fundamental frequencies. The study also examined the effect of altering the bandwidth of the analysis filter for both the spectral sections and the spectrograms. Table 2.1 shows the mean absolute error values across all the vowels and fundamental frequencies.

	Spectral Section		**Spectrogram**	
Analysis filter width (Hz)	**45**	**300**	**300**	**600**
Mean absolute error (Hz)	40	55	50	90
Spread (Hz)	-	60	70	150
Maximum error (Hz)	90	170	150	250

Table 2.1 Absolute formant measurement errors from spectral sections and spectrograms averaged over six synthetic vowels at six fundamental frequencies measured by five subjects. Adapted from Lindblom (1960 Table I-1).

The results show differences in performance both across the two methods and for the different filter widths. The study also presents plots for the errors across the range of fundamental frequencies, which show a poorer performance for the higher frequencies (up to 350 Hz). A number of factors are noted that influence the measurements, including the pre-emphasis filter, the location of the harmonics within the formant, and the relation between the width of the filter and the fundamental frequency. Whilst the results are not broken down by vowel, it is reported that ‘subjects consistently located

low formants much too high, e.g. F1 of [i:] at about 400 Hz instead of 240 Hz' (1960, p. 5). The conclusion to be drawn from this evidence is that measurement of formant frequencies on a spectrogram requires some degree of interpretation by the analyst. Therefore, the accuracy of the measurement method cannot be completely isolated from the skill and abilities of the analyst. Whilst this issue is not addressed in this specific study, a further report involving the same author (Lindblom et al 1960) does address this issue and it is considered in Section 2.1.2 where similar studies are also discussed.

A further work by Lindblom (1961) summarises his earlier findings (Lindblom 1960, Linblom et al 1960) and presents several observations relating to the sources of errors encountered in the analysis of vowels. A clear statement is made at the very start of the report which both encompasses the aim of measuring formants and some inherent limitations:

> When we measure the formant frequencies of a vowel we always aim at estimating the pole frequencies. Unless our measurements stand for poles they have no theoretical justification. This has sometimes been overlooked since neither the pole nor always the corresponding envelope peak have any direct spectrographic manifestations. (1961, p. 3)

This is followed by the important observation that some of the sources of error are inherent in the speech signal whilst others are a consequence of the analysis tool, in this case spectrography. Several sources of error are listed:

1. The higher the fundamental frequency the less information on the (spectral) envelope shape.
2. The asymmetry of a formant (in terms of the relative location of harmonics) may considerably increase the difficulties in formant frequency estimations.
3. In close vowels only the upper slope may be visible in the first formants.
4. The first two formants of back vowels are often badly defined since they are usually close together.
5. Close back vowels have only a slight amount of energy in the upper formants. Considerable high-frequency pre-emphasis may be needed to make them appear.
6. Zeros often interfere with the F-pattern and make accurate judgements difficult.
7. In non-stationary intervals the time position of the sample must be chosen more or less arbitrarily. (1961, p. 3)

The results from the earlier study (Lindblom 1960) are also revisited, and the magnitude of the errors is contextualised as often being larger than the difference limen for formant frequencies, which is given as approximately 3%. It is also observed that:

> The magnitude of these errors is to some extent dependent on the inter-relations between pole frequency and fundamental frequency, i.e. the further the strongest harmonic within a formant from the envelope peak, the larger the error (1961, p. 4)

A somewhat pessimistic statement is made that:

> the prospect of finding a formula that will be of general application and automatically give us the frequency of the pole are highly unfavourable (1961, p.4)

Following the widespread adoption of linear prediction as a method of speech analysis in the late 1960s and early 1970s, which achieves precisely what Lindblom suggested might not be possible, a number of studies examined the accuracy of the technique for measuring formants. Chandra and Lin (1974) compared the performance of the autocorrelation (stationary) and covariance (non-stationary) methods of determining the linear prediction model parameters. They analysed both synthetic and real speech, and examined the effects of LP order, the duration of the analysis segment and its location relative to pitch periods.

For synthetic speech they found that the covariance method produced almost perfect formant estimates when the duration of the analysis segment was less than one pitch period. In these conditions the autocorrelation method produced errors that were larger. As the duration of the segment increased past one pitch period the magnitude of the errors for both methods increased but then stabilised as the duration increased further. At the longer segment durations, greater than two pitch periods, the results from both methods were generally equivalent. For the real speech the findings were the same i.e. that for both methods as the analysis segment duration increased the formant estimates stabilised. They also found that for the autocorrelation method, when the segment duration was longer that a pitch period the particular windowing function that was applied to the segment influenced the results. They found that the application of a Hamming window produced a more accurate spectrum than when no window was applied. For the covariance method when the segment duration was less than a pitch period the precise alignment of the segment affected the results, with the best

performance when the glottis was closed. However, they note that ‘it is not always possible to isolate open and closed glottis conditions in real speech’ (1974 p.413).

The study also examined the influence of the LP order on the normalised minimum total square error, a measure of the similarity between the original signal and the signal predicted by the LP model. Their results showed that for both the real and synthetic speech the normalised error reduced as the LP order increased. They also investigated the influence of segment duration and found that for segment lengths longer than a pitch period the stationary and non-stationary formulations resulted in very similar errors especially for the synthetic speech. However, for the real speech, the covariance method resulted in much smaller errors for the all segment durations.

Both Chandra and Lin (1974) and Markel and Gray (1976, p. 187), in their summary of the former’s work, make comment that the advantage of testing synthetic speech is that the parameters are known, so an objective measure of performance can be obtained. But they also warn that for the results to be meaningful the synthesised speech must closely resemble real speech.

Another study, which is also summarised by Markel and Gray (1976, p. 188-189), examines the influence of voice periodicity on the accuracy of formant measurements (Atal and Schroeder, 1974). Again, synthesised speech is used and on this occasion the fundamental frequency is varied, as well as the formant frequency. The results are presented for signals with only one formant with a range between 200 and 700 Hz, with fundamental frequencies of 100, 200 and 400 Hz. The maximum errors for the three F0 conditions are 11, 30 and 67 Hz respectively. Like Lindblom (1961), the maximum errors are compared with difference limen of 3 to 5% (Flanagan 1972), and for 200 and 700 Hz they fall above them. The errors are not constant, but vary as the formant frequency changes. They oscillate around the true formant value and pass through zero when the true formant value is a multiple of the fundamental frequency. The authors report that similar results were found when synthetic speech was generated with two or more formants.

The more recent work by Vallabha and Tuller (2002) considers four sources of systematic errors in the LPC analysis of formants. They again predominantly use synthetic speech and note that since the ‘synthesiser satisfied all the assumptions of LPC analysis’ it ‘constituted a best-case scenario for the analysis method’. They begin

by examining the effect of fundamental frequency on formant measurement error. The problem is referred to as ‘F0 quantisation’, since it has already been shown that formant estimates tend towards the frequency of F0 harmonics. They summarise their findings as showing that the error increases linearly with fundamental frequency and that they oscillate more rapidly for the higher formants ‘because the small changes in F0 accumulate and cause larger shifts in the higher harmonics of F0’. They note that ‘if F0 is varying within a small range (e.g. F0 ± 10 Hz), the F1 estimate will vary slowly and, because F1 bandwidth is usually small, the error range will be quite large. For the same F0 fluctuation, the F2 estimate will fluctuate rapidly but because of the large bandwidth, the error range will be smaller than for F1’.

They next consider the errors due to the selection of the incorrect LPC order. They suggest that the usual rules of thumb for selecting the LPC order (twice the number of expected formants plus two or the sampling frequency in kilohertz) ‘ignore systematic between-speaker or between-vowel differences’. They propose and investigate a heuristic for determining the optimum order based on reflection coefficients, which can be derived from the LP coefficients and are equivalent to the acoustic reflection coefficients of an acoustic tube model of the vocal tract. The effects of altering the LPC order are investigated for five repetitions of two vowels by two speakers and the heuristic is shown to select more appropriate LPC orders than the rules of thumb.

The third source of systematic errors concerns the relationship between the frequency of the poles generated by the LP analysis and their equivalent spectral peaks. By manipulating the frequency and bandwidth of a pole, it is clearly demonstrated that the greater the bandwidth of the pole and the closer it is in frequency to another pole, the greater the divergence between the pole’s frequency and the spectral peak. The fourth source of errors relates to the alternative method of obtaining the formant estimates from the LPC coefficients, peak peaking, and specifically the use of interpolation to obtain more accurate estimates. The analysis found that estimates were ‘biased toward the nearest harmonic’ and that the errors were higher for formants with smaller bandwidths. To reduce the errors it is suggested that the length of the DFT is increased, which is used to obtain the spectrum from which the peaks are located.

The overall findings are summarised by way of a number of recommendations. These are:

1. The order of the LP filter should be matched to the utterance being analysed whenever possible. If this is not feasible, then the order of the filter should at least be matched to each speaker.
2. Root-solving should be used with caution for low formants or when formants are close to each other. In the latter situation, root-solving is best used to detect the existence of multiple formants. The locations of the roots bracket the locations of the spectral peaks and can thus guide the peak-picking algorithm.
3. When estimating the locations of the spectral peaks, the length of the DFT should be at least 512 (with parabolic interpolation) or 2048 (without interpolation). (Vallabha and Tuller 2002, p.156-157)

The magnitude of the errors encountered is again placed in context by comparing them to difference limen for trained listeners obtained by Kewley-Port and Watson (1994). For steady-state synthetic vowels they describe the thresholds as being relatively constant at around 14 Hz for frequencies less than 800 Hz, and increasing linearly above this frequency with a resolution of about 1.5%. The magnitude of formant errors encountered are summarised as being between 15 to 60 Hz, leading to the conclusion that the perceptual quality of resynthesized vowels may well be altered. They also note that when analysing real speech, with a fluctuating fundamental frequency, 'averaging of formant estimates over adjacent analysis frames can be effective in reduction the F0 quantization'. (2002, p. 158) A further comment relates to the analysis of diphthongs and the suggestion is made that, based on their experience, a single LPC order is sufficient for a given token, provided that the order is matched to the speaker.

A further study by Vallabha and Tuller (2004) expanded the testing of their heuristic for determining an optimum LPC order. The heuristic was applied to a relatively broad range of vowels for three speakers, two male and one female. For each speaker a different range of optimum values was established. To examine the effect of sampling frequency the similar speech material was recorded for a 13 year old male speaker at a sampling rate of 20 kHz. This was then down-sampled to 15 kHz and 10 kHz. Application of the heuristic again showed a range of optimum orders across the vowels and the ranges changed for the different sample rates. The higher orders were selected for the higher sampling rate and lower orders for the lower rates.

2.1.2 Analyst Variability

The studies summarised above focus on the accuracy of the analysis method. However, the analyst is an integral part of the formant analysis process, even if this is limited to the selection of the LPC order when conducted an LPC analysis. A number of studies have therefore examined the performance of analysts when measuring formants from spectrograms and some have also compared them with LPC analysis. These studies are discussed in this section.

In Lindblom et al. (1960) a study is reported that investigated the variability in the repeated measurements of vowel formants from spectrograms by five analysts. Real speech was examined, rather than synthetic, and the first four formants were measured from wide-band spectrogram for a total of six vowels, with an F0 of around 120 Hz. The average deviation from the mean reported for one analyst was for a spread of 10 to 30 Hz, averaged over the four formants. It was found that 'the systematic disagreement between subjects was maximally 130 Hz with an average values of the order of 50 Hz' (1960, p. 12). Karlsson (1975) conducted a similar study but concentrated on the vowels of female speakers. Eight analysts measured the formants of five synthetic vowels, again using a wide-band spectrogram. The average deviation was found to be 31 Hz with a standard deviation of 32 Hz, across all 4 formants, with the errors for F3 tending to be on average higher at 41 Hz. A slight increase in the errors with pitch was also observed. Comparison of the errors across the eight analysts revealed a spread of mean errors from 36 Hz below the true value to 26 Hz above it. Four of the eight analysts also performed measurements on 20 natural vowels spoken by 10 female speakers. Comparison of the variation of measurements, rather than their accuracy, revealed different patterns of deviation from those found for the synthetic speech. Also, the deviations were greater for the real speech. Overall, the results are comparable with those reported in Lindblom et al (1960) for male speakers.

Monsen and Engebretson (1983) report on a comparative study between the accuracy of spectrographic measurements by three experienced analysts, and the equivalent results from an LP analysis. The speech material comprised 90 synthesised tokens with a range of fundamental frequencies, from 100 to 500 Hz, with formants that represented a range of different vowel qualities. The bandwidths of all instances of one vowel were also varied from 50 to 400 Hz. The spectrographic measurements were made from wide-band representations that were accompanied by narrow-band sections from the centres

of the vowels. The LP analysis was conducted using the Speech Microscope (Vemula et al., 1979) set of computer programs. The LP order used was 22, which seems surprisingly high. No information is provided about the sampling frequency used so it is difficult to assess its suitability. However, it was selected on the basis of a short pilot experiment that tested a range of orders from 12 to 30, which showed that below order 20 the absolute error increased, whilst above 20 it remained relatively constant.

A detailed analysis of the results is presented, with a separation made between those obtained for tokens with a fundamental frequency between 100 and 300 Hz and those from the higher range of 350 to 500 Hz. This was done as the performance between the two sets was markedly different. Table 2.2 shows the mean absolute error (MAE) values over all tokens for each formant for both analysis methods. For F1 and F2 both the LP and spectrographic methods produced similar results within each of the F0 ranges. However, for F3 the performance of the spectrographic method is much worse in both. For vowels where formants were closely spaced, the performance for the LP analysis decreased, whereas no such pattern was seen for the spectrographic analysis. For both methods, the performance did decrease for increasing formant bandwidth.

F0 Range	Method	F1 MAE (Hz)	F2 MAE (Hz)	F3 MAE (Hz)
100 - 300 Hz	LP	69	57	50
	Spectrographic	70	40	111
350 - 500 Hz	LP	143	105	111
	Spectrographic	143	123	174

Table 2.2 Combined mean absolute error values for F1, F2 and F3 from LP and spectrographic measurements from 90 synthetic vowel tokens separated by two F0 ranges. The measurements were made by 3 analysts (adapted from Monsen and Engebretson, (1983), Tables 4 and 5).

The variability of the performance across the analysts was also considered, with their absolute error across the three formants in the 100 to 300 Hz F0 range reported as 64, 79 and 79 Hz. This increased to 141, 166 and 133 Hz in the 350 to 500 Hz F0 range.

The study also demonstrates the change in performance of the LP analysis for a different synthesis method. The findings reported above were from a parallel synthesis approach, which was chosen as it was considered to be more representative of real speech than serial synthesis. The LP analysis was re-run on speech generated with the same formant values but using serial synthesis. The results were reported for the 100 to 300 Hz F0 range as 31 Hz, 40 Hz and 26 Hz for F1 to F3 respectively, which shows a

marked improvement compared with the results from the parallel approach of 69 Hz, 57 Hz and 50 Hz seen in Table 2.2.

A study by Wood (1989) compared formant measurements from spectrograms and LP analysis for real speech. The speech material consisted of 60 words spoken in Bulgarian. Whilst it is acknowledged that the true formant values cannot be known, the findings are assessed in light of those from Monsen and Engebretson, (1983). The results showed that F1 for the LP analysis, on average, produced results that were 34 Hz higher than the spectrographic measurements for stressed vowels and 26 Hz higher for unstressed vowels. Since Mosen and Engebretson (1983) reported that the spectrographic analysis underestimated the true frequency by about 10%, it is presumed that these results show that the LP analysis 'underestimated F1 in the Bulgarian vowels by about 5%'. The F2 results show the LP analysis to be on average 9 Hz lower for stressed vowels and 21 Hz lower for unstressed vowels. The results from Mosen and Engebretson (1983) show a 3% overestimate of F2 values with spectrographic analysis, so again, the LP is analysis is assumed to be closer to the true value. No spectrographic measurements were made for F3 as it was not well defined for a number of vowels, especially the unstressed one. However, the LP analysis did produce estimates for nearly all vowels, which, again based on the performance of LP in the Mosen and Engebreston (1983) tests, are assumed to be relatively accurate.

In the studies that have been summarised so far, the effects of analyst variability have only been considered for spectrographic analysis. Where LPC analysis has been used, it has generally been applied in a systematic and controlled way. However, this does not necessarily reflect the real world usage of LPC analysis tools when being used in an interactive way i.e. where decisions must be made concerning where in time a measurement should be made and what analysis parameters should be used. The study by Duckworth et al. (2011) investigates this issue by comparing the formant measurements from three analysts for real speech material both before and after agreeing a common measurement procedure. The speech material consisted of six repetitions of six monophthongs from a total of 40 male speakers, separated in to two sets with 20 speakers per set. The measurements were made using the Praat software for the initial set after agreeing some very general principles, but the analysts were free to choose the measurement method, either LPC, spectrogram or spectral slice. Following the first set of measurements, the analysts agreed a common strategy which restricted

them to LPC analysis only and a clearly defined method for locating the point in time at which to take the measurements. These procedures were then applied to the measurements made for the second set of speakers. The measurements were compared across the two speaker sets in a pair-wise way between the three analysts for each vowel category. As predicted, after agreeing a common strategy, the measurements from the second set showed less between-analyst variation. Overall, the variation for F3 was the greatest in both sets, whereas the most consistent results were for F2 in the second set. The measurements for individual speakers were examined and it was found that for many of them the three analysts produced very similar formant estimates. However, a number of speakers showed very large differences, which contributed to the overall variation. Whilst the location in time of the measurements and choice of LPC order was not analysed, the study does make recommendations that such information should be retained. This is particularly relevant in the forensic context where the close scrutiny of formant measurements may occur if the results from different analysts are divergent.

2.1.3 Technical Characteristics of the Speech Signal

The research considered in the previous sections mainly concerns the effects of the measurement process, including the analyst, on the accuracy or variability of formant measurements. Another factor which has been shown to affect formant measurements is the technical characteristics of the speech signal. The studies summarised below all focus on this issue.

One of the early works that considers the impact of the technical characteristics of the speech signal is the study by Künzel (2001), which examines the 'telephone effect'. Since telephone channels act as filters, with a pass-band measured in this study from approximately 400 to 3400 Hz, it was hypothesised that the reduction of speech energy at the lower frequencies would result in an artificial upshift in formants. Ten male and ten female speakers were recorded reading 'The North Wind and the Sun' in German in to a standard digital telephone handset whilst being simultaneously recorded via a microphone at the near end of the line and at the far end of the line. The F1 and F2 values were measured in both recordings for the 29 vowel tokens. The measurements were made from spectrograms using the KAY Multi-Speech Software. Attempts were made to use the LPC formant tracking function of the software but this was found to be unreliable up to 50 per cent of the time, so it was not used. Comparison of the two conditions found that for F1 the difference between them was significant, with the

values always being higher in the telephone recordings. The results for F2 overall showed no significant differences, confirming the hypothesis. Examining the results from the different vowels revealed that those with the lower F1 values tended to have the greatest differences. When the results were examined for the individual speakers, a range of variation was found, meaning that it was not possible to apply a general rule to compensate for the upshift in F1 values.

A study by Byrne and Foulkes (2004) replicated the work of Künzel (2001), but used a mobile phone at the speaker's end of the line rather than a landline. A standard text was read by six male and six female speakers whilst being simultaneously recorded at both ends of the phone line. Measurements for F1 to F3 were made via LPC spectra in the Sensimetrics SpeechStation2 software using the narrowest bandwidth setting available. Some problems were reported where the first and second formants were too close to resolve and where F3 in the mobile phone recordings could not be located at all. Only F1 showed a significant difference between the two conditions, with an average upward shift of 29% in the mobile phone recordings compared with the microphone recordings. Again, the upward shift was greatest for vowels with the lowest F1 values. The F2 values showed little change across the conditions as did the majority of F3 values. However, the highest F3 values in the direct recordings did show a large downward shift in the mobile phone recordings. Variability was found across tokens and speakers precluding a compensatory algorithm.

In order to examine in greater detail the effects on formant measurements of the GSM AMR codec used in mobile telephones, Guillemin and Watson (2006, 2008) conducted a controlled study in which recorded speech was processed via the codec at different bit rates. In order to remove other potential variables that could be introduced by using a real telephone network, the speech was encoded and decoded within a computer. Formant measurements were made on the original unprocessed and the GSM processed speech using WaveSurfer with default settings, including an LPC order of 12. The preliminary findings suggest that the overall tendency is for formant frequencies across all three formants to decrease in the processed versions. A difference is seen between the male and female speakers, with the changes for female speakers being greater. They also report that the behaviour is unpredictable and that no patterns emerge across the different bit rates tested. However, a similar study by Enzinger (2010), reached different conclusions. The same approach of applying the GSM codec to recorded speech at

different bit rates within a computer was used and in this study the telephone network band-pass filtering characteristics were also simulated. This study again found the raising of F1, caused by the band-pass filtering and concluded that the effects of the GSM codec were small relative to the filtering effect. A small tendency was found for F3 to be lower in the encoded signals. The study concludes that the codec does cause problems for the automatic tracking of formants resulting in the incorrect assignment of formants and in some cases missing them completely, but after applying manual corrections the differences were 'rather small'.

It is likely that a significant proportion of the differences seen by Guillemin and Watson (2006, 2008) are a consequence of tracking errors rather than the codec significantly altering the position of the formants. The default settings of WaveSurfer would have restricted the bandwidth of the unprocessed signal to 5 kHz, whilst the codec required the speech to be down sampled, resulting in an upper frequency of 4 kHz. Furthermore, the highest two bit rates restrict the upper frequency to approximately 3.6 kHz, the middle range bit rates limit it to approximately 3.4 kHz, whilst the lowest two show upper limits of 2.8 kHz and 3 kHz respectively. Whilst it is acknowledged by the authors that the limited bandwidth will affect the higher formant frequencies and that the tracker clearly has difficulty in locating some of the formants, no mention is made of the suitability of the analysis parameters selected, particularly the LPC order (2008, 213).

Another technical aspect of the speech signal that has received attention in relation to formant measurement errors is the recording process. The study by Livijn (2004) examines the influence of different recording devices on formant measurements. A modified version of 'The North Wind and the Sun' was read in Swedish, in an anechoic room, whilst being simultaneously recorded via a condenser microphone connected to a computer, via a dynamic microphone connected to a standard cassette recorder, by a microcassette recorder and via a mobile telephone that was being recorded at the distant end of the line. Formant frequencies F1 to F3 were measured at the start, middle and end of all the 18 vowels in each recording using Praat. It is assumed that an LPC analysis was used although it is not stated and no settings are given. The values from the recording made directly to the computer were considered as the reference set and the measurements from the other devices were compared with them. The largest deviations were found for the microcassette recording, followed by the mobile telephone, then the

standard cassette. The largest divergences for the microcassette are reported as 13 %, 12.4 % and 10.6 % for F1 to F3 respectively. However, a number of large differences are attributed to artefacts of the measurement process rather than an inherent shift caused by the recording method. Upward shifts in F1 are again observed for the mobile phone.

The impact of lossy compression algorithms is the focus of the work by van Son (2005). Four male and four female speakers were recorded simultaneously via a fixed condenser microphone and a head-mounted dynamic microphone to an audio CD recorder. The speech material was read and retold versions of the 'The North Wind and the Sun', in Dutch. The recordings from the condenser microphone were re-recorded to a MiniDisc player so that the material would be compressed using the ATRAC3 method, as well as separately being subjected to MP3 encoding at a bit rate of 192 kbps and Ogg Vorbis encoding at 80 kbps and 40 kbps. Formant measurements for F1 to F3 at vowel midpoints were made using Praat's Burg tool with default settings to emulate a naïve user. The formant measurements were converted to semitone values to allow a direct comparison across the three formants, with a difference of 1 semitone being approximately equivalent to 6 % within the range of 0.25 to 3 semitones. The measurements from the compressed recordings and different microphone were initially compared with those from the reference CD in order to locate large differences in formant values for each vowel token. Differences larger than 9 semitones for individual tokens were removed from the further analysis of the results. Of the 2415 tokens, the most were removed for the different microphone condition, approximately 3.8%, 2.4% and 0.2% for F1 to F3 respectively. The percentage of rejected tokens for the compressed recordings were much lower than for the microphone change and were relatively consistent across codecs (0.8% for F1 and F2, 0.1% for F3). The least rejections occurred for the MP3 encoded material. Analysis of the remaining errors, with the outliers removed, showed that in terms of RMS errors, expressed in semitones, the microphone change produced the largest errors of approximately 1.7 semitones for F1, 1.3 semitones for F2 and 1.2 semitones for F3. For each codec, the performance was similar across the three formants, with the Ogg Vorbis 40 kbps material performing worst, with an error of approximately 0.8 semitones RMS error for each formant, and the MP3 material performing the best with errors of approximately 0.3 semitones RMS. The MiniDisc and Ogg Vorbis 80 kbps results were similar at around 0.6 semitones RMS.

The results from van Son (2005) showed that the greatest differences in formant measurements were introduced by using a different microphone. Results from a preliminary study using three different microphones indicate large changes can occur in formant values measured by LPC analysis (Hansen and Pharao, 2006). Almost 15% of the 252 measured values from one speaker showed differences between 5 and 10%, whilst a further 12% showed differences greater than 10%. A further experimental procedure using more microphones at a number of distances and with more speakers is described but no detailed analysis was reported.

In addition to the microphone, other factors that influences the frequency spectrum of the signal are the acoustic environment in which the speech occurs and the distance from the speaker to the microphone. A small-scale study by Vermeulen (2009) aimed to investigate these effects on formant measurements. Twelve synthetic steady state vowels were generated with a fundamental frequency of 100 Hz and formant frequencies that coincided with the harmonics. The vowels were replayed via a loudspeaker in three acoustic environments, a semi-anechoic room, a long corridor and a domestic living room. Recordings were made at a range of distances from the loudspeaker. Initial formant measurements made via an LPC analysis in Praat of the original synthetic vowels revealed an average error across the four formants of the order of 4%. It was decided that LPC analysis would not be used to analyse the recordings so the relative amplitudes of the harmonics of the fundamental measured from FFT spectra were considered instead. Statistically significant differences were observed across the spectra for the distances in each acoustic environment but the interpretation of the results in relation to how the changes affected the appearance of formants was problematic.

2.1.4 Contextualising Formant Variation & Errors

Several of the studies summarised above relate the magnitude of the errors found to difference limen in order to contextualise the results (e.g. Lindblom 1960, Monsen and Engebretson 1983, Vallabha and Tuller 2002). Whilst this is a useful technique, the significance of formant measurement errors is different across applications. The errors themselves may be of little relevance; it is how the measurements are interpreted and what conclusions are drawn from them that are most important. This issue is considered in the context of sociophonetics by Woehrling and Mareüil (2007). The study compares

the performance of Praat and Snack[1] on two large corpora of French, one consisting of face to face recordings, the other containing telephone speech. In addition to comparing the performance of the software, the aim of the study is to determine if it is possible to discriminate two varieties of French based on the formant measurements.

For both Praat and Snack the frame advance was set to 10 ms and a frame width of 50 ms was chosen. Given the limited bandwidth of the telephone recordings the upper analysis frequency in Praat was limited to 3000 Hz for male speakers and 3300 Hz for females and the number of formants to extract was set to 3. However, it is not clear if the tracker function in Praat was used or if the Burg tool was used and the setting of 3 corresponded to an LPC order of 6. No mention is made of the LPC order in Snack or whether the upper analysis frequency was reduced, but it is stated that the other parameters were set to default. Formants were measured for 10 phonological vowels that had been automatically segmented in the corpora. Whilst the number of tokens is not provided the telephone material contained approximately 70 speakers per region with an average conversation duration of 14 minutes. A similar amount of material was available in the face to face corpus.

Formant measurements that were outside a range of ± 500 Hz from a set of reference values for each vowel category were discarded. For each vowel category, correlation coefficients were calculated for mean F1 and F2 values for each speaker and distances were calculated between the means within each corpus. Overall, the correlations between the F1 values for the male and female speakers in each corpus were all greater than 0.85. Summed absolute differences for F2 were less than 50 Hz for all circumstances but were greater than 50 Hz for F1 for the telephone speech. In general for individual vowels the correlations were good with only 16% being under 0.7. However, some weak correlations and large distances were noted for the telephone corpus, with Praat's F1 values being consistently higher, which resulted in a vertical shift in plotted vocalic triangles. They therefore warn:

> 'Praat and Snack exhibit substantial differences, especially on F1 and certain vowels. Therefore comparisons among vowel spaces stemming from different signal processing tools must be taken with caution.' (2007, p. 1008)

[1] The software WaveSurfer is built on the Snack toolkit and produces identical measurements. This is discussed in more detail in Section 7.2.2.

The study goes on to examine various differences found between the northern and southern varieties of French in the corpora. Finally, decision tree techniques were applied to the measurements from the vowel /ɔ/ to see if they could be used to discriminate between speakers from the north and south of France. The performance of the system was between 73% and 97% depending on the software, sex of the speakers and the corpus used. This performance was considered sufficient to 'outline a spreading linguistic change: /ɔ/ fronting in northern French' (2007, p. 1009). Whilst the study clearly demonstrates the differences in the measurements between the software used, its aim of discriminating two varieties of French was achieved.

A study by de Castro et al. (2009) examines the performance of a forensic speaker comparison method based on the statistical modelling of features extracted from formant measurements obtained via an automatic measurement process and compares it with those made by an analyst. A clear motivation for using an automatic measurement approach is that it is much quicker and also allows the analysis of more material, resulting in statistical models that are more representative. Whilst no information is provided in relation to the actual measured values and the differences found between the two methods, it is accepted that analysts will produce more accurate measurements. The outcome of the tests of the speaker comparison system revealed that even though the performance based on the automated formant measurements is worse than with the human measured values, the performance is still acceptable.

A similar, but more extensive set of tests are reported by Zhang et al (2012). They tested the performance of 5 formant trackers and 4 analysts by fusing the results from the formant measurements with a baseline MFCC system and assessed whether the addition of the formant data provided an improvement over the MFCC system on its own. The comparisons were also conducted using different quality recordings from telephones. Again, information concerning the differences in the actual measurements is not provided but an assessment of the within-analyst reliability is reported showing relatively good within and cross analyst agreement. The fusion of the human-supervised measurement results with the baseline system always led to an increase in performance over the baseline system on its own. The pattern of results from the automatic measurement systems was more complex with some trackers in some conditions improving the performance of the combined system, whilst others did not. However, the

study concludes by questioning whether the degree of improvement from the analyst based measurements is justified given the time required to make the measurements.

2.2 Software Performance & Guidance

The studies discussed above provide many useful insights into the measurement of formants, including factors that can influence the measured values and the magnitude of errors or variation that can be encountered. However, a number of specific issues are not addressed by the literature, but which are relevant not only to forensic speech scientists, but the wider phonetics community. Several of the studies examine the variation in performance of analysis methods when parameters such as LPC order are altered, but they have not been conducted using software that is currently in widespread use. It is therefore not certain how the findings might relate to these specific implementations of the measurement methods. Those studies that do use current software tend to have a different focus and the performance of the software is not addressed at a level of detail sufficient to yield any significant insights that might assist analysts when making their own measurements. There are some studies which provide a comparative analysis of the performance of current software, and these are discussed in Section 7.4, but the results only serve as a benchmark against which a novel approach is being assessed. These studies also highlight a number of problems when interpreting the reported performance, such as insufficient detail concerning the methods followed and the presentation of results in ways which makes them difficult to compare across studies. Comparisons of the performance of commonly used software have been conducted from a forensic perspective (Schiller and Köster 1995 and Howard et al 1993), but they only concern the measurement of fundamental frequency. A further work by Morris and Brown (1996) also addresses the accuracy of fundamental frequency estimates from a general speech analysis viewpoint. At present, no similar studies exist either in the forensic field or within phonetics more broadly that directly address the performance and variability of commonly used formant measuring tools.

A further shortcoming of the studies reviewed above is their limited attention to the performance for different speakers. Whilst the work by Vallabha and Tuller (2002, 2004) does propose an approach for selecting an optimum LPC order for an individual speaker, this is only demonstrated on a small number of speakers. Also, many of the studies consider measurements from a range of vowels but little attention is paid to variation across the vowel space.

In terms of practical guidance that is available to analysts when making formant measurements, very little is available. Earlier work such as Makhoul (1975, p. 574) discusses the issue of determining an optimum LPC order for the overall representation of the spectrum rather than for specifically measuring formants. The suggested approach is to examine the system error, i.e. the difference between the original signal and the LPC signal, for increasing LPC orders until the error no longer significantly decreases with increasing LPC order. Whilst this may appear to be a sensible approach, it is difficult to implement in modern analysis software and it is not apparent if it would result in the most accurate formant measurements. One of the earlier studies that does address obtaining formant measurements from an LPC analysis (Markel, 1972) discusses the issue of selecting appropriate analysis parameters. For LPC order it states that it 'is not a strong function of the particular speech sound' but 'it is a strong function of the system sampling rate' and therefore the maximum analysis frequency (1972, p. 134). It recommends that a suitable LPC order can be calculated as the sampling rate measured in kHz plus 4 or 5. So for a sampling rate of 10 kHz (giving a maximum analysis frequency of 5 kHz) the optimum LPC order is 14 or 15. This advice is repeated in Markel and Gray (1976, p.154) and is often considered as a general rule of thumb for determining a suitable LPC order.

A slightly different rule of thumb is provided by Ladefoged (1996, p.212) and suggests taking the sample rate in kHz and adding 2. However, he describes choosing the correct LPC order as being 'somewhat of an art' (1996, p. 212) and ultimately suggests trying several LPC orders and then seeing which provides the 'most interpretable results'. Harrington and Cassidy (1999, p. 221) recommend that the minimum LPC order for voiced male speech is equal to the sample rate in kHz. For a recording with a sample rate of 10 kHz or a specified maximum analysis frequency of 5 kHz, the three rules of thumb suggest a range of LPC orders from 10 to 15. The suitability of the orders within this range and the sensitivity of measurements across it has not been subject to empirical testing using modern software implementations.

One acknowledgement in a forensic text of the variation in performance between formant analysis software appears in Rose (2002, p. 265-267). Based on this variable performance Rose states that 'it is mandatory to carry out comparison of questioned and suspect material on the same equipment, with exactly the same settings' (p. 267). Whilst on the face of it this may appear to be sensible advice, it contradicts the

recommendation given by Vallabha and Tuller (2002, p. 156) that 'the order of the LP filter should be matched to the utterance being analysed whenever possible' and where that is not possible it 'should at least be matched to each speaker'. Such contradictory advice is clearly problematic for analysts attempting to determine what might be considered as 'best practice' based on the guidance of others.

Another source of information on LPC analysis is textbooks concerned with speech analysis. They frequently contain descriptions of the principles of LPC analysis, the limitations and the pitfalls, but by their nature any advice or suggested settings are very general and not software specific. At the other extreme the manuals or help files for software packages may provide a description of the algorithm or analysis process, the available analysis parameters and default values, without providing any detailed guidance in their usage.

2.3 Present Research

2.3.1 Motivation

It is apparent from Section 1.3 that formants are considered to be an important speech feature in the field of forensic speech science and that they are measured and analysed in a significant proportion of cases. It is also clear from previous sections that formants are subject to many sources of variation from both a speech production perspective and from technical factors including the type of signal transmission and the measurement method. These factors mean that formant measurements used in forensic analysis will contain inaccuracies and errors. If the sources of error and the likely reliability of the measurements are not understood and accounted for then analysts are at risk of misinterpreting the data. This has the potential to influence and ultimately alter the outcome of individual forensic tests, which in turn can affect the final conclusion reached by a forensic scientist concerning the identity of a speaker or the interpretation of an utterance. In the most extreme situation it is possible for the misinterpretation of erroneous formant measurements to be a significant contributory factor to a miscarriage of justice. Whilst the author is not aware of any examples of this having occurred, based on other instances of misinterpreted data that have been encountered in casework, it is possible to envisage scenarios in which it could.

The potential impact of formant estimation errors is a function of both the magnitude of the errors and the weight or reliance placed on the measurements as part of an

individual test and in reaching a final conclusion. Whilst highly inaccurate measurements have the potential to lead to the most significant misinterpretation of the data, they should also be more easily identifiable as erroneous, allowing them to be rejected. It is the measurements which are moderately inaccurate that pose the greatest risk of misinterpretation.

The impact of errors also depends on whether they are random or systematic. Since formants are subject to natural variation in speech production, caused by differences in articulatory movement and co-articulation effects, multiple tokens of vowels from the same category are usually analysed to obtain a distribution of measurements. Random errors in the measurement process will cause these distributions to be artificially wide. In speaker comparison analysis the distributions are compared across samples. If the random nature of the errors is the same across the samples, and there are a sufficient number of representative tokens in each sample, then the overall influence on the two distributions should be similar. Whilst this is of limited significance for distributions that genuinely display a high degree of overlap, for non-overlapping or partially overlapping distributions the random errors may artificially increase the degree of overlap. This is problematic for the interpretation of the data as the extent of the overlap may be incorrectly attributed to the degree of similarity between the samples rather than inaccurate measurements.

Similar issues arise when measurements from an individual token are compared with a distribution from another sample, as is often done when analysing disputed content. Random errors will result in an artificial widening of the distribution and even where the measurements from the disputed token are not truly part the distribution they may fall within it. This again leads to the potential for the data to be misinterpreted.

Systematic errors result in predictable shifts in measured values, such as that caused by the filtering effect of telephone transmissions (Künzel 2001 and Byrne and Foulkes 2004). If the process that caused the shift applies equally to the measurements being compared, then this type of error is potentially unproblematic as both are affected to the same extent. However, if only one set of measurements is affected or the two are affected differently then there is the potential for the results to be misinterpreted. If formant measurements from two samples originating from the same speaker are affected differently, their distributions may appear separate when in fact they should overlap.

The opposite could also occur where distributions from different speakers are shifted so they overlap when in fact they should be separate.

The potential impact of errors is also influenced by the number of tokens analysed. In general, the greater the number of tokens the better the representation of the true distribution of values. A single erroneous value in a well-represented distribution will have a smaller overall impact than in a sparse distribution. Within a well-represented distribution a single erroneous value may appear as an outlier allowing it to be rejected or reanalysed. However, forensic samples are often short and of poor quality which limits the number of tokens available for analysis. In the case of disputed utterance analysis the acceptance or rejection of an interpretation can be strongly influenced by the measurements from a single token in the word in question. In situations such as these with very limited data the potential impact of errors is at its greatest.

To arrive at a conclusion for a forensic case, either on the identity or non-identity of a speaker or the acceptance or rejection of the interpretation of an utterance, the outcome of a formant analysis is assessed subjectively in conjunction with the outcome of other examinations, such as an auditory analysis. Since these processes are subjective, there are no fixed thresholds for determining the results of individual analyses or the final conclusion. Reaching a conclusion can be particularly difficult when the results of a formant analysis are at odds with the results of other tests or where the formant analysis outcome is unclear. Coupled with this difficulty, forensic scientists are also susceptible to cognitive bias when making measurements and interpreting findings (Kassin et al 2013). Confirmation bias, the tendency to find features, make measurements or interpret findings in such a way as to support an opinion that has already been formed, can affect the measurement and interpretation of formants. This type of bias can manifest itself as analysts being less critical of measurements which appear to support their already formed opinion. Alternatively, analysts could be overly critical and use measurement errors as an explanation for findings that do not fit their conclusion. Analysts may also be overly reliant on formant measurements as they consider them to be superior to other forms of analysis as they provide what appears to be an objective result. This can occur with material of poor quality, which is a common attribute of forensic recordings, where formants are unclear in spectrographic representations but measurements from a LPC analysis are nevertheless accepted as accurate.

The discussion above highlights the potential impact that formant measurement errors can have on forensic analyses. For forensic speech scientists to be able to properly analyse and interpret formant measurements and reach conclusions based on them, they must have knowledge and understanding of the sources of variation and the resulting reliability of the measurements. Whilst some of this knowledge can be acquired through personal experience it must also be obtained from empirical studies. Some research has focussed on formant accuracy and variation from a forensic perspective such as Duckworth et al (2011) concerning the influence of analysts' decisions on measurements and Byrne and Foulkes (2004) on the telephone effect. A very limited number of studies consider the difference in performance of automatic speaker comparison systems when using automatic versus manual measurements (Zhang et al 2012 and de Castro et al 2009). The literature does provide some insight into the reliability of LPC derived formant measurements but there is a lack of information concerning the performance and behaviour of tools currently used by analysts. This is a significant shortcoming as it cannot be assumed that all software implementations will behave in the same way. Furthermore, there is very little empirically derived advice available to analysts concerning the measurement of formants. It is these issues that are the motivation for the research presented in this thesis.

2.3.2 Research Goals

The ultimate goal of the thesis is to provide guidance and information that will be of assistance to forensic speech scientists when making and interpreting LPC derived formant measurements. This guidance and information will be based on the empirical study of the behaviour and accuracy of formant measurements made by software currently in use by forensic speech scientists. The insight gained from these investigations should allow analysts to better understand some of the factors that can influence the accuracy of formant measurements and therefore make better informed decisions when making and analysing measurements. As well as facilitating greater accuracy of measurements, the results of this work have the potential to improve the performance of speaker comparison and disputed content determination methods.

2.3.3 Research Questions

In order to focus the investigations presented in this thesis, three research questions are posed.

RQ 1. What influence does the LPC formant measuring tool have on the accuracy of formant measurements?

RQ 2. How does altering the LPC analysis parameters affect formant measurement accuracy?

RQ 3. To what extent does the accuracy of LPC formant measurements vary across speakers?

The questions concern three important factors that can affect the accuracy of formant measurements. The influence of the measuring tool and how it is used is addressed in Question 1. For analysts to reliably use specific software, its behaviour and performance must be understood. It is not sufficient to simply assume that results reported in the literature are universally applicable to all software, so tools currently used by analysts must be tested empirically. Since different analysts may use different software, it is important to understand how measurements may vary between them, especially if the results from one piece of software may be compared with those from another. Considering different tools is particularly important as guidance derived from the results for one may not be applicable to others.

Question 2 concerns the influence of the analysis parameters. Understanding the effects that altering the parameters can have on measurements is important since it is the means by which analysts interact with the software and can influence the measurements. Such understanding will allow analysts to use tools in the most appropriate ways in order to make more accurate measurements and develop better analysis strategies. These effects must be investigated across software tools as the findings from one may not be applicable to others.

The effect of the speaker is the focus of Question 3. Since the LPC method relies on a simplified model of speech production it is to be expected that the behaviour and accuracy of formant measurements will vary across speakers because for each one the degree of correspondence with the model will be different. This source of variation is a further factor that analysts must consider when making and interpreting measurements. It is of particular relevance for the forensic speaker comparison task where formant measurements are compared across recordings from potentially different speakers.

2.4 Summary

The present chapter has reviewed the literature concerning formant measurement errors arising from a range of sources, namely the measurement method, analyst variability and the technical characteristics of the speech signal, as well as considering how the errors may be contextualised. The limited guidance on measuring formants and issues concerning the performance of software are also discussed. This was followed by a presentation of the motivation and goals for the current research and the research questions. The following chapter describes a pilot study concerning the variability of formant measurements across three software packages (Harrison, 2004) and a supplementary analysis of the results.

Chapter 3 Variability of Formant Measurements Across Current Software

This chapter is a summary of the research conducted for the author's MA dissertation (Harrison, 2004) and is supplemented by a further analysis of the data carried out after the dissertation was completed. This further work was presented at the conference of the International Association for Forensic Phonetics and Acoustics (IAFPA) in 2006 (Harrison, 2006).

3.1 Introduction

The review of the literature in the previous chapter revealed that little attention has been paid to the behaviour and performance of formant measuring tools currently used by speech analysts. This chapter begins to address this issue by analysing and comparing formant measurements obtained from three commonly used LPC analysis tools for two speakers across of range of analysis parameters. The analysis of the results from these experiments addresses the first two research questions, which ask what influence the software and the analysis settings have on the accuracy of formant measurements. This is achieved firstly by examining how the measurements vary as the analysis parameters change, and secondly by considering the proportion of accurate measurements obtained at different LPC orders. Since the measurements were only obtained for two speakers, the experimental results provide a limited answer to the third question concerning how the accuracy varies across speakers. Despite the limitations of the study, the results highlight the importance of empirically testing software and provide the basis for some important guidance for analysts when making formant measurements.

3.2 Methodology

3.2.1 Determining Accuracy

The nature of formants means that it is problematic to determine the accuracy of formant measurements as there is no sufficiently reliable or accurate method that can be used to obtain 'true' values which can be compared with measured values. In recognition of this fundamental issue, the approach chosen for this pilot study was to consider the relative variation of formant values across analysis settings rather than attempting to determine the absolute accuracy of the measurements.

3.2.2 Speech Data

In order to have control over the speech material, recordings were made specifically for the study rather than relying either on forensic case materials or recordings made for another purpose. To replicate some of the range of quality that can be found in forensic recordings, the speech material was recorded simultaneously via a microphone (Shure SM58) and at the far end of a landline telephone to landline telephone line connection. The recordings were made to a Tascam DA-40 digital audio tape recorder at a sampling rate of 44.1 kHz and 16 bit resolution. The simultaneous recording process allowed the measurements from the two channels to be directly compared without the differences in production which would have occurred had the material been repeated for the second channel.

To further control the speech material to ensure a sufficient number and range of vowel tokens, a word list was compiled. The words are shown in Table 3.1. The list contains real words, mainly in a CVC structure, with an initial /h/ due to its open articulation requiring minimal articulatory movement to reach the vowel target. The vowels were selected to represent the four extremes of the vowel space generally utilised by speakers of most accents of English and have the lexical headwords FLEECE, TRAP, PALM and GOOSE (Wells, 1982). A neutral vowel, NURSE/lettER, was also included. The final consonant was controlled to allow an investigation into whether this factor had any influence on the measurements. However, this aspect of the data analysis was not undertaken.

Final Consonant	**Vowel Category**				
	FLEECE	**TRAP**	**PALM**	**GOOSE**	**NURSE/lettER**
Zero	*he*	*ha*	*Har*	*who*	*hisser*
/t/	*heat*	*hat*	*heart*	*hoot*	*hurt*
/d/	*heed*	*had*	*hard*	*who'd*	*herd*
/s/	*cease*	*pass*	*Haas*	*Soos*	*hearse*
/z/	*he's*	*has*	*SARS*	*who's*	*hers*
/n/	*seen*	*Hann*	*Hahn*	*Hoon*	*Hearn*

Table 3.1 Word list arranged according to final consonant and vowel category.

The word list was presented to the subjects with the word order randomised to remove any ordering effects, and filler words were included at the start and end of the list to combat any list effects. The list was read three times resulting in 18 tokens per vowel category and 90 tokens in total. The subjects were two male native British English speakers, including the author.

3.2.3 Software

It was observed in the literature review that the majority of previous studies of formant measurement accuracy have not been carried out on implementations of the LPC algorithm in software currently used by phoneticians. As one of the aims of the study was to test currently used software, members of IAFPA were contacted by email in 2004 and asked what software they used when making formant measurements. Sixteen of the fifty-six members responded. The three most commonly used programs were Praat (8 users), Kay CSL/Multi-Speech (8 users) and WaveSurfer/X Waves (5 users). Based on these results the three programs used in the study were Praat (Boersma, 2001), the Snack Sound Toolkit (Sjölander, 1997) and Kay Multi-Speech (Kay Elemetrics, 2004). The Snack Sound Toolkit is the underlying software that WaveSurfer is built on and the two systems produce identical measurements.

3.2.4 Analysis Settings

The formant analysis tools within each of the three programs require a number of different analysis settings to be specified. Analysing the effects and interactions of all possible settings would have made the study prohibitively large. Therefore, a subset was chosen based on two criteria: settings which are likely to be adjusted by an analyst, and settings which are sufficiently similar across the programs. The settings chosen were LPC order (specified via the 'number of formants' setting in Praat), pre-emphasis and frame or analysis width (or length). To restrict the complexity of the study, the effect of each setting was examined independently, i.e. when one setting was varied the other parameters were kept at their default values. The settings used are listed in Table 3.2.

Multi-Speech			Praat			WaveSurfer		
LPC	**Width (s)**	**Pre-Emph**	**Formants = LPC**	**Width (s)**	**Pre-Emph (Hz)**	**LPC**	**Width (s)**	**Pre-Emph**
6	0.005	0.0	3 = 6	0.005	1	10	0.01	0.0
8	0.010*	0.3	4 = 8	0.010	25	11	0.02	0.1
10	0.015	0.6	5 = 10*	0.015	50*	12*	0.03	0.3
12*	0.020	0.9*	6 = 12	0.020	75	13	0.04	0.5
14	0.025	1.1	7 = 14	0.025*	100	14	0.049*	0.7*
16	0.030	1.3	8 = 16	0.030	125	15	0.06	0.9
18		1.5	9 = 18	0.035	150	16	0.07	
				0.040		17	0.08	
				0.045		18	0.09	
				0.050			0.10	

Table 3.2 Analysis parameters selected as variables and the settings used for each program. Asterisk denotes the default values.

The numerical values of the settings were selected to provide a degree of comparability across the software and also cover a range around the default values that an analyst may choose. Some restrictions were imposed by the software, such as Multi-Speech only permitting even numbered LPC orders. A complicating factor was that the pre-emphasis parameter is not equivalent across the programs. For Praat it specifies the frequency above which pre-emphasis is applied, whereas for WaveSurfer, and presumably Multi-Speech, the value is the coefficient for the pre-emphasis filter.

The remainder of the analysis parameters were kept at their default settings except for the Number of Formants setting in WaveSurfer which was reduced from 4 to 3, as this was the number of formants to be logged. See Section 7.3.2.4 for a discussion of the influence this analysis parameter has on the measurements in a different set of recordings. Also, the Maximum Format parameter in Praat was set to 5,000 Hz, since the default value of 5,500 Hz is more suitable for female speakers, according to the manual. There is no equivalent parameter in Multispeech as the analysis is performed across the entire frequency range of the signal. In order to ensure consistency across the programs the recordings were resampled at 10 kHz, to give a signal bandwidth of 5 kHz, before being analysed in Multispeech.

3.2.5 Measurement Process

To ensure that the same central steady state section of each vowel was analysed across the three programs, a start and end time was determined based on a visual inspection of a broad-band spectrogram in conjunction with the waveform. These timings were used

for all of the measurements at each analysis setting within each of the programs. The small time offset between the synchronised telephone and microphone recordings, caused by the slight delay introduced by the telephone transmission, was determined so that the same timings could be used for each set.

The standard LPC formant measuring tool was used in each of the programs. Within WaveSurfer this is a tracker, whereas for Praat the standard Burg function does not perform any tracking and assumes that the first pole frequency is F1, the second is F2 and so on. Based on the results presented in Figure 3.6 to Figure 3.9, certain aspects of the behaviour for Multi-Speech are the same as that seen for Praat, so it is assumed that the Multi-Speech tool is also a simple formant measurer, not a tracker. The measurement values obtained from each program were the mean of the measurements from all the frames within the analysis time period specified for each vowel. The mean values for the first three formants were logged.

3.2.6 Script Automation

To facilitate the large number of measurements made over the range of analysis settings and tokens, scripts were used to automate the measuring and logging process where possible. This also reduced the potential for mistakes to be made during these processes. This was relatively easy to accomplish for Praat and the Snack Toolbox. However, it was not possible to automate these processes in Multi-Speech, so the formant measurements were manually copied from the software and logged to a spread sheet.

3.3 Initial Analysis of Results

3.3.1 Raw Formant Plots

To obtain an overall impression of the data the mean formant values for each token were plotted. Separate plots were generated for each speaker, analysis parameter and recording channel for F1, F2 and F3 as well as all three formants combined. An example plot of the F1 values from the microphone recording of speaker 1 obtained from Praat whilst varying LPC order is shown in Figure 3.1.

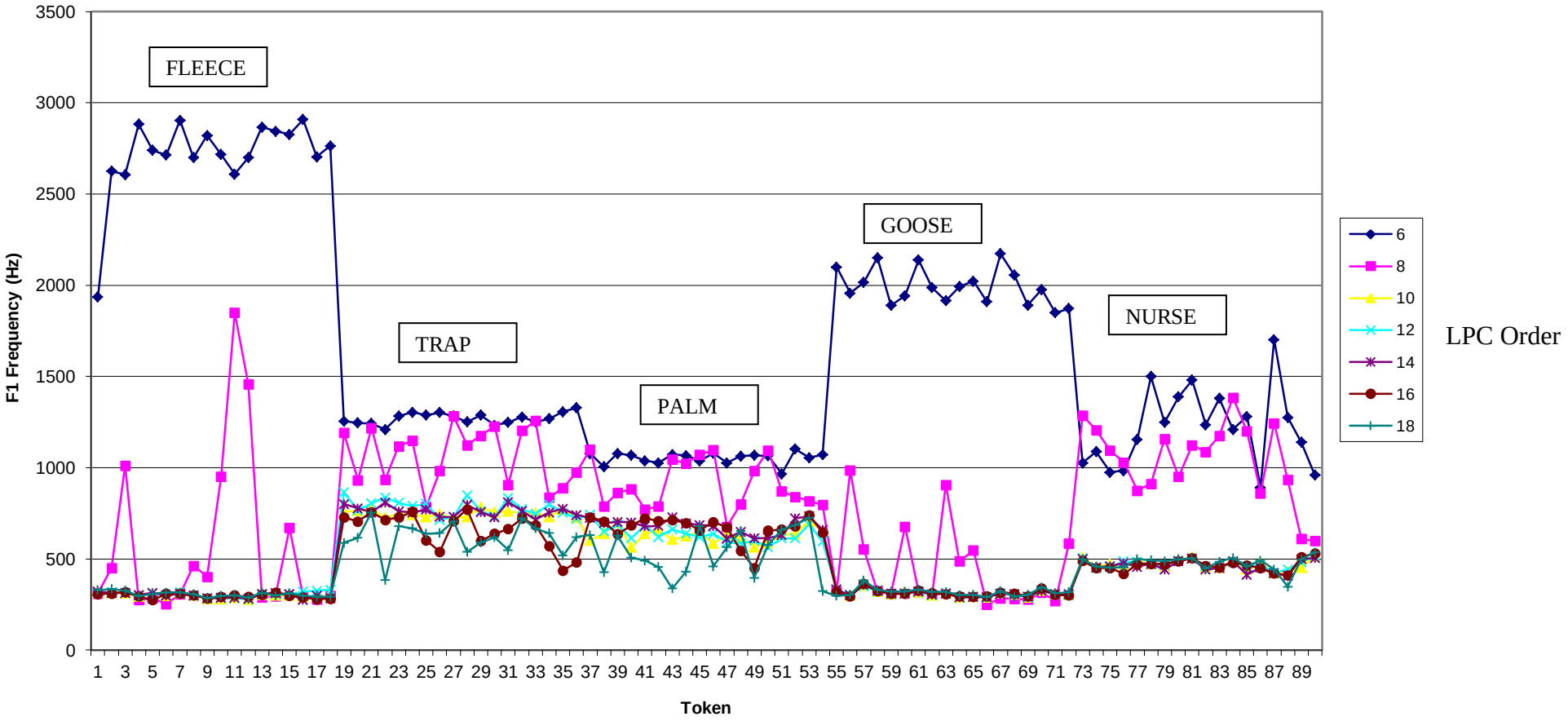

Figure 3.1 Mean F1 values for all of speaker 1's 90 vowel tokens from the microphone recording obtained at different LPC order settings in Praat. The vowel categories are labelled.

The variation in the measurements across the LPC orders can be seen clearly in Figure 3.1. The sets of measurements from the five vowel categories are easily distinguishable due to the differences in the measured values and the different patterns of variation in each category across the LPC orders. Within the categories of FLEECE, GOOSE and NURSE the measurements obtained with LPC orders 10 to 18 are particularly consistent. For TRAP and PALM the LPC orders 10 to 14 show consistencies, while orders 16 and 18 display some variation in the measurements. Across all categories the measurements obtained at LPC order 6 are very different from those at the other orders. This is a consequence of the LPC model having an insufficient number of coefficients to adequately model the speech spectrum. This effect is demonstrated in Figure 1.9. For all the tokens at LPC order 6 the measured F1 value, i.e. the frequency of the lowest pole in the LPC model, does not correspond to the first formant in the speech signal. In the case of TRAP and PALM the measured F1 values correspond to the second formant of the vowel.

The plots of the measurements obtained for the other formants, analysis parameters, software, speaker and recording channel exhibit differing degrees of variation and patterning in the results. It is apparent that the measurements are, to some extent, influenced by all of these variables. Given the range and complexity of variation present, it is difficult to summarise the data meaningfully in this form. However, one very clear result is that variation of LPC order has a much greater influence on the measured formant values than pre-emphasis or frame width. The pattern of variation caused by varying the LPC order is also different for the three formants across the three programs. This finding is discussed further in Section 3.4.

3.3.2 Quantitative Analysis

To reduce the complexity of the data and attempt to reveal any clear patterns, a quantitative analysis was conducted. As previously discussed it was not possible to consider the measurements in terms of absolute accuracy as no true formant values could be obtained. To assess the variation in the measurements across analysis settings the values obtained with the default analysis settings in each program were used as a set of reference measurements. The measurements obtained when one parameter was varied could then be expressed in terms of a difference from those reference values. The absolute differences were calculated for all measurements so that positive and negative differences would not cancel each other out when calculating the mean difference. The

mean differences were calculated for all tokens at each parameter setting and for each vowel category. Table 3.3 shows the average difference results for the data shown in Figure 3.1. Again, the same patterns in the results can be seen, with greatest variation in LPC orders 6 and 8, and greatest stability in the measurements for FLEECE, GOOSE and NURSE in the LPC orders above 10.

	LPC Order						
Vowel	**6**	**8**	**10 (Default)**	**12**	**14**	**16**	**18**
FLEECE	2416	280	0	8	5	9	9
TRAP	522	317	0	45	28	99	130
PALM	425	278	0	39	53	68	114
GOOSE	1681	152	0	2	3	6	13
NURSE	748	570	0	8	15	15	19
All	1158	320	0	20	21	39	57

Table 3.3 Mean F1 absolute difference values (Hz) by vowel category, and all tokens combined, for variation in LPC order in Praat from speaker 1's microphone recording.

Despite analysing the data as mean absolute differences, the results still exhibited complexity, especially across vowel categories. However, the analysis did confirm the trends observed in the raw formant plots. For all programs, when varying LPC order, the mean absolute differences for F1 were smaller than for F2, which were in turn smaller than for F3. For pre-emphasis and frame width the differences across the formants were less pronounced. In general, the mean absolute differences for altering pre-emphasis were less than those from frame width, with the greatest being for variation in LPC order. The results from Praat from altering both pre-emphasis and frame width showed very little variation in any of the formants. In the case of pre-emphasis this could be accounted for by the fact that the parameter operated differently from those in WaveSurfer and Multi-Speech. The complete set of results from the microphone recordings are presented and discussed in Harrison (2004).

3.3.3 Default Settings Measurements

One clear result which emerged from the data was the difference between the measurements obtained with the default settings for each program. Figure 3.2 to Figure 3.4 show the average formant values by vowel category for speaker 1 for F1, F2 and F3 respectively. For the vowel categories TRAP and PALM there are considerable differences between the results from each program, which suggests that the default settings must be resulting in inaccurate measurements in some of the programs. This

illustrates the problem of simply accepting the default analysis settings. As the Number of Formants parameter in WaveSurfer was set to 3, and given the findings discussed in Section 7.3.2.4 from other tests conducted with WaveSurfer, it is likely that the higher values for F3 seen in Figure 3.4 for TRAP, PALM and GOOSE are a consequence of this being not being on the most appropriate setting. It is not clear if this setting caused the behaviour seen for F1 for TRAP and PALM in Figure 3.2.

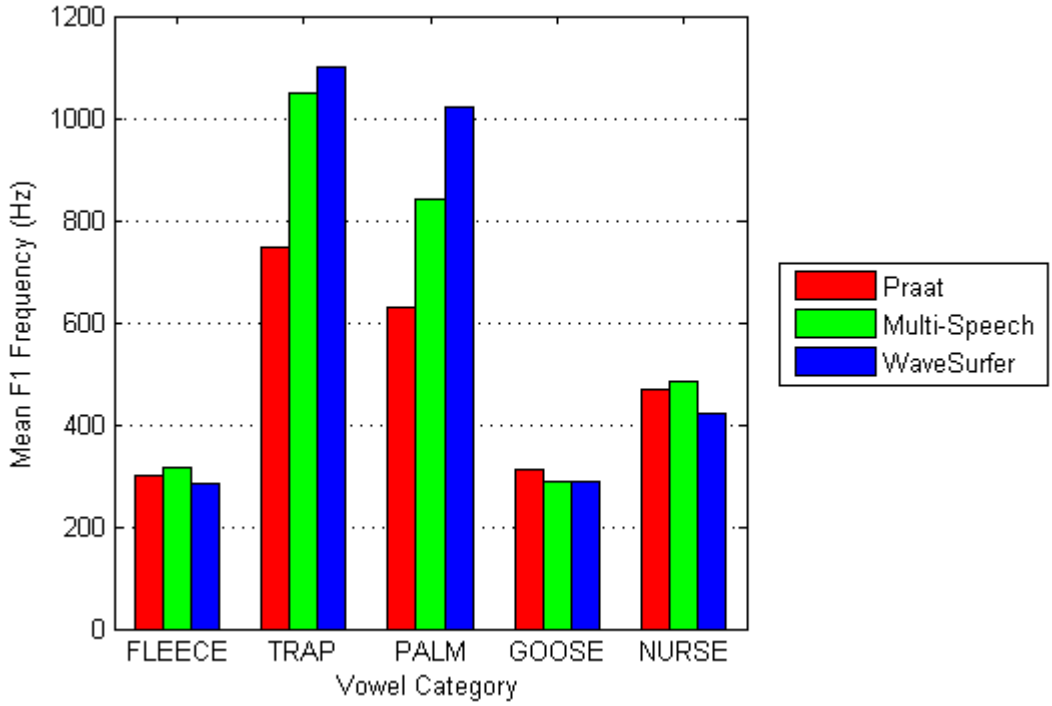

Figure 3.2 Mean F1 values by vowel category obtained with default analysis settings for all three programs for speaker 1's microphone recording.

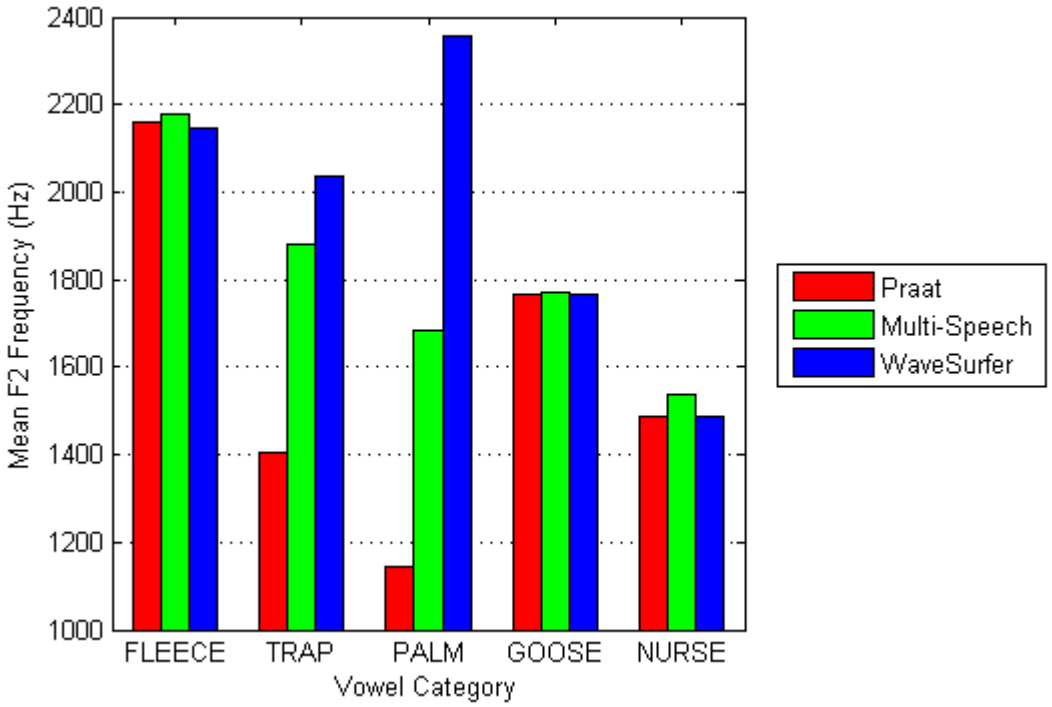

Figure 3.3 Mean F2 values by vowel category obtained with default analysis settings for all three programs for speaker 1's microphone recording.

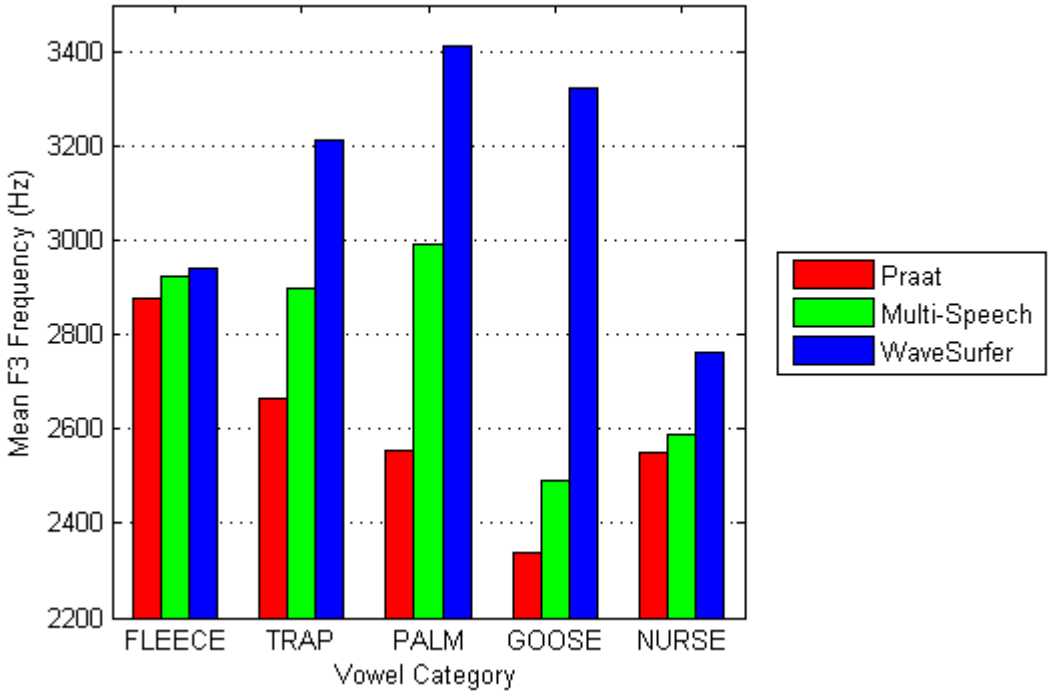

Figure 3.4 Mean F3 values by vowel category obtained with default analysis settings for all three programs for speaker 1's microphone recording.

3.3.4 Summary

Even though the pilot study was limited in its scope, the results showed that considerable variation does occur in formant measurements from software most

commonly used by analysts. The results began to address the research questions posed as they demonstrated variation across the software, the analysis parameters and the speakers. When embarking on the pilot study, it was hoped that it would be possible to propose a clear set of guidelines or recommendations to assist practitioners when making formant measurements. Given the complex set of results and the dependency on each of the experimental variables and the limited data set, this was not possible. However, two general recommendations were made. Since the programs tested are capable of producing inaccurate results, the first recommendation was to compare all formant measurements with spectrographic representations to assist in identifying inaccurate measurements. Secondly, it was suggested that owing to the variation in results obtained across vowel categories, that within a recording the same LPC order should be used consistently for a particular vowel category. A further general comment was made that analysts should be aware of the effects that altering analysis parameters can have on formant measurements.

3.4 Further Analysis of Data

3.4.1 Analysis Method

As discussed in Section 3.3.2, the first quantitative analysis of the data revealed a complex set of results with limited general patterns or trends. The formant measurements were analysed in terms of their difference relative to the measurements obtained with the default analysis settings and no consideration was given to their accuracy. The results in Section 3.3.3 demonstrated that the default analysis settings for some vowel categories and formants resulted in measurements that were very different across the software. The use of these values as reference data in the previous analysis could have resulted in an incorrect impression of the behaviour and performance of the software or a masking of patterns.

Even though it is not possible to obtain true formant values which can be used to calculate the accuracy of the measurements, it is possible to make a judgment about their accuracy more generally. Analysts often make decisions about the acceptability of formant measurements by visually comparing values that are overlaid on spectrograms. Determining the proportion of values that are reasonably accurate would give an indication of performance and would provide a more grounded analysis of the results compared to the previous approach.

In order to make such an assessment of accuracy a criterion must be established for accepting and rejecting measurements as being sufficiently close to the true value. Analysts will generally reject formant measurements as inaccurate if they are overlaid on a broad-band spectrogram and do not visually align with a formant in the spectrogram. To determine this band, several spectrograms of the recorded speech tokens were examined and an impressionistic 300 Hz band that aligned with the visual centre of each formant (i.e. 150 Hz above and 150 Hz below) was chosen as being a reasonable bandwidth within which to classify measurements as being acceptable.

For each token the upper and lower limits of a 300 Hz acceptable band were determined for the first three formants. This was done through the examination of spectra and the measurement of spectral peaks. This proved to be the most successful method, having attempted using spectrograms, LPC derived bandwidth measurements and spectrograms with overlaid formant values.

The spectra were generated with a bandwidth of 260 Hz in order to make the formants visible rather than the harmonics of the fundamental frequency. This bandwidth was chosen as it is the default bandwidth for the spectrogram display in Praat. The spectra were generated over the same material that was used to obtain the mean formant measurements. It was not possible to determine the 300 Hz band for every formant, as some peaks were not clear, so these tokens were ignored in the analysis. In some instances double peaks were present in the location of the formant so the frequency of the peak with the highest amplitude was chosen as the centre of the band.

Each formant measurement was then considered against the acceptable 300 Hz band for that token and was either rejected or accepted. This was only carried out for the measurements obtained from varying the LPC order as the extent of variation present in the measurements from varying pre-emphasis and frame width was relatively small. The percentage of accepted measurements for each vowel category at each analysis setting was then calculated.

3.4.2 Results

The percentage of acceptable formant measurements were plotted across LPC order for each vowel category in the form shown in Figure 3.5.

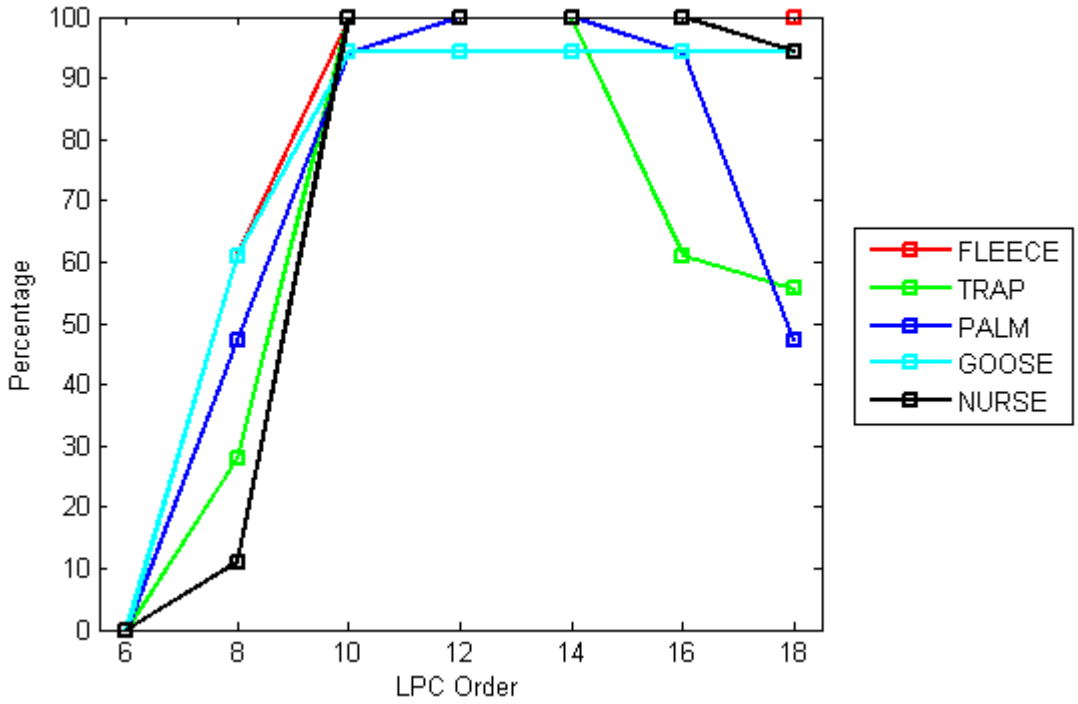

Figure 3.5 Percentage of F1 measurements for each vowel category falling within a 300 Hz acceptable band across LPC order for the microphone recording of speaker 1 from Praat.

The results in Figure 3.5 can be divided in to two groups. The results from vowel categories TRAP and PALM exhibit an inverted U shaped curve whilst FLEECE, GOOSE and NURSE rise to a plateau as LPC order increases. This behaviour is explained below.

All the results for the two speakers, for both the microphone and telephone recording condition, across all formants and in all three programs are shown in Figure 3.6 to Figure 3.9.

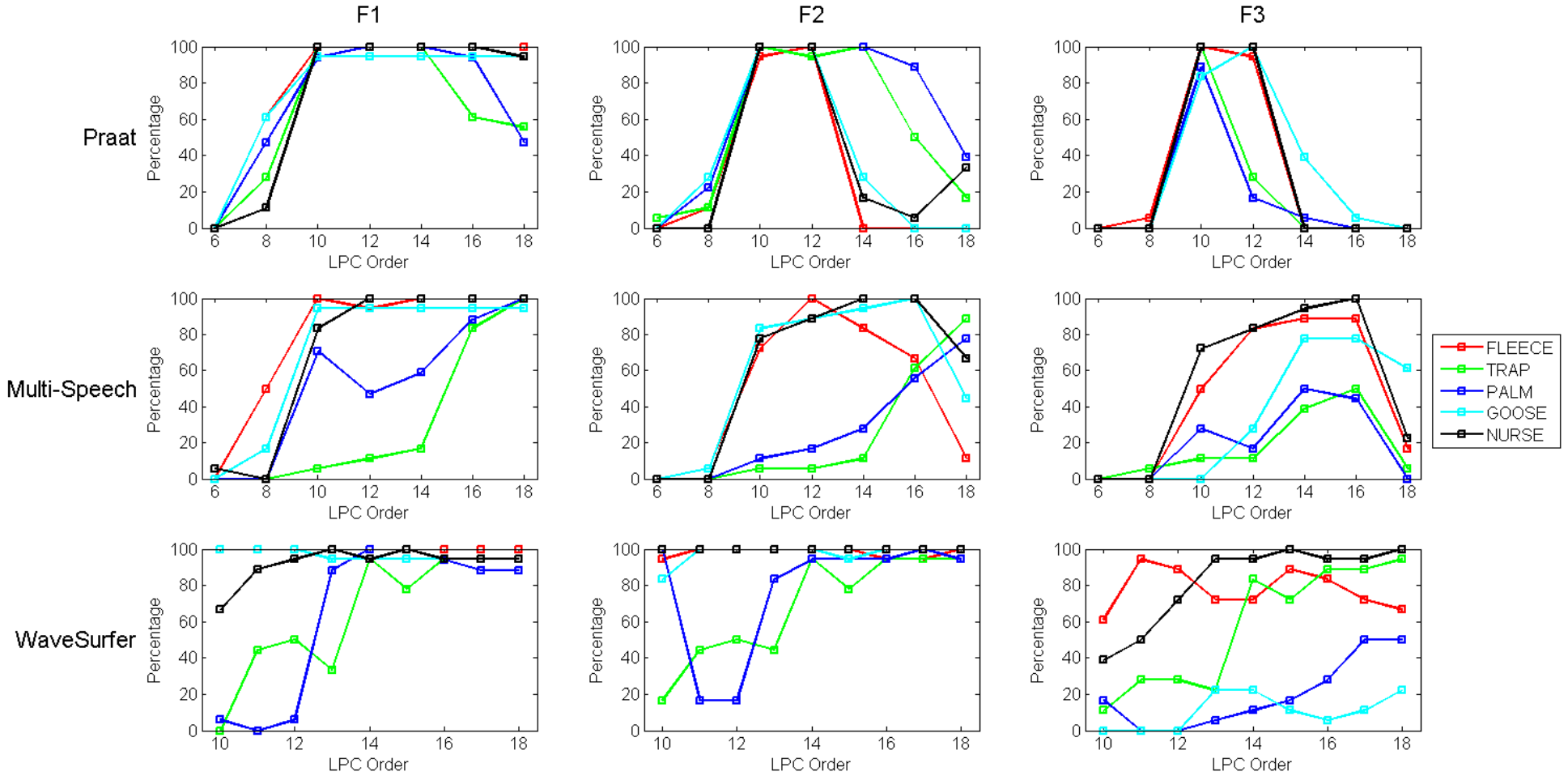

Figure 3.6 Speaker 1 microphone recording – percentage of acceptable formant measurements for each vowel category across LPC order for F1, F2 and F3, across all three programs.

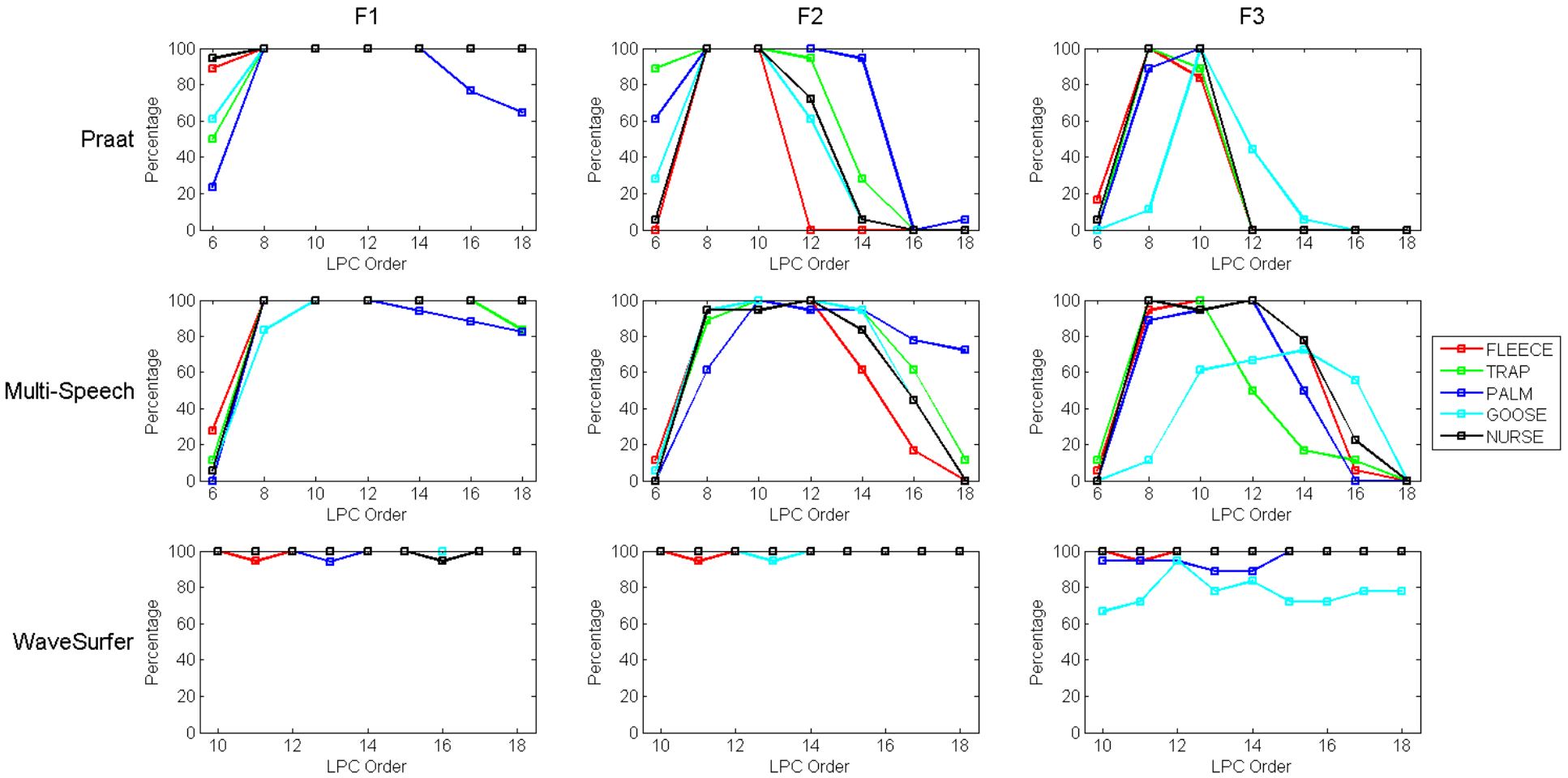

Figure 3.7 Speaker 1 telephone recording – percentage of acceptable formant measurements for each vowel category across LPC order for F1, F2 and F3, across all three programs.

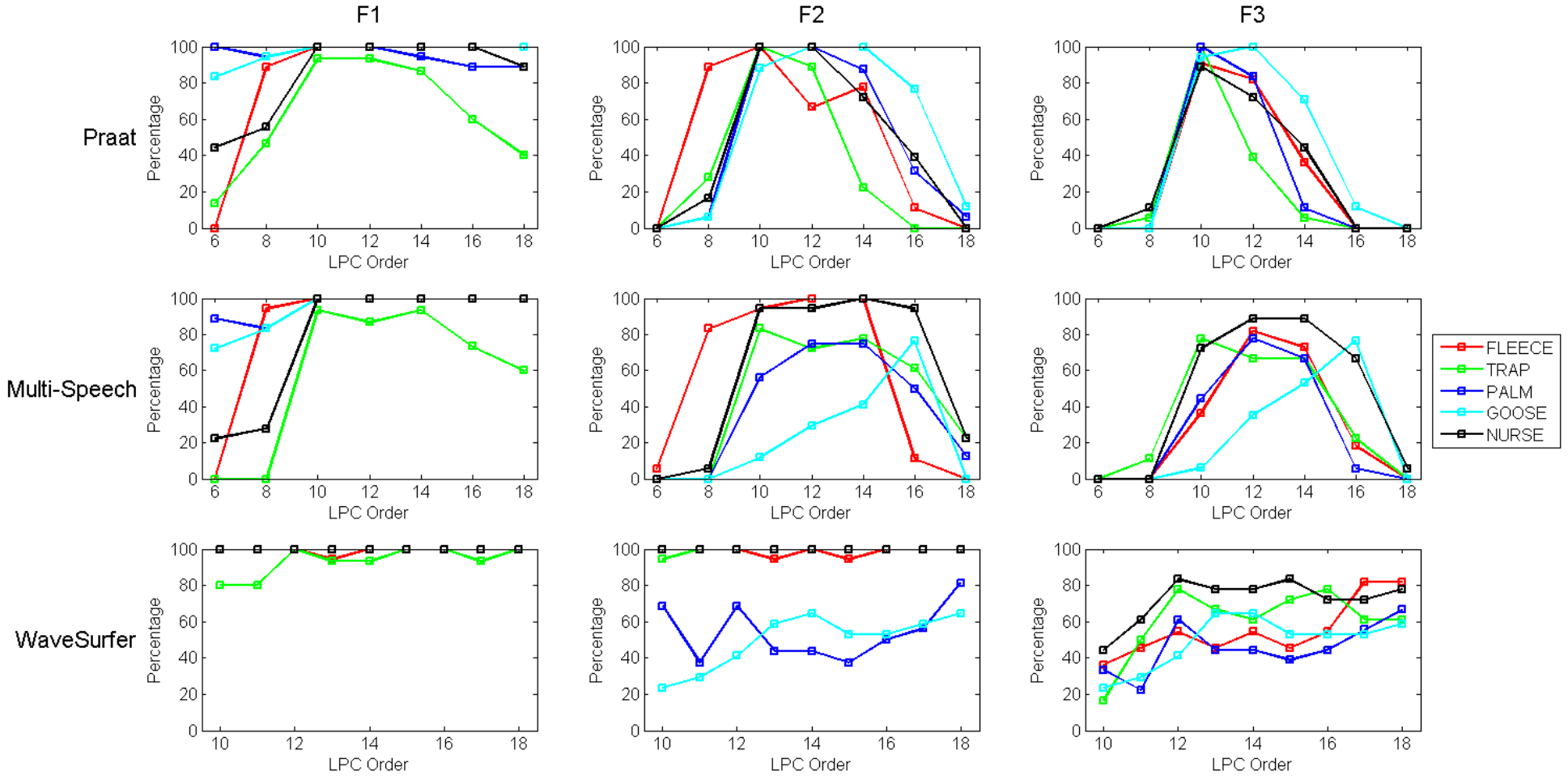

Figure 3.8 Speaker 2 microphone recording – percentage of acceptable formant measurements for each vowel category across LPC order for F1, F2 and F3, across all three programs.

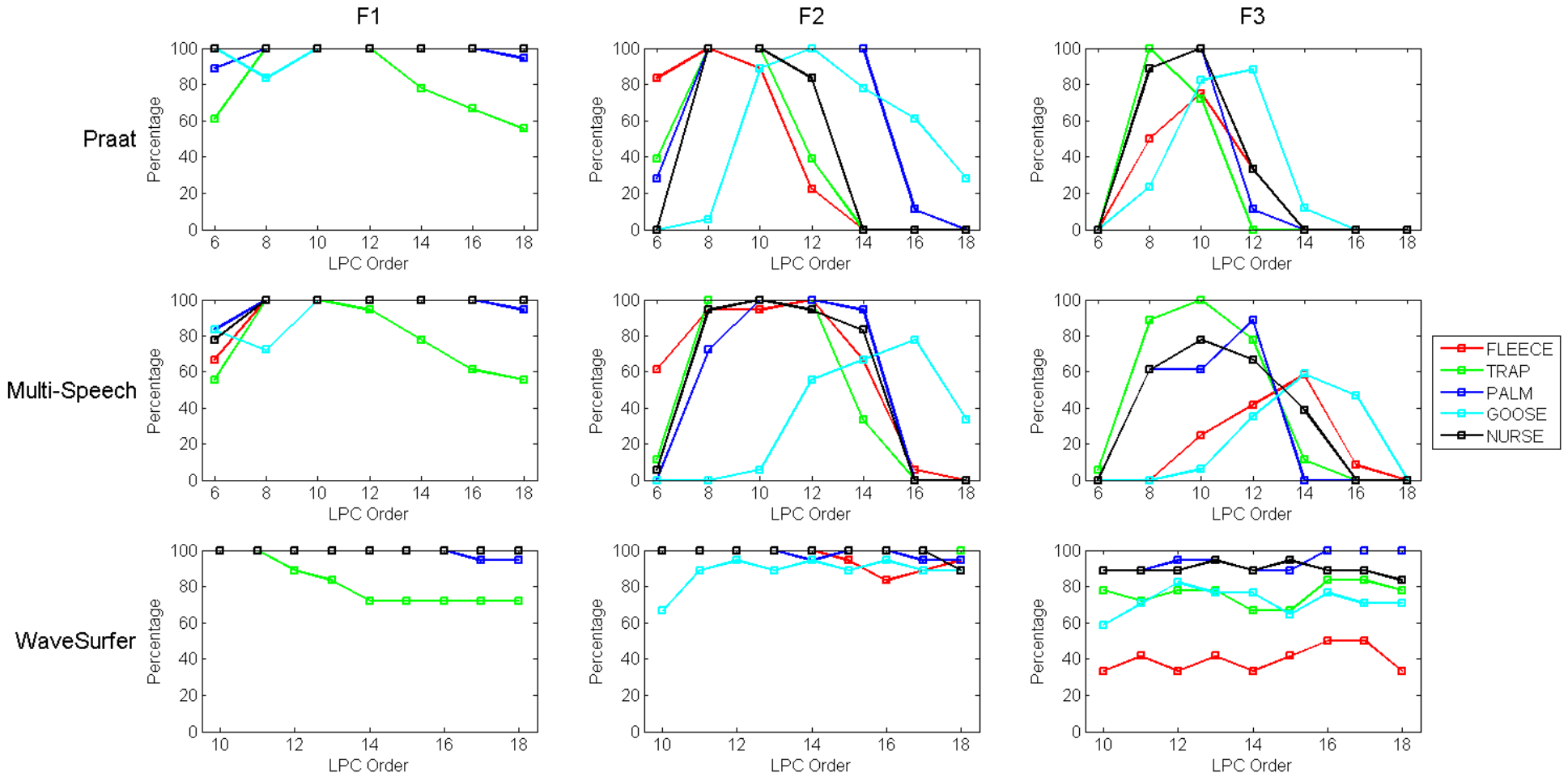

Figure 3.9 Speaker 2 telephone recording – percentage of acceptable formant measurements for each vowel category across LPC order for F1, F2 and F3, across all three programs.

Whilst there is still complexity to the results, some clear patterns do emerge which were not apparent from the previous analysis. Comparing the three programs, the results from Praat and Multi-Speech form either an inverted U shaped curve or they rise to a plateau as the LPC order increases, whereas for the majority of results from WaveSurfer the lines are relatively horizontal. This shows that for Praat and Multi-Speech the general accuracy of the measurements is sensitive to LPC order, whereas for WaveSurfer there is much less influence from LPC order. Overall, for Praat the LPC order that gives the most accepted measurements is 10. For Multi-Speech there is a range from 10 to 14 which produces the most acceptable results.

For Praat and Multi-Speech the results for F1 tend to exhibit a rise and plateau form, whereas the F2 and F3 results have an inverted U shape. This behaviour, and the different pattern of results from WaveSurfer, is a consequence of the way each program derives the formant measurements from the LPC analysis. Praat and Multi-Speech assume that each pole of the filter model defined by the LPC coefficients corresponds to a formant, with the lowest frequency pole being F1, the next one being F2 and so on. When the LPC order is too low the model contains fewer poles or peaks than the speech signal and the poles tend not to align with the formant peaks in the speech so the measurements are often incorrect. This accounts for the lower percentage of accepted measurements at the lowest LPC orders for Praat and Multi-Speech across all three formants. At higher LPC orders the LPC analysis produces a better model of the speech signal and the poles of the model correspond to the formants, resulting in a high percentage of accepted measurements for all formants. As the LPC order increases further, extra poles appear resulting in peaks in the spectrum of the LPC model that do not correspond to formants. These additional poles tend to appear above the first formant of the speech signal, rather than below it, so F1 retains a high percentage of accepted measurements at the highest LPC orders. If an extra pole occurs in the LPC model between the true location of F1 and F2 then the software will return the frequency of this pole as the measurement for F2 since it simply returns the frequency of the second lowest pole. This will also influence the accuracy of the F3 measurement since the third lowest frequency pole will potentially be aligned with the second formant. This results in a low percentage of accepted measurements at the highest LPC orders for F2 and F3. The effect on the LPC model spectrum and the number of peaks as the LPC order increases is demonstrated in Figure 1.9.

WaveSurfer, unlike the other two programs, employs a formant tracking process, which is described in Section 7.2.2.1. At the higher LPC orders the LPC model contains more poles and peaks than the speech signal, just like with Praat and Multi-speech, but the tracker analyses all of the poles in an attempt to determine which ones correspond to formants and which ones do not. The tracking process results in a more consistent performance across the LPC orders tested. The results do not show a low percentage of accepted measurements at the lowest LPC order because WaveSurfer imposes a lower limit on the LPC order to ensure that enough poles are in the model for the tracker to function. The limit is twice the number of formants to be tracked plus four. For these tests the number of tracked formants was three so the minimum LPC order was ten.

Comparison of the results from the microphone recordings with the telephone recordings for Praat and Multi-Speech show a leftward shift in the inverted U curves for F2 and F3, i.e. the highest percentage of accepted measurements occurs at lower LPC orders for the telephone recordings. The results also show an increase in performance for F1 at the lower LPC orders. This is again a consequence of the measurement approach used by the software. Because telephone signals have a reduced bandwidth, and therefore fewer formants, the speech in the telephone recordings is modelled better at a lower LPC order.

For WaveSurfer, the telephone recordings show better performance than the microphone recordings. This is most dramatic for speaker 1 where the microphone recording performance was particularly poor, especially for F3, and the telephone recording achieved almost 100 percent acceptance at all LPC orders, for all formants and across all vowels. This degree of improvement was not as marked for speaker 2, especially with the limited change in the results for F3. This shows that the performance is to some extent speaker dependent. Speaker 2 actually shows a reduction in performance from the microphone to the telephone material for the F1 TRAP measurements made from LPC order 12 upwards.

In general, the level of acceptance is higher for the telephone recordings than the microphone recordings. In the case of WaveSurfer this is likely to be because the Number of Formants setting was at 3, which is better aligned with the expected number of formants that will be found in the frequency band limited telephone recordings. In relation to the other software it is possible that the analysis settings used are also better

suited to the reduced bandwidth signal or that the telephone filtered speech is a better fit to the simple production model assumed for LPC analysis.

3.5 Best Software

An obvious question to ask when comparing different algorithms or programs is "Which one is the best?" In order to answer the question a criterion must be specified against which they can be judged. This could be the accuracy of measurements, the consistency of measurements across analysis parameters or the performance across speakers. Given the range of factors which influence the performance, and the limited number of speakers, there is not sufficient data on which to reliably answer the question. However, this question is returned to in Chapter 7 where the performance of three formant trackers is assessed.

3.6 Summary

The initial results from the pilot study and their further analysis begin to address the first two research questions concerning the effect of software and analysis parameters on formant measurements. Whilst the methodology employed did not allow an assessment of absolute accuracy, it effectively demonstrated the complex variation of formant measurements across three analysis parameters and three programs for two speakers. By considering the general accuracy of the measurements, the further analysis showed quite clearly that the behaviour of the measurements is influenced by the formant measurement method used by the software and that this behaviour is affected by the LPC order. The results also showed variation in the behaviour of measurements across vowel categories. In terms of the analysis parameters, across all programs, LPC order was found to have a much greater influence on the measurements than frame width or pre-emphasis.

The study provided limited insight into the third research question, which concerns the variation of measurements across speakers, since only two were considered. However, notable differences were seen in the measurements across the speakers, which suggest that the research question is well founded.

An important aspect of the study was that the software tested was in common use by analysts. The fact that differences in performance were seen across the programs highlights the need for such empirical testing as it demonstrates that programs do not all

behave the same and that generic guidance may not be applicable to all programs. The results show that the wide range of LPC orders from 10 to 15 suggested by the rules of thumb discussed in Section 2.2 is not universally applicable. Praat's performance was relatively good over LPC order 10 and 12, whereas Multi-Speech was generally consistent from order 10 to 14. The fact that WaveSurfer employs a tracker function meant the performance was relatively unchanged across all LPC orders tested, rendering the rules somewhat redundant.

In terms of translating the findings into guidance or advice for analysts, two aspects of the results are particularly important. Firstly, the differences seen in the formant measurements obtained at the default settings across the three programs show that it cannot be assumed that default settings will give accurate measurements. Whilst the default settings produced generally accurate measurements in some situations, in others they did not. This suggests that the tailoring of settings could lead to more accurate measurements. Secondly, the difference in behaviour across the programs shows that analysts should understand the way in which particular programs operate and appreciate how altering the LPC order may influence the results. Despite the limited scope of the study the results serve to illustrate the variation found across the variables investigated and can raise awareness of it, even if they cannot be used to form more specific guidance. In order to provide more detailed answers to the research questions and provide more specific guidance, the accuracy and behaviour of formant measurements require further study. This is done in the following chapters.

Chapter 4 Formant Measurement Errors From Synthetic Speech

4.1 Introduction

Chapter 3 focused on the effects that varying analysis parameters had on formant measurements across three commonly used speech analysis tools. One of its limitations was that the absolute accuracy of the measurements could not be determined since it was not possible to obtain true reference values. The approach adopted in this chapter overcomes this limitation by using synthetic speech to investigate measurement variability. As the true formant values are specified during the synthesis process, measurement errors can be calculated accurately. A simple source-filter synthesis method is employed with the first and second formants and fundamental frequency as the primary variables, whilst the measurements are made using Praat's formant tool across a range of LPC orders.[2] This method provides results that mainly address the second research question:

RQ 2. How does altering the LPC analysis parameters affect formant measurement accuracy?

The measurements and analysis only concern a single synthetic speaker, so limited insight is gained into the variation in accuracy across speakers, which is the focus of the third research question.

RQ 3. To what extent does the accuracy of LPC formant measurements vary across speakers?

However, some insight is gained from the influence of fundamental frequency on the measurements. The issue of speaker performance is addressed in greater detail in Chapter 5, where measurements from multiple synthetic speakers are examined.

4.2 Motivation for Using Synthetic Speech

It is clear from Chapter 3 that formant values derived from an LPC analysis are dependent on many factors, including the speaker, the software and the chosen analysis parameters. However, the absolute accuracy of the measurements could not be

[2] The approach employed in this chapter has been presented and published with preliminary data (Harrison 2007, 2008a, 2008b, 2008c).

determined due to the methodology employed. In order to calculate the accuracy of a measurement the true value of the quantity being measured must be known. As discussed in Section 1.2, because of the widely spaced harmonics of the glottal sound source, measuring formants by FFT spectra, spectrograms or LPC is problematic. None of these approaches can be considered as satisfactory for obtaining true formant values. Techniques such as x-rays (Fant 1960), MRI scans (Clément et al 2007) or impulse reflectometry (Gray 2005) can be used to determine the resonance characteristics of the vocal tract independently of the speech signal but they are not sufficiently accurate to provide reference values to compare with other methods as they also rely on models and assumptions to obtain formant values. Furthermore, they can only be used when the method is applied simultaneously with the recording of the speech signal.

One way in which true formant values can be known is to specify them during the production of synthetic speech. The measured values can be compared with the 'ground truth' values used in the synthesis process, and the resulting measurement error can then be calculated. Other studies in which this method has been used are discussed in Section 2.1.1.

Using synthetic speech has other advantages. In addition to being able to specify formant centre frequencies and bandwidths, many other speech production variables can also be controlled and manipulated. These include parameters relating to the glottal source, with one of the most important being fundamental frequency. Since speech synthesis is generally performed by computer software, all the synthesis parameters can be specified and controlled precisely. This allows a degree of precision in the speech output that could not be achieved by a human. For instance, as described in Section 4.3, evenly sampled vowel spaces can easily be generated with various fundamental frequencies.

Since the synthesis is conducted within computer software, the process lends itself to being automated, allowing many speech tokens to be generated without analyst intervention. As described in Section 4.3.14, this can also be combined with an automated analysis process allowing the entire procedure to be carried out automatically. This permits many thousands of tokens to be generated and analysed, which would take a considerable amount of time and effort if done manually.

4.3 Methodology

4.3.1 Speech Synthesis Methods

There are three general speech synthesis methods, namely concatenative synthesis, model-based synthesis, and articulatory synthesis. Concatenative synthesis entails combining strings of short segments of pre-recorded speech to produce the required speech output. Model-based synthesis techniques generally rely on the source-filter model of speech production, where a sound signal representing the vocal sound source is passed through a filter that reflects the spectral characteristics of the vocal tract, resulting in the speech signal (Klatt and Klatt, 1990). Articulatory synthesis involves the construction of a mathematical model of the vocal tract based on the acoustic properties and locations of the articulators within it. Then the airflow through it is modelled to produce a speech signal at the lips.

Of these three approaches the model-based method is the most suitable for investigating formant measurement errors as it relies on the assumption that the vocal tract filter is independent of the source, and the filter can be constructed from specified resonance or formant frequencies. This is the method adopted by other studies that use synthetic speech to investigate formant measurement errors. The following sections describe the implementation of this method.

4.3.2 Praat's Source-Filter Synthesiser

The specific source-filter synthesiser used in this chapter is a relatively simple all-pole cascade synthesiser that can be implemented easily in Praat and is described in the manual (Boersma 2001). It was chosen for several reasons. Firstly, this method most closely aligns with the assumptions of the speech model on which LPC analysis is based. Therefore, this represents a best-case scenario for an LPC-based formant measurement system, and it is assumed that such measurement methods will achieve their best performance with this type of synthetic material. Secondly, this implementation allows the relevant parameters to be specified directly and the synthesis process can be controlled easily through Praat's native scripting language. Also, in this implementation the number of required parameters is relatively small, reducing the number of potential variables and allowing the study to be relatively constrained. Since Praat was the software used both to synthesise the speech and measure the formants, both steps could be integrated in a single script.

The synthesiser does not exist as a single function within Praat but uses several different standalone functions. The process of combining these functions and the various options are described within the software's manual (Boersma, 2001). The following sections discuss the different stages of the process and the settings used.

4.3.2.1 Generating the Glottal Source

The first stage of the synthesis process is to generate a glottal source signal. Praat has the capability to allow a high degree of control over the glottal source signal generation in order to replicate, to some degree, the diversity found in real speech. However, in the first instance, a simplified representation of the signal, known as a pulse train, was used. It consists of a series of pulses: sounds with very short onset, duration and offset, which represent the sound made by the vocal folds as they open and close during phonation. A more complex model and realistic representations are employed in Chapter 5.

To generate the pulse train signal within Praat, a PitchTier object with a defined duration is first created. This is effectively a container for pitch points, which represent a pitch contour. Each pitch point is defined by a time and a fundamental frequency value. If a single point is added to the tier then the entire contour over the duration of the PitchTier is flat. If multiple points are added at different times and frequencies then the contour is dynamic.

The next step is to use the pitch contour information in the PitchTier to generate a sound. In this instance the 'To Sound (pulse train)…' command is used, which creates a Sound Object containing a pulse train with the fundamental frequency contour and duration defined by the PitchTier. Several parameters must be specified, including the sampling frequency of the sound to be generated. A sample rate of 44.1 kHz was used for all tokens. The remaining parameters are of little relevance in the configuration being used and were kept at their default values.

The waveform shown in Figure 4.1 is a pulse train generated using this method, with a fundamental frequency of 100 Hz at a sample rate of 44,100 Hz. The period of the waveform, i.e. the time between each pulse, is 10 milliseconds.

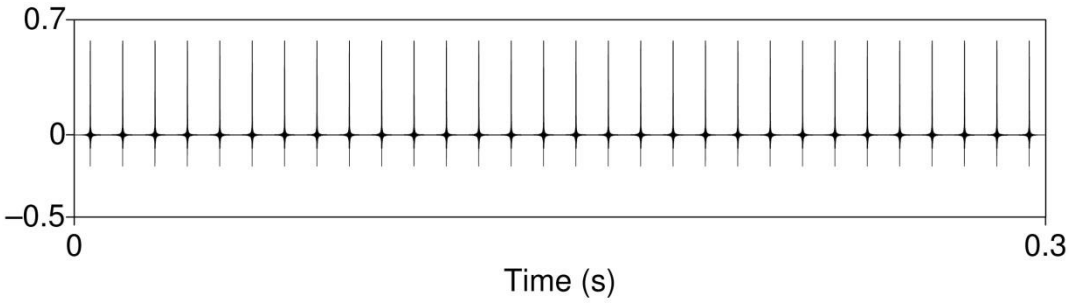

Figure 4.1 An example of the pulse train waveform used to produce synthetic speech, shown with a fundamental frequency of 100 Hz.

The frequency spectra of pulse trains generated using this method contain the fundamental frequency plus harmonics at all multiples of it. There is no roll-off in the amplitude of the harmonics as the frequency increases so each harmonic has the same amplitude as the fundamental. The pulse train can be described as a buzzing sound with prominent higher frequencies.

The spectrum of the pulse train in this form is not a particularly good approximation of that of real glottal source signals. The amplitude roll-off characteristics of normal glottal source signals are generally accepted to be –12 dB per octave (Klatt and Klatt, 1990), meaning that the amplitude of the harmonics decreases by 12 dB for every doubling in frequency. However, as the sound waves pass the lips and leave the mouth the change in acoustic impedance results in a boost to the higher frequencies of the order of +6 dB per octave. In order to replicate these effects within the synthesis process a –12 dB per octave filter could be applied to the pulse train before it is subject to the vocal tract filter and then a +6 dB per octave filter can be applied after the vocal tract filter. However, given the assumption of linearity of the source-filter model, these effects are often combined as a single –6 dB per octave filter which can be applied to the glottal pulse signal before it is filtered by the vocal tract filter. The resulting speech output signal then has an overall spectral slope of –6 dB per octave.

The –6 dB per octave filter was applied using Praat's 'De-emphasize' filter command. The only parameter that can be specified is the frequency above which the filter will be applied. The setting used was the default value of 50 Hz.

4.3.2.2 Specifying the Vocal Tract Filter Parameters

The second stage of the synthesis process is to specify the characteristics of the vocal tract filter. This information is stored as formant points, which are within a FormantGrid

container object with a specified duration. Each point is defined by a time, a formant number, a centre frequency and a bandwidth. This information represents a formant contour across the duration of the FormantGrid. Like the PitchTier, if points exist at different frequencies at different times then the contour is dynamic. The specified centre frequency and bandwidth values determine the location of the poles that define the vocal tract filter.

4.3.2.3 Filtering the Source

The third and final stage of the synthesis process is to filter the glottal source signal to produce the speech output. This simply involves selecting both the Sound object containing the –6 dB per octave filtered pulse train and the FormantGrid object with the filter parameters specified as formant centre frequencies and bandwidths, then executing the 'Filter' command. There are no parameters or options to specify.

Praat generates several IIR (infinite impulse response) filters, one for each formant, whose properties are determined by the centre frequency and bandwidth values provided. The glottal source signal is then filtered by each one in turn in a cascade process. The final output signal is the synthesised speech.

A conceptual representation of the source-filter speech production and synthesis process is shown in Figure 1.1. It shows an idealised glottal source spectrum, the frequency response of the vocal tract filter and the spectrum of the resulting speech signal.

4.3.3 Synthesis Variables & Parameters

The simple source-filter synthesis method and pulse train glottal source within Praat have a limited number of parameters that can be specified, which limits the possible variables. The three independent variables chosen were fundamental frequency, and the first and second formant centre frequencies. For the study to be relevant the specified values must, as closely as possible, reflect real speech. In the following sections the specific values for each of the relevant parameters is presented as well as the reasons for selecting them.

4.3.4 Vowel Variability & Duration

The vowel synthesis process in Praat allows both fundamental frequency and formants to vary with time. In order to restrict the study, all the vowel tokens were generated with

a constant fundamental frequency and formant frequencies. They were therefore all monotone monophthongs.

The duration of each synthesised token must be specified as part of the synthesis process. A duration of 300 ms was chosen so that a sufficiently large number of glottal pulses would be included in each generated token.

4.3.5 Fundamental Frequency

Previous studies have shown that formant measurement errors are influenced by fundamental frequency (Atal and Schroeder, 1974, Vallabha and Tuller, 2002). The fundamental frequency of the glottal source pulse train was therefore used as a variable for this study. No other aspect of the glottal source was varied, but in Chapter 5 other parameters are considered.

It was decided that all vowel tokens would be generated with fundamental frequency values ranging from 70 to 190 Hz at 5 Hz intervals. This covers a range of frequencies that could be readily produced by a typical adult male speaker (Baken and Orlikoff 2000, p. 175, 188). In order to constrain the study, the frequency range was not extended to cover the higher frequencies typically produced by women and children. Also, it is known that such higher fundamental frequencies tend to make measuring formants more problematic (Traunmüller and Eriksson, 1997) as there is relatively less spectral information in the speech signal due to the greater spacing of the harmonics.

4.3.6 Vowel Qualities

As previously discussed in Section 1.3.1, the vocal tract is capable of producing a wide range of different vowel qualities. In a descriptive framework, vowels are often labelled or categorised according to the two main parameters of height and frontness, which describe the relative position of the tongue body within the oral cavity. The first and second formant frequencies are relatively well correlated with vowel height and frontness respectively and these two measured values alone are often used to characterise vowel realisations (Fant, 1960).

The first and second formant frequencies therefore provide a convenient way to define a large set of vowel qualities. It is for this reason that the first and second formant frequencies were chosen to define the synthetic vowel realisations used in this study.

The choice of specific values and the associated higher formants and bandwidth values are discussed below. The formant centre frequency and bandwidth values, used to generate the vocal tract filter, allow the measurement error to be calculated. In order to differentiate them from the measured formant values they will be referred to as the 'specified formants' or 'specified values'.

In order to constrain the study and utilise a simple speech production model, other aspects of vowel quality, such as nasalisation and roundedness, were not considered.

4.3.7 F1~F2 Vowel Space

The range of F1 and F2 values used to define the vowel space were selected to be typical of an adult man. The specific values were obtained from a vowel perception study (Nearey, 1989) where they were used to generate synthesised vowels, and represented the 'baseline' condition. They are based on average male values from Peterson and Barney (1952) and Fant (1973). The F1 values ranged from 250 to 700 Hz, whilst F2 ranged from 750 to 2250 Hz. Constraints were applied to the F1~F2 pairs in order to remove certain combinations that fall outside of the normal vowel quadrilateral. Whilst these are not explicitly stated in Nearey (1989), they must have been applied as the plots showing the vowel space have certain combinations removed (1989, p.2096, Figs 1 & 2). The constraints which were applied in the current study can be represented mathematically as follows:

> [1] F1 + F2 <= 2500 Hz – removes low front vowels (lower left corner of vowel space)
>
> [2] F2 – F1 >= 350 Hz – removes low back vowels (lower right corner of vowel space)

The resolution of the vowel space was chosen to be 10 Hz for F1 and 20 Hz for F2. This produced a total of 2,858 F1~F2 combinations or vowel tokens. Other resolutions were tried initially but these were found to be a reasonable compromise between processing time and detail within the vowel space.

The resulting F1~F2 vowel space is shown in Figure 4.2. The directions of the axes have been reversed and their positions swapped in order for the orientation of the vowel space to replicate the representation of the vowel space or vowel quadrilateral commonly used in phonetics and other areas of speech research.

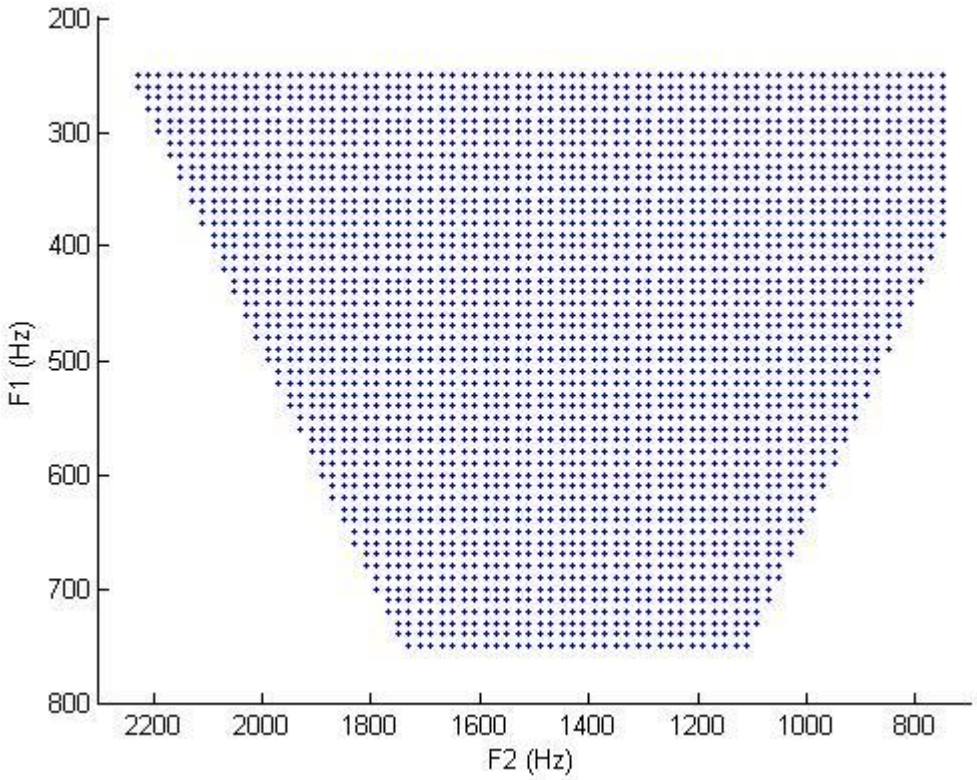

Figure 4.2 Arrangement of the 2,858 synthetic vowel tokens over the F1~F2 vowel space.

4.3.8 F3 Calculation

Broad and Wakita (1977) found that based on 778 steady-state tokens of 30 vowels from one female speaker, the measured F3 values were distributed in such a way that they could be calculated with reasonable accuracy from the corresponding F1 and F2 values. The F3 values existed in one of two planes that formed a front-back split in the F1~F2 vowel space, shown in Figure 4.3. This same approach was also adopted by Nearey (1989), who again used the data from Peterson & Barney (1952) and Fant (1973) to calculate the coefficients to represent the planes, as well their line of intersection.

In Nearey (1989), the intersection between the planes corresponds to the line:

$$[3]\ F2 = (0.17 \times F1) + 1463$$

If the F2 value in the F1~F2 pair is less than the value calculated by [3] then it is classed as a back vowel; if greater, then it is a front vowel. The following equations are then used to calculate the F3 values for each F1~F2 pair according to whether the vowel is front or back:

[4] $F3_{front}$ = (0.522 x F1) + (1.197 x F2) + 57

[5] $F3_{back}$ = (0.7866 x F1) – (0.365 x F2) + 2341

The resulting F3 values used in the current study are between 1994 Hz and 2862 Hz. Figure 4.3 shows a three dimensional plot of the F1, F2 and F3 combinations used. The two planes on which the F3 values lie are reasonably apparent. The value of F3 is represented by both the colour indicated in the colour bar and the height of the point in the vertical or z axis.

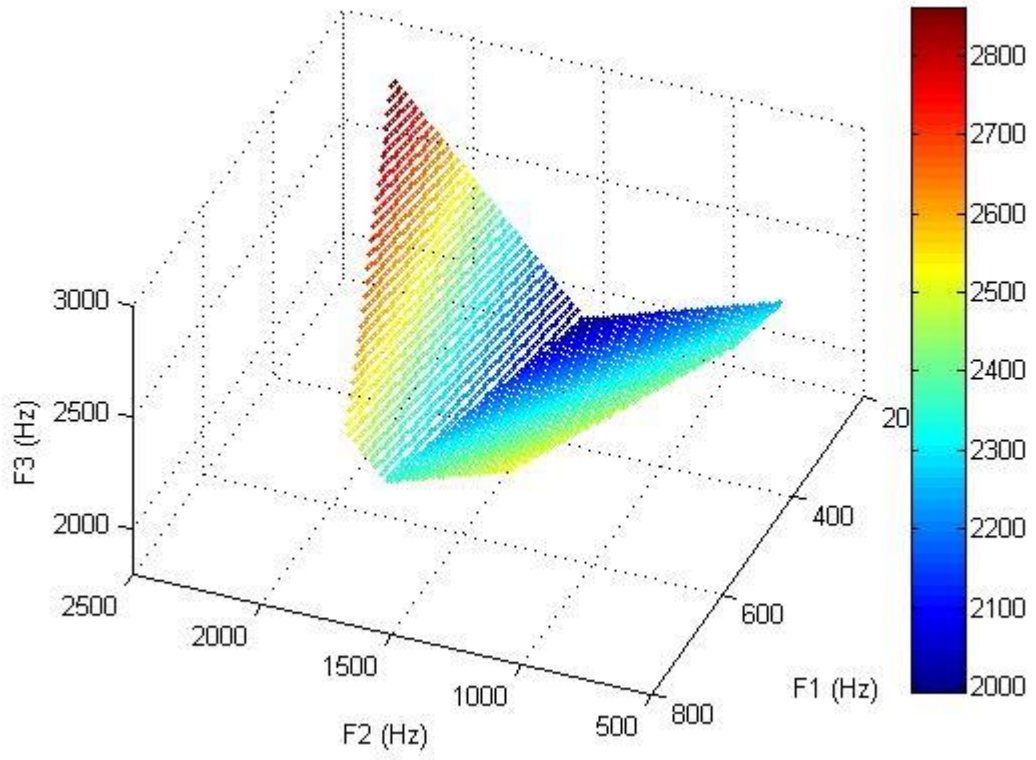

Figure 4.3 Three dimensional representation of the F1~F2~F3 synthetic vowel space with F3 represented by both height in the z axis and colour.

4.3.9 F4 & F5 Determination

The approach adopted by Broad and Wakita (1977) to calculate F3 is not extended to the higher formants F4 and F5, and no comment is made about them. However, in Nearey (1989), F4 and F5 were required for the synthesis process used and their values were held constant at 3500 Hz and 4500 Hz respectively. The justification for these figures was that they correspond approximately to the fourth and fifth resonant frequencies of a uniform tube with a length of 17.5 cm that is open at one end and closed at the other, which is often considered to be the equivalent of an average male

vocal tract in a neutral position. The values used in Nearey (1989) have been adopted here.

4.3.10 Bandwidth Values

Nearey (1989) does not discuss formant bandwidth values. It was therefore necessary to locate an alternative source for this data. A suitable study was Fant (1972), in which empirical data were used to derive a series of formulae for calculating the bandwidths of the first three formants from their centre frequencies. The formulae (numbered 56 to 58 in Fant, 1972) are as follows:

$$[6]\ B_1 = 15(500/F_1)^2 + 20(F_1/500)^{1/2} + 5(F_1/500)^2$$

$$[7]\ B_2 = 22 + 16(F_1/500)^2 + 12000/(F_3-F_2)$$

$$[8]^3\ B_3 = 25(F_1/500)^2 + 4(F_2/500)^2 + 10F_3/(F_4-F_3)$$

Table 4.1 summarises the properties of the calculated bandwidth values.

Formant	Mean (Hz)	Std Dev (Hz)	Min (Hz)	Max (Hz)
F1	46.9	9.7	39.4	75.4
F2	51.8	8.7	33.9	73.3
F3	79.6	24.0	33.6	131.2

Table 4.1 Mean, standard deviation, minimum and maximum values for formant bandwidths calculated using Fant (1972) formulae for F1, F2 and F3.

No formulae are provided in Fant (1972) for the calculation of bandwidths for F4 and F5. Since the F4 and F5 centre frequency values do not vary over the F1~F2 vowel space it was decided to simply select constant bandwidth values as well. These were chosen as 200 Hz and 300 Hz respectively. These values align well with the plot in Hawks and Miller (1995, p. 1343, Fig 1) of average bandwidth values against formant centre frequencies derived from empirical data.

4.3.11 Single Synthetic Speaker

The specific formant and bandwidth values described above form a single set from an effectively limitless number of sets that could have been generated. Altering the range of the F1~F2 space, the method to calculate F3, the values chosen for F4 and F5 or modifying the bandwidth formulae would have created a different set. Each set of

[3] In Fant (1972) there is an error in equation number 58. The final term should contain a division operator as shown in [8] rather than the multiplication operator.

values can be considered as a single 'synthetic speaker'. The remainder of this chapter is concerned with measurement errors for this single synthetic speaker.

4.3.12 Formant Measurement Method

The analysis undertaken in this chapter is primarily concerned with addressing the second research question concerning the effects of altering analysis parameters, and as a consequence only a single program is used to measure formants. The Praat software was chosen as it was found to be the single most used piece of software in the survey of forensic phoneticians described in the previous chapter, it is the main speech analysis tool used in the author's forensic laboratory and its scripting capabilities enable the synthesis and analysis processes to be entirely automated.

The specific function within Praat that was used was 'Sound: To Formant (burg)…'. As noted in Section 3.2.5 previously, this function is not a tracker since it does not track formant values from one analysis frame to the next, nor does it make decisions about whether measurements are likely to correspond to formants or not. It simply carries out an LPC analysis and returns the pole frequencies as the measured formant values with only those below 50 Hz and those 50 Hz or less below the maximum analysis frequency being disregarded as unlikely formant values.

Again, in order to constrain the study, LPC order was chosen as the only analysis parameter that would be varied. In the previous chapter and in other studies (Chandra & Lin, 1974, Vallabha & Tuller, 2002), LPC order has been shown to be an analysis parameter that has a significant influence on measured formant values. The remaining parameters were held at their default values for male speech. These were:

- Time step = 0.00625 s
- Maximum Formant = 5000 Hz
- Window Length = 0.025 s
- Pre-emphasis from = 50 Hz

The range for LPC order was chosen to be from 6 to 20 in steps of 1. For Praat's 'Sound: To Formant (burg)…' function LPC order is not specified directly. Instead, the parameter 'Maximum number of formants' is used and this value is equal to half the

LPC order. As a consequence the 'Maximum number of formants' parameter can be specified in steps of 0.5, e.g. a setting of 4.5 corresponds to an LPC order of 9.

From the perspective of the user Praat's formant measurement process involves executing the 'Sound: To Formant (burg)…' function with the specified analysis parameters on a Sound object containing a synthesised vowel token. Praat then executes several processes to obtain the formant measurements. Firstly, the sound is resampled with a sampling frequency that is twice the specified Maximum Formant value. Pre-emphasis of +6 dB per octave is then applied to the sound above the frequency specified in order to flatten the frequency spectrum of the sound by adjusting for the –6 dB per octave roll-off. The sound is then considered in terms of individual analysis frames with duration and locations determined by the 'Time Step' and 'Window Length' settings. A Gaussian-like window function is applied to each frame to reduce of the effect of the discontinuity in the waveform at the start and end of the frame. An algorithm, developed by Burg (Childers, 1978) (as cited in Boersma, 2002), is then applied to each frame, which calculates the LPC coefficients. The pole frequencies are obtained from the LPC coefficients and are subsequently converted to formant centre frequency and bandwidth values. Any formants with a centre frequency either below 50 Hz or within 50 Hz of the 'Maximum Formant' settings are considered as artefacts of the LPC algorithm rather than true formant values, and they are rejected. The formant centre frequency and bandwidth values for each analysis frame are made available to the user in a Formant Object within Praat's Objects List.

A range of queries can be run on the Formant Object to obtain information about the measurements, including formant values from specific frames and statistical measures. In this study a series of queries were run to obtain the mean centre frequency and mean bandwidth for F1 to F5 from time 0.1 seconds to 0.2 seconds from each 0.3 second token. The average value is obtained rather than a value from a single frame because the analysis frames do not coincide with pitch periods (i.e. it is not a pitch synchronous analysis) and there may be differences in the measurements across frames. Taking the mean reduces any potential effect of this variation.

4.3.13 Calculation of Measurement Error

The final stage in the speech synthesis and measurement process is the calculation of the measurement error for the formant centre frequencies. This simply involved subtracting the specified value from the measured value. This can be expressed as follows:

$$[9]\ F_{error} = F_{measured} - F_{specified}$$

Calculating the error in this way means that if the error value is greater than zero then the measured value is greater than the specified value and if the error value is negative then the measured value is less than the specified value.

The calculated error values are expressed in Hertz, as are the specified formant values and the measured values from Praat. Since formants span a range of frequencies from F1 to the higher formants that is approximately a factor of 10, percentage errors were also calculated to allow the errors to be compared across the formants. This was done using the following formula:

$$[10]\ F_{\%\ error} = ((F_{measured} - F_{specified})/\ F_{specified}) \times 100$$

Another measure which could be used to represent errors is the cent, which is one hundredth of a semitone. It is a logarithm unit which is used to measure the interval between frequencies, most commonly for musical notes. The use of the scale is not widespread within phonetics and is very infrequently used in the field of forensic speech science. Use of the unit in the present study would present two main problems. Firstly, since the measure is not in widespread use by the intended readership of this work, unfamiliarity with results expressed in cents would make their interpretation difficult. Secondly, the results presented in other published research concerning formant measurement errors are generally expressed in Hertz or, less commonly, as percentages. If cents were used then comparisons with these studies would not be possible.

4.3.14 Implementation

Several stages were involved in the calculation of the formant values, the generation of the synthetic speech, the subsequent formant measurement and error calculation. The first of these was to generate the specified formant and bandwidth values. This was done via a single Praat script using the formulae and constant values described above. The script produced a table, as a plain text file, with 2,858 rows, one for each vowel in

the vowel space, containing each token's formant and bandwidth values. The table was generated with empty columns to store the measured values and calculated errors so that all the data for a given LPC order and fundamental frequency would be stored in a single table file.

All the remaining stages were performed by a single script. This consisted of a nested loop structure, where the main body of the script, which performed the synthesis, measurement and logging, would be executed for each combination of fundamental frequency and LPC order. For each combination the script first read the specified formant value table file into Praat. It cycled through each row generating a vowel token with the required fundamental frequency, and then performed the formant analysis at the specified LPC order. The mean centre frequency and bandwidth values were then obtained, the measurement error was calculated, and measurements and error values were inserted in the table. When the script had worked through every row in the table it was saved with a new filename indicating the fundamental frequency and LPC order used to obtain the measurements. The script then returned to the start of the loop and reloaded the original specified formant table and started the process again with a different combination of fundamental frequency and LPC order. With the parameters specified above, over 1 million vowel tokens were synthesised and measured, which took almost 20 hours to complete.

4.4 Analysis

The following sections describe and present the results of the analysis that was conducted on the data generated using the method described above. It is limited to the centre frequencies of the first three formants, as these are the parameters most often considered within forensic casework and in phonetics more generally. All of the analysis was carried out in Matlab (The MathWorks, 2007). This was used as it has powerful data analysis, processing and plotting capabilities. Commands entered in the software can be combined easily in scripts or functions, allowing many of the processes to be automated and easily repeated. The data within each of the text files generated by the Praat synthesis and formant analysis script were imported in to Matlab to allow the analysis to be undertaken.

4.4.1 Error Surface Plots

The initial analysis of the data involved plotting the measurement error for each formant over the entire F1~F2 vowel space for each combination of fundamental frequency and LPC order. This allowed a quick impression to be gained of how the measurement error varied across the vowel space as well as what influence the fundamental frequency and LPC order had.

One way to display this data would have been as three-dimensional scatter plots, with the x and y axes showing the specified F1 and F2 values with the measurement error on the z-axis. Whilst this method worked to some extent, the data could be difficult to interpret, especially in some orientations of the plots, as the spatial relationship between data points was not readily apparent. This occurred because data points could be seen in the gaps between other data points and their spatial proximity could not be easily determined. In view of this problem the method was not used.

To overcome this problem surface plots were generated, where a surface is fitted to the data points. These are created by way of a command that calculates triangles (a process known as Delaunay triangulation; Delaunay 1934) between the data points. These triangles are then plotted and filled in with colour to give a continuous opaque surface. A colourmap is applied to the plot so that the colour at any point on the surface also represents the height of the surface in the z axis. This colourmap is interpolated over the surface so that the transitions between data points are smooth and continuous in all directions. The only problem with this type of representation is that because the surface is opaque certain sections of it can be hidden by other parts. However, within Matlab the surface can be rotated easily so that it can be seen from any angle. This is therefore only a problem when producing a static representation of a plot.

Figure 4.4-Figure 4.6 show the measurement error surface plots for F1 to F3 at a fundamental frequency (F0) of 100 Hz and an LPC order of 8. At this F0 and LPC order the errors are particularly stable and are some of the smallest obtained across all three formants. The orientation of each of the plots is different in order to provide the best overall impression of the surface from a single viewpoint. The range on the z-axis is also different for each plot, as is the range represented by the colourmap, since the range of the errors is different for each formant.

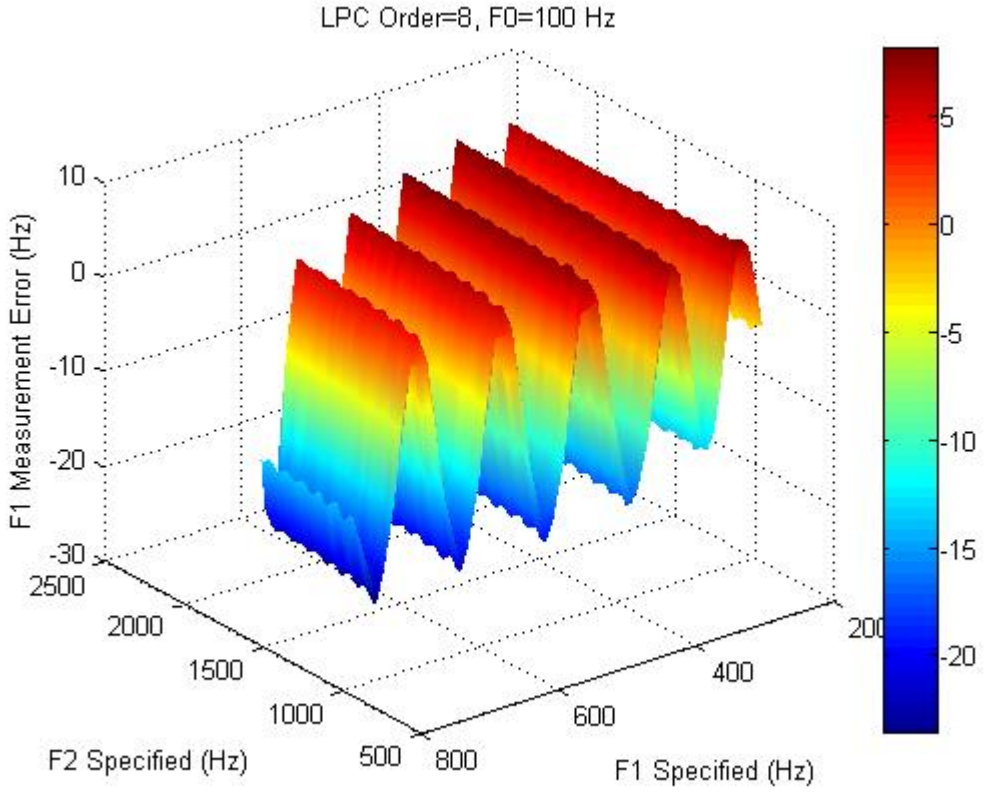

Figure 4.4 Surface plot representing F1 measurement error from synthetic speech with a F0 of 100 Hz measured in Praat with an LPC order of 8.

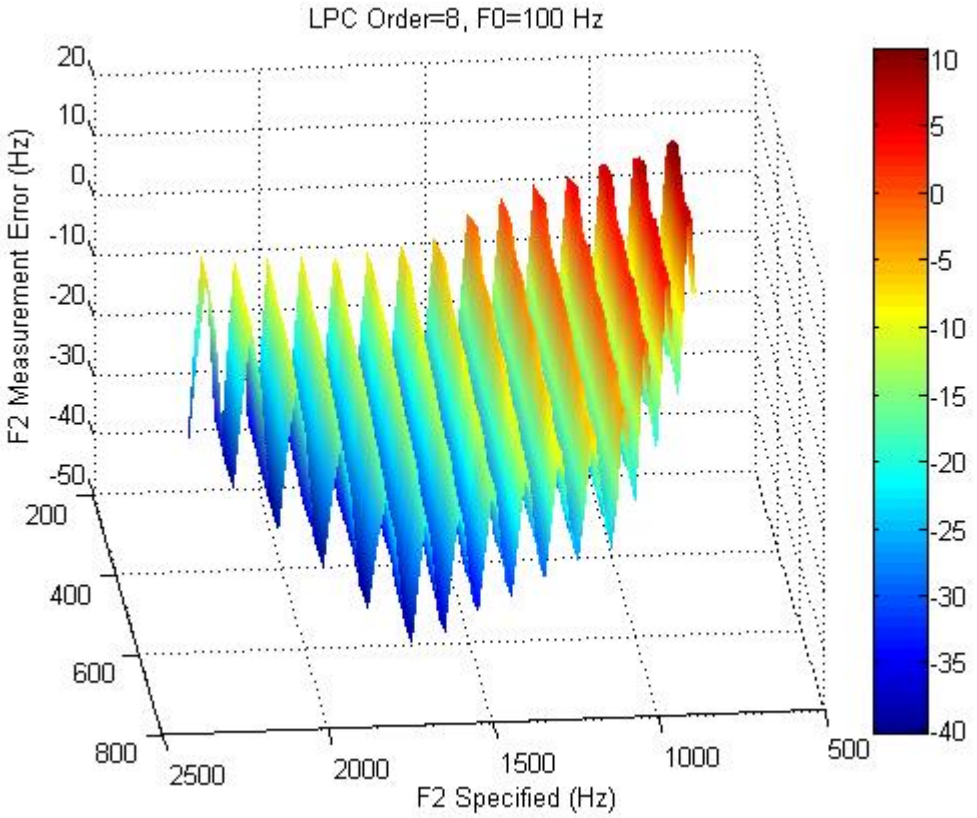

Figure 4.5 Surface plot representing F2 measurement error from synthetic speech with a F0 of 100 Hz measured in Praat with an LPC order of 8.

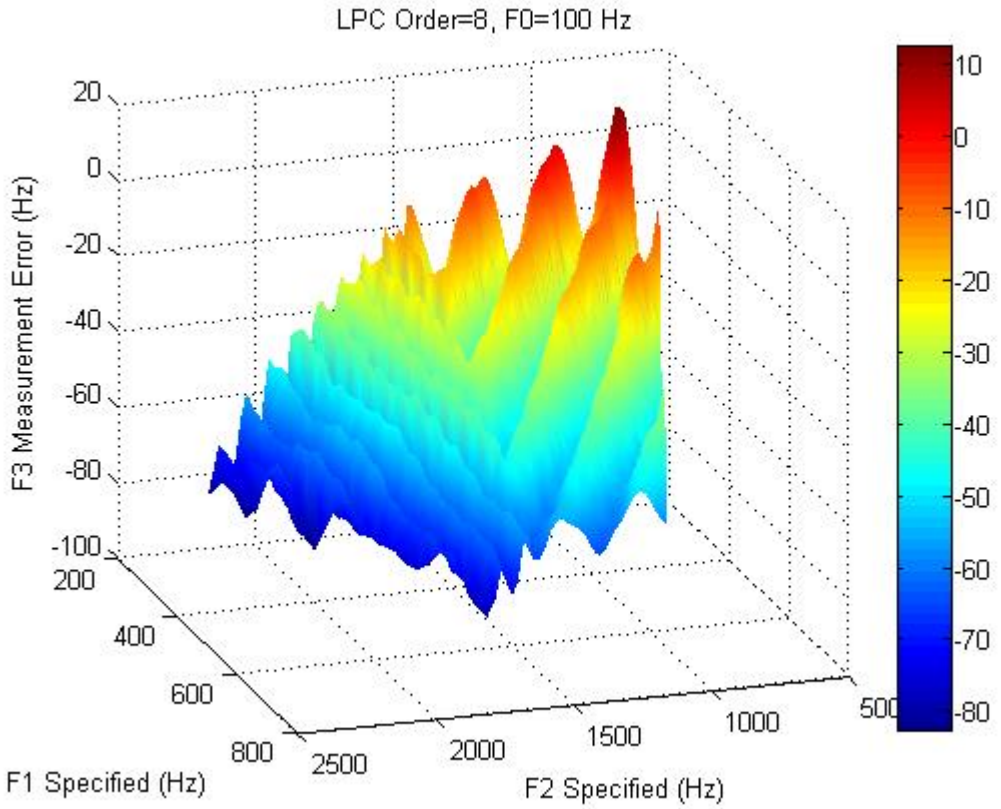

Figure 4.6 Surface plot representing F3 measurement error from synthetic speech with a F0 of 100 Hz measured in Praat with an LPC order of 8.

Perhaps the most obvious feature of all three plots is the cyclic or repetitive nature of the error surfaces. In Figure 4.4 the F1 error surface shows 5 repetitions of a sine wave type shape over the specified F1 values, which cover a range of 500 Hz. Therefore the effective period of each cycle is approximately 100 Hz, which corresponds to the F0 of the synthetic vowel tokens measured. This association is considered in greater detail in Section 4.4.5. There also appears to be some dependency of the error surface on F2 with slight cyclic variation as F2 changes. This is most noticeable at the peaks and valleys. Overall, the range of errors is relatively small at 32 Hz, with a minimum of –24 Hz and a maximum of 8 Hz. A feature which is not particularly clear in this orientation of the plot, although it can be seen as a change in the darkness of the blue in the troughs, is that the range of error variation within a cycle increases as F1 increases.

The F2 measurement error surface, shown in Figure 4.5, is also cyclic. The effective period appears to be shorter than for F1. However, this is a consequence of the range of the F2 axis being greater than the F1 axis. The specified F2 values cover a range of 1500 Hz and the number of cycles shown in the plot is 15. Again, the period of variation corresponds to the F0. There is also some variation in the F1 direction and again this is relatively small. What is more apparent in this plot is the trend for the measurement

errors to become larger in a negative direction as F2 increases. The overall range is small at 51 Hz, with a minimum of –40 Hz and a maximum of 11 Hz.

Figure 4.6 shows that the F3 measurement error surface is somewhat different in that there are two distinct regions, which both exhibit cyclic behaviour. These correspond to the two planes that make up the specified F3 values (see Section 4.3.8). The dependence of the specified F3 values on F1 and F2 make the periodicity in the two regions of this plot somewhat harder to interpret. However, if the error values are plotted against the specified F3 values then it becomes clear that the cyclic dependency on the specified values is the same as for the other formants, i.e. it is dependent on the fundamental frequency. Again, the errors overall are relatively small, with a range of 95 Hz, from a minimum of –83 Hz, to a maximum of 12 Hz.

4.4.1.1 Animated Error Surfaces

The error surface plots only show the behaviour at a single LPC order at one fundamental frequency. The measurement errors were calculated for each formant of each vowel token across a total of 15 LPC orders and 25 fundamental frequencies. This gave a total of 375 error surface plots for each formant. It would obviously be impractical to generate and view each one individually. However, one way in which an overall impression of the data was gained was by making animated error surfaces to see the effects of varying either LPC order or fundamental frequency. For example, the LPC order would be held constant whilst the error surface for each fundamental frequency was displayed in turn. The process effectively added a fourth dimension to the plots.

From examining these animations two main trends became apparent. The first was that as fundamental frequency increased the effective period of the repetition in the error surfaces increased, confirming the initial impression that the cyclic property is linked to fundamental frequency. Secondly, as LPC order increased, the magnitude of the errors also increased in a negative direction, showing that the measurements were under-estimating the true formant values. This was most apparent for F2 and F3. Also, the increase in errors was not always uniform across the vowel space, with certain localised regions showing much better performance than others.

4.4.2 Distribution of Errors

To provide an overall impression of the variation in performance across LPC orders, the errors from all fundamental frequencies for each LPC order were combined. This arrangement of the data more closely reflects the realities of human speech since it occurs across a range of fundamental frequencies. A convenient way to summarise the behaviour of the errors across LPC orders was to generate box plots for each formant. Figure 4.7 to Figure 4.9 show boxplots generated from the errors across all fundamental frequencies for F1 to F3 respectively. At each LPC order, the horizontal red line represents the median value, the lower and upper edges of the blue box are the 25th and 75th percentile, and the black whiskers extend to the limits of the data that are not considered outliers. Outliers have been determined as values that fall outside a range defined as being 1.5 times the interquartile range above the 75th percentile and 1.5 times the interquartile range below the 25th percentile, where the interquartile range is the difference between the 75th and the 25th percentile (Tukey, 1977). If the data were normally distributed then these limits would encompass 99.3% of the values. All of the outliers are shown in the plots as red crosses. Since the size of the errors cover a large range, each plot includes a detailed view of the region where the errors were smallest.

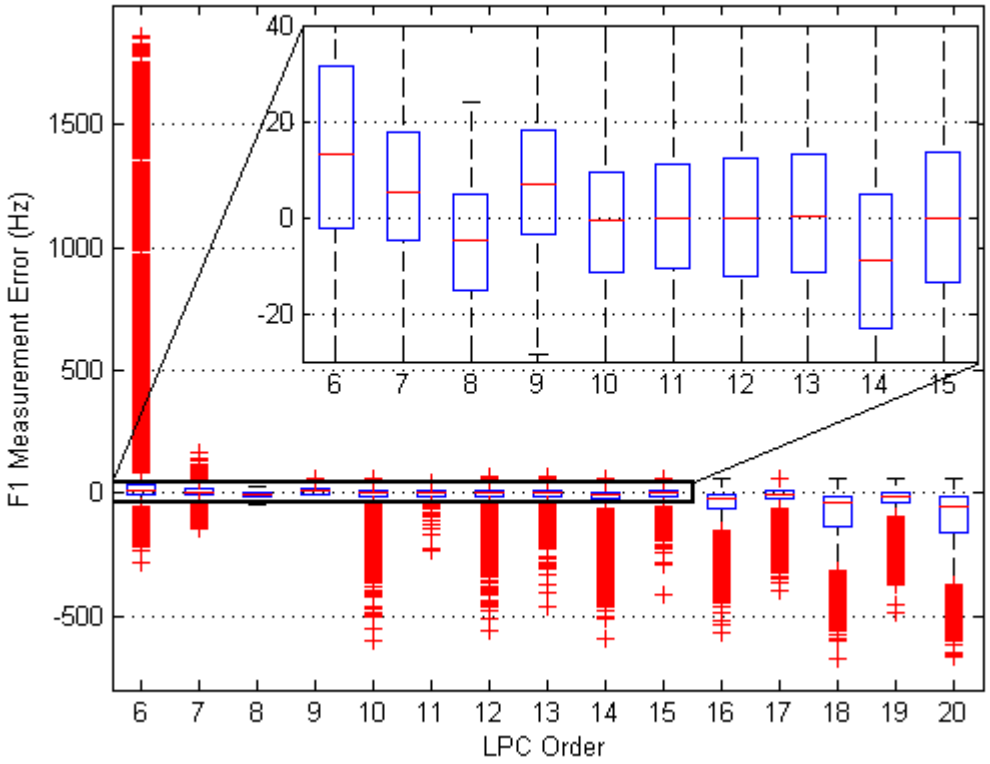

Figure 4.7 Boxplot showing the distribution and variation of F1 measurement errors from synthetic speech for all fundamental frequencies across LPC order.

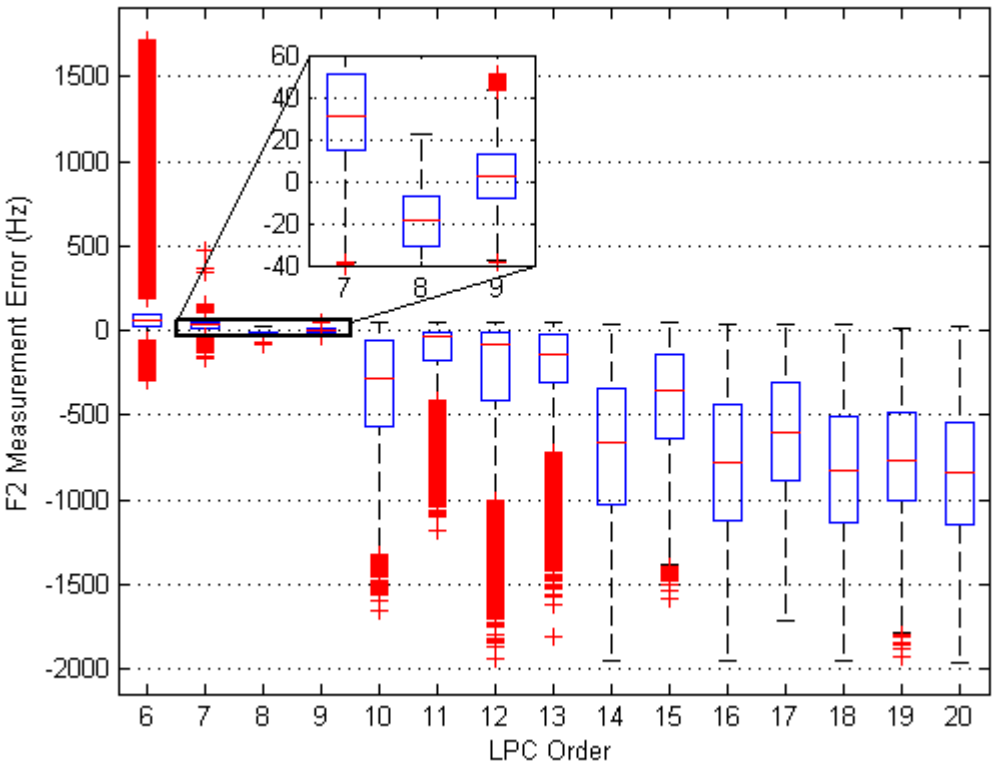

Figure 4.8 Boxplot showing the distribution and variation of F2 measurement errors from synthetic speech for all fundamental frequencies across LPC order.

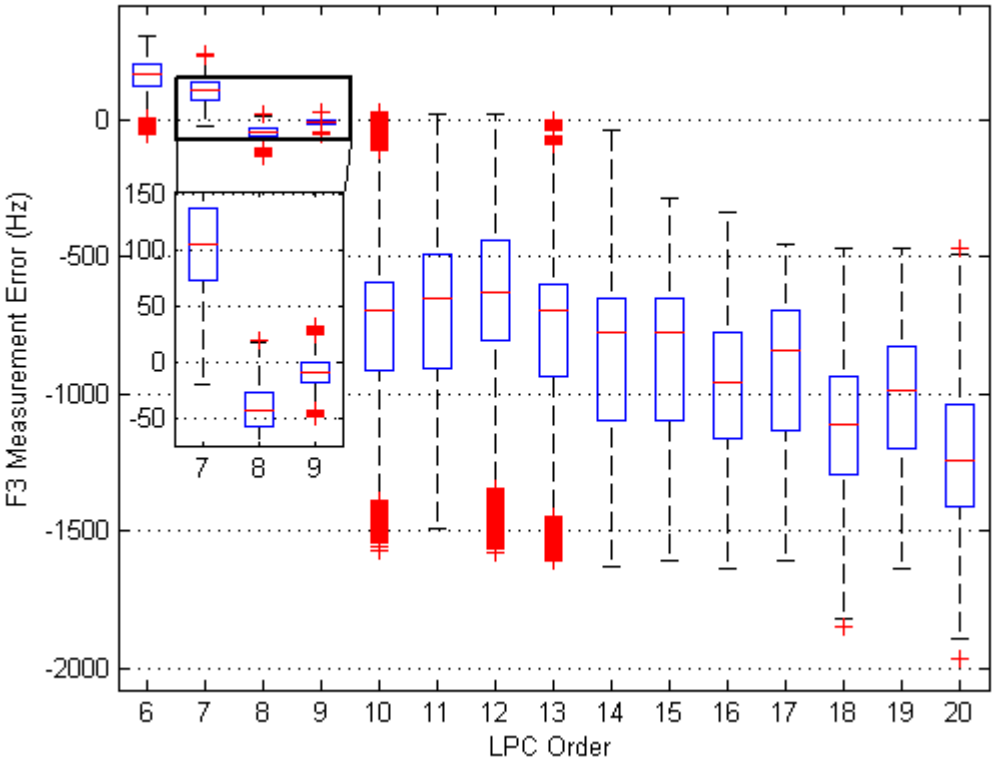

Figure 4.9 Boxplot showing the distribution and variation of F3 measurement errors from synthetic speech for all fundamental frequencies across LPC order.

The boxplots reveal a significant amount of information about the data and the behaviour of Praat's formant measuring tool. In Figure 4.7, the central tendency of the F1 errors, represented by the median, remains relatively similar across LPC orders 6 to 15, as does the interquartile range. Above LPC order 15, the even numbered orders have larger negative errors and the interquartile range increases. Also, the distributions become negatively skewed. The odd orders 17 and 19 are more similar to the lower orders. The results from LPC order 6 show a large number of outliers that extend well above the central range of results and a number below. At order 7 the number and range of outliers has decreased and at order 8 there are none. As the order increases to 10 and above, the majority of the outliers are below the central band of results. The most accurate and least variable results occur at LPC orders 8 and 9. Overall, at the lowest LPC order the F1 measurements tend to be overestimates. As the order increases the distribution of the errors centre around 0 Hz and at higher orders the measurements tend to be underestimates. Given the way that the formants are defined by Praat, i.e. the lowest frequency pole is always F1, and the way that extra peaks appear in the spectrum as the LPC order increases, this behaviour is expected and it aligns with the patterns observed in the previous chapter.

In the case of the results for F2 and F3 shown in Figure 4.8 and Figure 4.9 the effect of the LPC order on the errors is much greater than for F1. For F2, the results at LPC order 6 are not dissimilar to those for F1. At orders 7 and 8 the variability of the F2 errors is reduced, as are the number and dispersion of outliers. The most accurate and least variable measurements exist at LPC order 9. At LPC order 10 and above, the magnitude and the variation in the errors become much greater, and the measurements are generally underestimates. This behaviour can again be accounted for by the way Praat extracts formant measurements and the behaviour of the LPC analysis. As the order increases and more poles/peaks appear in the LPC model, the second pole/peak, which Praat assumes corresponds to the second formant, often no longer corresponds to the second formant and instead lies somewhere below it. Similar behaviour is seen in the results from F3 in Figure 4.9 and it can be explained by the same mechanism.

In addition to generating boxplots, the distributions of errors were also examined via histograms. For F1, the form of the distributions aligned very well with the impression given by the boxplots. Ignoring the outliers, the distributions were approximately symmetric, with negative skewing only occurring at order 14 and above. The

histograms of the errors for F2 revealed behaviour that was somewhat harder to interpret in the boxplots. At orders 7, 8 and 9 the distributions are roughly symmetrical and similar to those for F1 at the same orders. At higher orders, from 10 to 13, the distribution of the F2 errors is very different. This is also shown by the configuration of the boxplots at these orders as the 75th percentile at the top of the box lies very close to the upper whisker. Figure 4.10 shows the distribution of F2 errors at LPC order 11. The distribution is very negatively skewed and a tall narrow peak occurs just below 0 Hz.

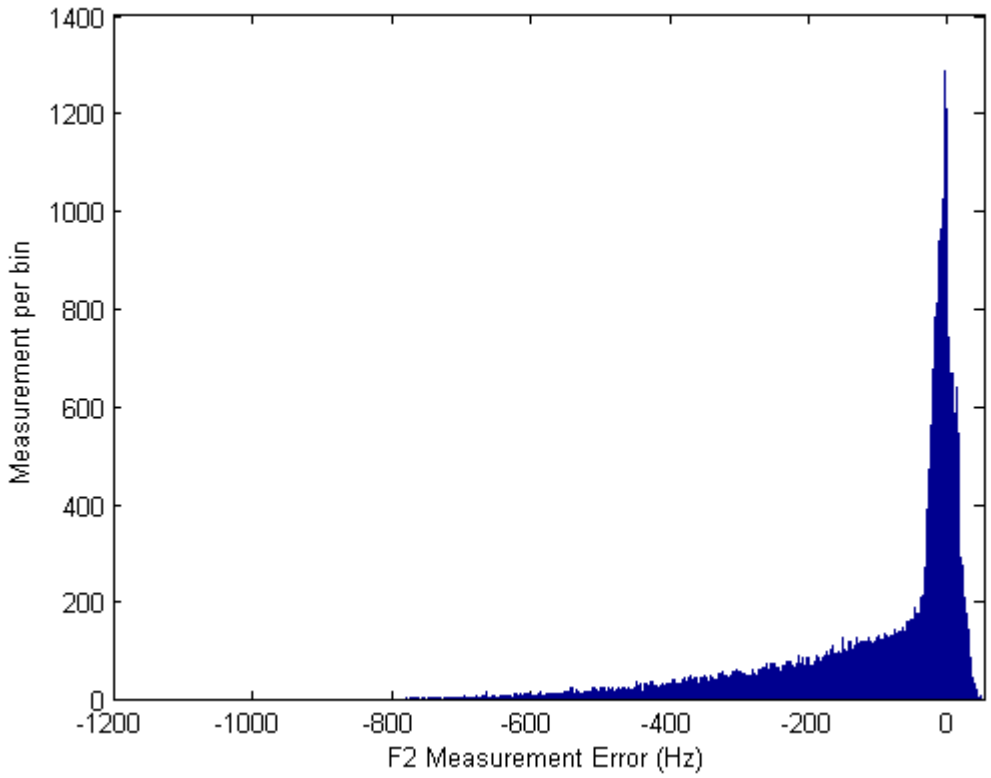

Figure 4.10 Histogram of F2 errors at LPC order 11 from synthetic speech across all fundamental frequencies, with a bin width of 1 Hz.

The median F2 error at LPC order 11 is -33.1 Hz, which corresponds to the transition point from the narrow peak to the gradual slope in Figure 4.10. This means that 50% of the measurements are very close to the true value. The distributions of the F2 errors at LPC order 10, 12 and 13 are similar in form to the distribution at order 11. At LPC orders 14 and above the distributions are different again and are roughly symmetric or gently skewed, as shown in the boxplots. This sudden change in the behaviour of the measurements as the LPC order increases corresponds to a large proportion of the F2 measurements no longer corresponding to the true F2 value.

The distributions of F3 errors at LPC orders 7, 8 and 9, which produce the most accurate measurements, are also approximately symmetrical and consist of a tall narrow peak. At order 10 and above the behaviour changes but it is different from that seen for F2. Rather than retaining a tall narrow peak around 0 Hz the entire distribution becomes relatively broad and a tall peak occurs around -600 Hz. A smaller narrow peak is also present just below 0 Hz. Figure 4.11 shows the distribution of F3 errors at LPC order 11.

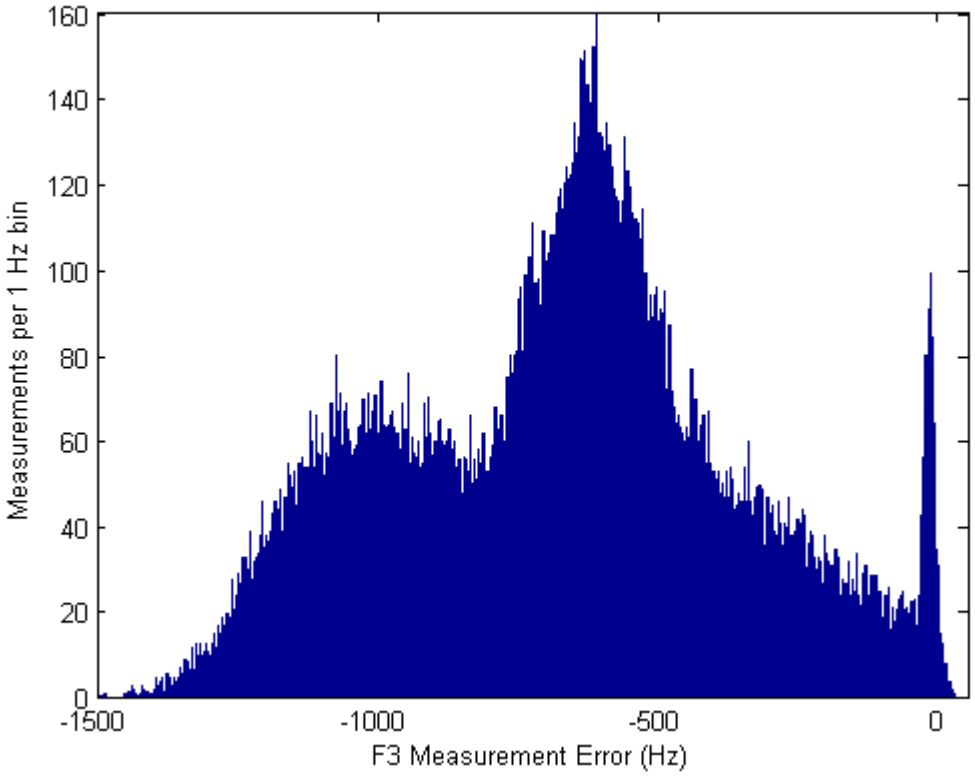

Figure 4.11 Histogram of F3 errors at LPC order 11 from synthetic speech across all fundamental frequencies, with a bin width of 1 Hz.

In Figure 4.11 the narrow peak just below 0 Hz represents the F3 measurements that are relatively accurate. Most of the measurements in the larger central peak and some in the area to the left are F3 measurements that correspond with the true F2. This is a consequence of how Praat assigns pole frequencies to formant measurements. Since extra peaks appear in the LPC spectrum at higher LPC orders the third pole often corresponds with F2 and a higher pole aligns with F3. This effect is not seen for the F2 errors as it is less common for the second pole to align with F1 since the first pole rarely occurs below the true F1. This type of behaviour could be considered as a formant numbering error, rather than a measurement error, since the LPC analysis has accurately

represented the F2 peak but it has not been assigned the correct label by the software. Formant numbering errors are considered further in Section 4.4.7.

The distributions of F3 errors are similar up to LPC order 17 and all show a peak around -600 Hz. Above this order the distributions become approximately symmetrical or skewed and the peak is no longer present.

4.4.3 Mean Error & Mean Absolute Error

The majority of studies on formant measurement errors report performance as either mean error or mean absolute error (ignoring the sign or the direction of the error). Mean absolute error is a useful measure as it removes the potential effect of positive and negative errors cancelling each other out and gives a better sense of the overall size of the errors, whether they are under or over estimates. Mean or mean absolute errors may also be expressed as a percentage, which is helpful when comparing the magnitude of errors across formants. Summary statistics were calculated for the results discussed above to compare them with the results presented in the following chapters as well as other published studies. They were calculated for the errors from all fundamental frequencies at LPC order 9, as this setting produced the most accurate measurements. The standard deviation was also calculated to provide a measure of the variability of the errors. The mean and standard deviation are valid summary statistics at this LPC order as the examinations of the distribution of errors showed them to be sufficiently symmetric. At this LPC order the mean or median do not reflect the overall magnitude of the errors since the centre of the distributions are located near to 0 Hz. The summary statistics were also calculated for the combined errors from the three formants to give an overall measure of performance. The values are shown in Table 4.2.

		F1	**F2**	**F3**	**F123**
Mean Error	(Hz)	7.48	2.49	-9.58	0.13
	(%)	1.73	0.26	-0.40	0.53
Mean Absolute Error	(Hz)	13.04	11.97	13.16	12.72
	(%)	3.00	0.91	0.56	1.49
Standard Deviation	(Hz)	14.54	14.52	12.61	15.65
	(%)	3.43	1.15	0.53	2.29

Table 4.2 Summary statistics for each formant and three formants combined for measurement errors from synthetic speech for all fundamental frequencies at LPC order 9.

Table 4.2 demonstrates the differences between the mean error and the mean absolute error. The mean error varies between the three formants, showing a different central tendency for each, whereas the mean absolute error is much more similar, meaning overall the errors are of comparable magnitude. The standard deviation shows little different for the three formants indicating a similar degree of variation of the errors. In percentage terms the performance is worst for F1 as the errors are proportionally larger in comparison to the true values.

4.4.4 F0 Influence on Errors

The analyses in the previous two sections did not consider the influence of fundamental frequency on the formant measurement errors. However, it is clear from the examinations of the error surface plots in Section 4.4.1 that fundamental frequency does affect the pattern of the errors. In order to quantify this effect, summary statistics were calculated for the errors for each formant at each fundamental frequency. These were restricted to LPC orders 7, 8 and 9 as these were the orders that produced the most accurate measurements across all three formants. The statistical measure which is most revealing is the mean absolute error. This is plotted against fundamental frequency in Figure 4.12 for all three formants with F1 represented as crosses, F2 as circles and F3 as stars, at LPC orders 7 (red points), 8 (green points) and 9 (blue points).

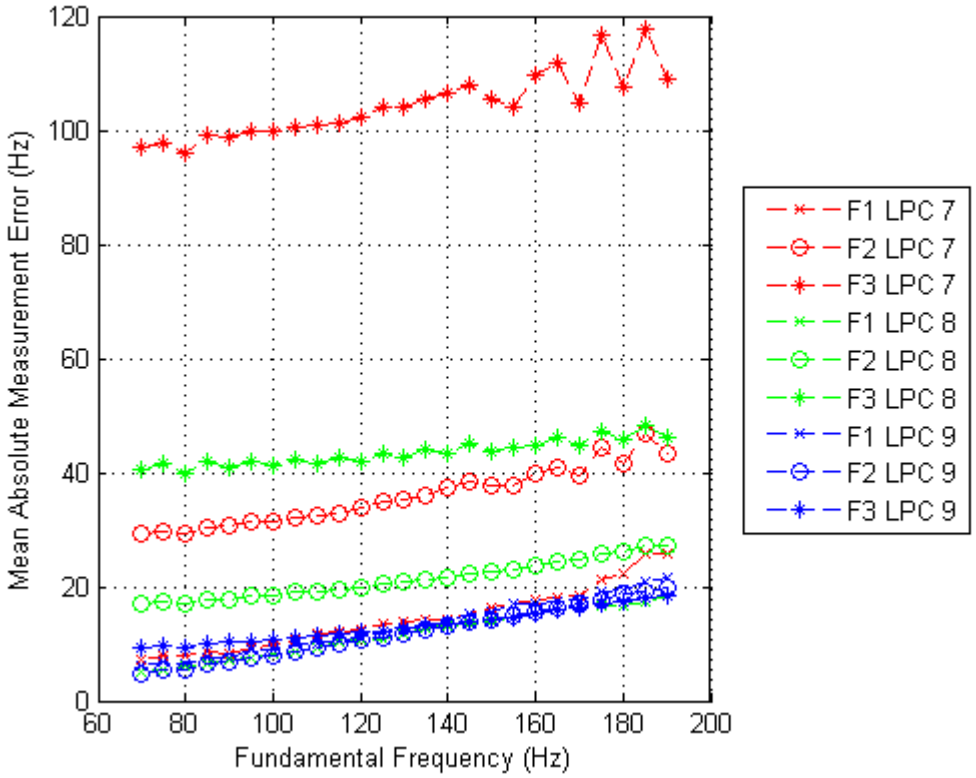

Figure 4.12 Mean absolute error from synthetic speech across fundamental frequency for LPC orders 7 (red), 8 (green) and 9 (blue) for F1 (crosses), F2 (circles) and F3 (stars).

The results in Figure 4.12 show that there is a clear relationship between fundamental frequency and mean absolute error across all three formants at the LPC orders examined. For each formant and LPC order the errors are smallest at the lowest fundamental frequency and increase linearly as the fundamental increases. The slope of each line is similar, showing that the effect is consistent across formants and LPC orders. The only deviation from this pattern is for LPC order 7 at the higher fundamental frequencies. The plot also shows the difference in overall performance between the three LPC orders and the three formants. The standard deviation of the errors was also calculated and plotted across fundamental frequency and this showed the same pattern. The standard deviation increased linearly for all formants as the fundamental frequency increased, showing that the variation or spread of the errors is greatest at higher fundamentals.

4.4.5 F0 Influence on Individual Vowels

An alternative way to examine the effect of fundamental frequency on measurement error is to consider how the error varies for specific vowel tokens as fundamental frequency changes. Figure 4.13 to Figure 4.15 show the measurement error for F1, F2

and F3 for a single vowel at LPC order 9 as fundamental frequency increases. The example lies centrally in the vowel space and has the specified values of F1 = 500 Hz, F2 = 1510 Hz and F3 = 2183 Hz.

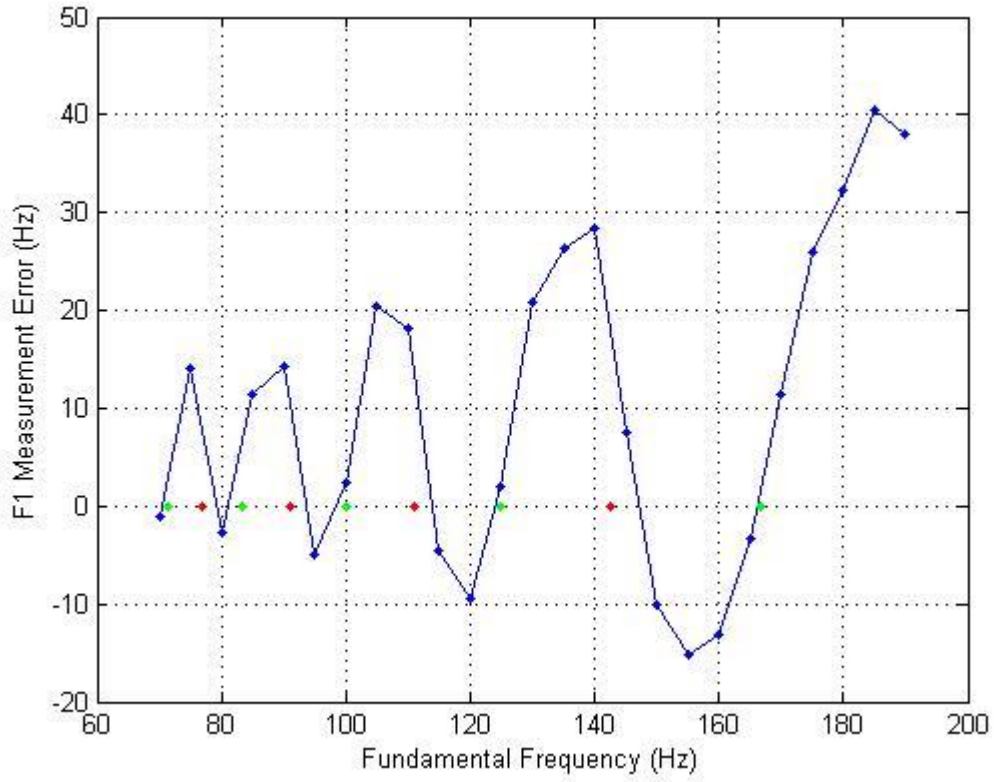

Figure 4.13 F1 measurement error across fundamental frequency for specified F1 formant frequency of 500 Hz at LPC order 9 from synthetic speech. Green dots represent fundamental frequencies that are integer multiples of 500 Hz and red dots represent ones that are half integer multiples.

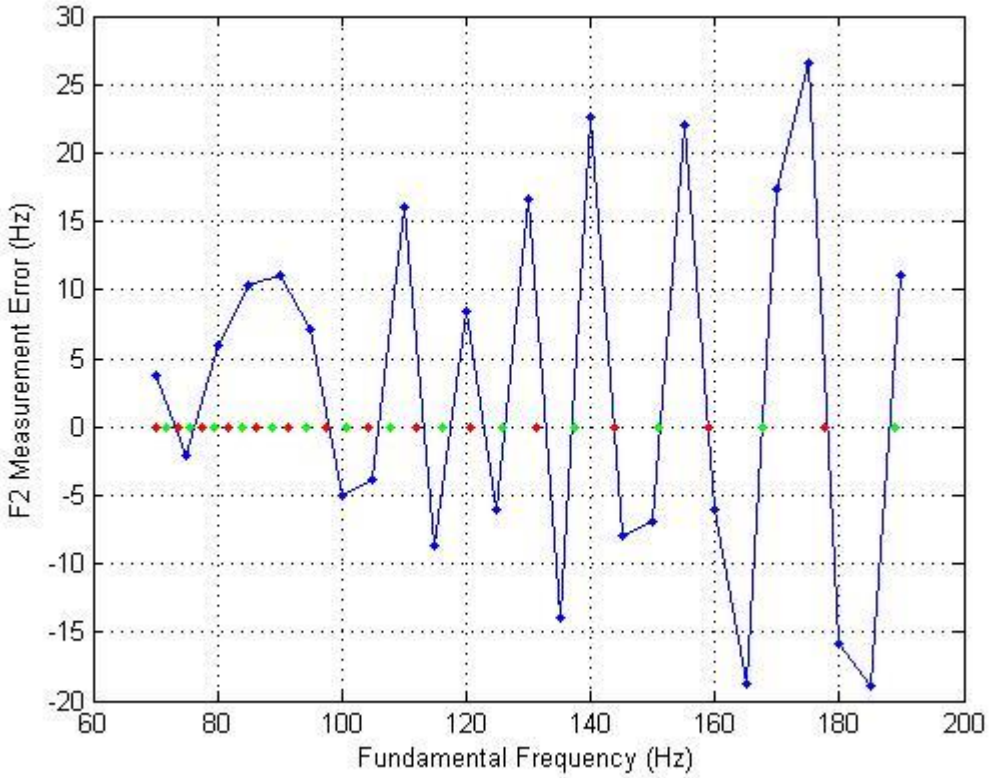

Figure 4.14 F2 measurement error across fundamental frequency for specified F2 formant frequency of 1510 Hz at LPC order 9 from synthetic speech. Green dots represent fundamental frequencies that are integer multiples of 1510 Hz and red dots represent ones that are half integer multiples.

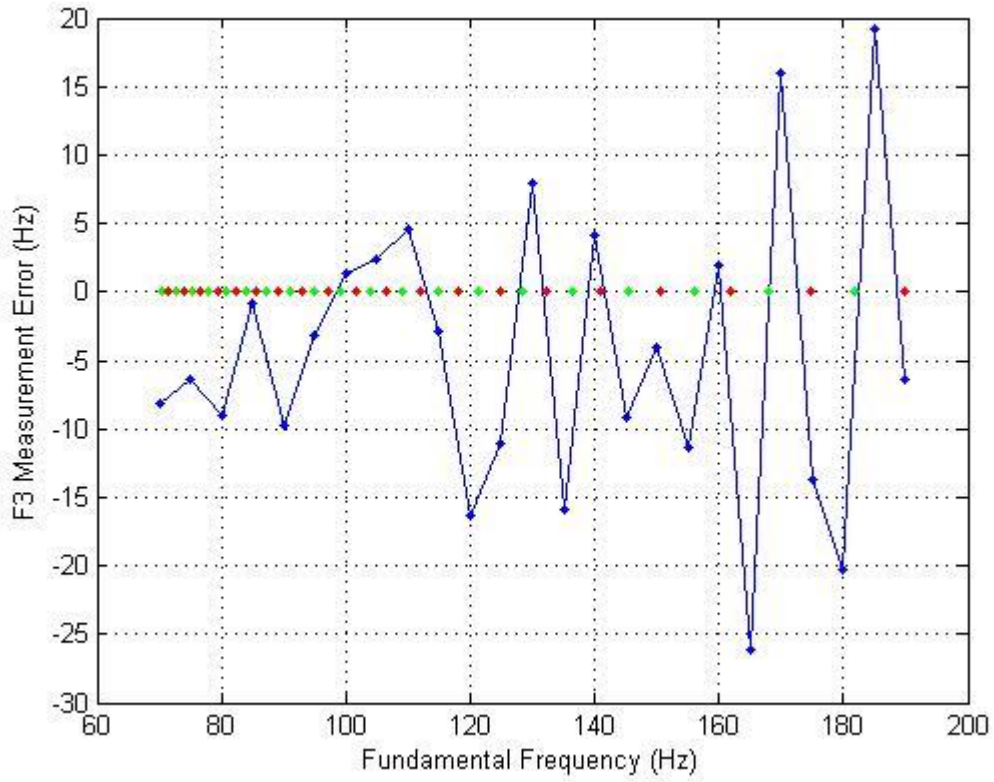

Figure 4.15 F3 measurement error across fundamental frequency for specified F3 formant frequency of 2183 Hz at LPC order 9 from synthetic speech. Green dots represent fundamental frequencies that are integer multiples of 2183 Hz and red dots represent ones that are half integer multiples.

What is apparent in the plots, and is clearest in Figure 4.13 for F1, is that the measurement error oscillates as fundamental frequency changes. A consequence of the oscillation is that there are certain frequencies where the measurement error is zero. This occurs around the point where the fundamental frequency is either an integer multiple of the specified formant value or a half integer multiple. For example, for F1, 166.7 Hz multiplied by 3 and 142.6 multiplied by 3.5 are both equal to 500 Hz, the specified formant value, and it is near these frequencies that the measurement error is zero. The points where the fundamental frequency is an integer multiple of the specified formant frequency have been marked on the plots as green dots, whilst the half integer multiples are red dots.

The closest alignment of the measured error with these points is for F2 at the higher fundamental frequencies. At lower fundamental frequencies the spacing of 5 Hz between measurements is not sufficient to capture the oscillations in the measurements.

This starts to occur below about 110 Hz. It is even more marked for F3 with the lack of data occurring below a fundamental of 160 Hz.

The oscillating behaviour of the errors can be explained by considering the alignment of the resonant formant peaks with the harmonics of the fundamental. The formant measurements are most accurate when a harmonic of the fundamental frequency coincides with the specified formant frequency. In this situation the harmonic can be seen as reinforcing the location of the resonant peak. When a harmonic does not align with the specified formant value, the measured value is pulled away from the true resonant peak by the influence of the nearest harmonic. When the fundamental is a half integer multiple of the formant frequency then the harmonic peaks are located equidistant from the formant centre frequency so the effective pull of the harmonics is cancelled out, resulting in a near zero error.

This effect also accounts for the oscillations in the error surfaces. As the specified formant values increase they periodically become aligned with integer and half integer multiples of the harmonics of the fundamental frequency resulting in near zero errors. The measurement errors systemically increase and decrease as the specified formant values move between these points.

4.4.6 F1~F2 Vowel Space Distortion

Another way of examining the influence of the harmonics of the fundamental frequency on the formant measurements is to generate scatter plots of the measured F1 and F2 values for various combinations of LPC order and fundamental frequency. Figure 4.16 and Figure 4.17 show the measured F1 and F2 values at an LPC order of 8 at a fundamental frequency of 100 Hz and 150 Hz respectively.

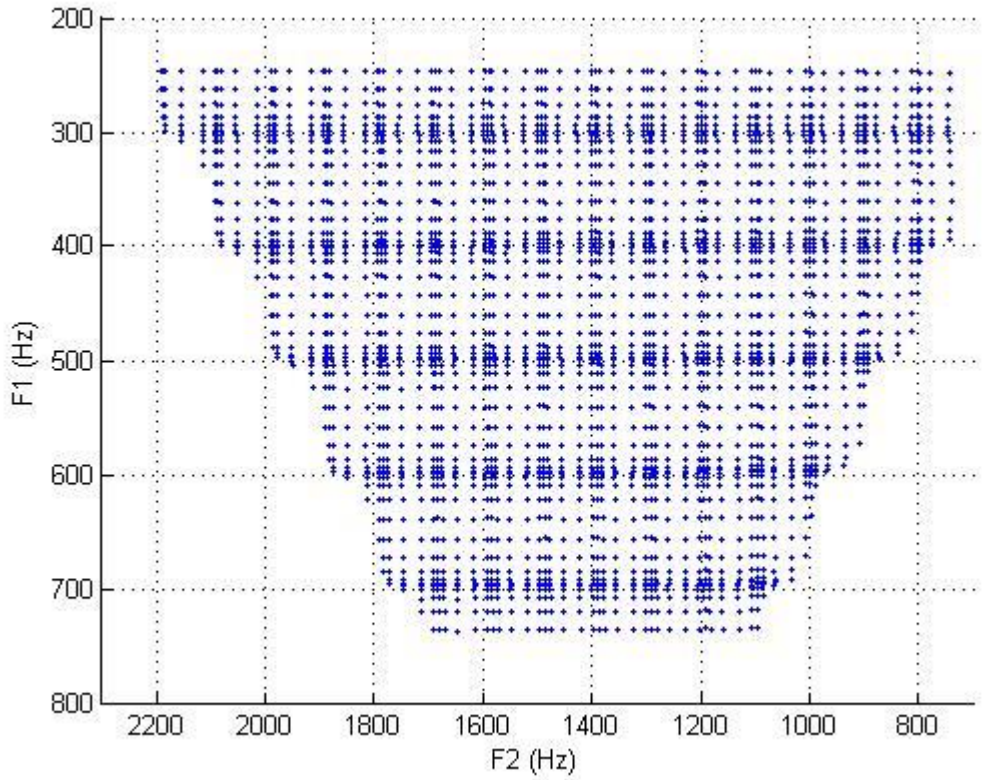

Figure 4.16 F1 and F2 measurements from synthetic speech at LPC order 8 and fundamental frequency of 100 Hz showing the effective distortion of the F1~F2 vowel space.

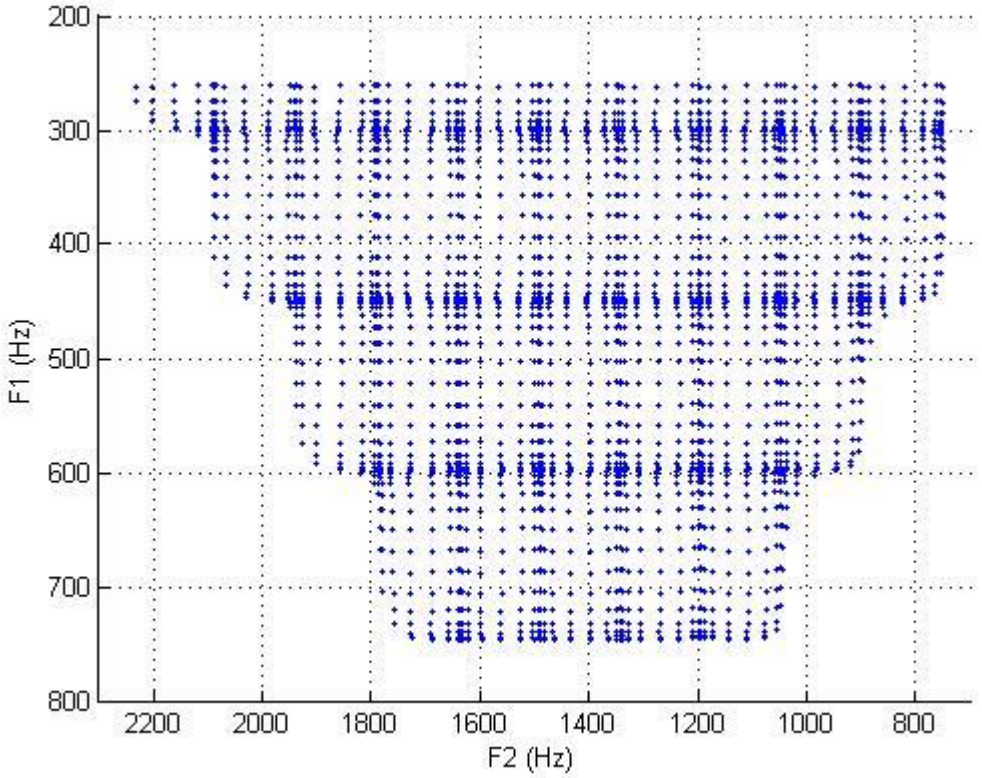

Figure 4.17 F1 and F2 measurements from synthetic speech at LPC order 8 and fundamental frequency of 150 Hz showing the effective distortion of the F1~F2 vowel space.

Both figures show the distortion that has occurred to the evenly sampled vowel space that is shown in Figure 4.2. The distortion manifests itself as the bunching of measurements. This is most apparent in the F1 direction. It is also visible in the F2 direction but owing to the higher spacing between specified F2 values (20 Hz versus 10 Hz for F1) the effect is less pronounced. The bunching of the measurements is centred on the harmonics (i.e. multiples) of the fundamental frequency. In Figure 4.16 with a fundamental frequency of 100 Hz the F1 values are bunched around 300, 400, 500, 600 and 700 Hz. In Figure 4.17, where the fundamental frequency is 150 Hz, the bunching of F1 measurements is around 300, 450 and 600 Hz. However, as the formant frequency increases the bunching tends to occur slightly lower than the harmonic. These plots clearly demonstrate the influence that the harmonics have on the formant measurements.

An alternative way of visualising the bunching effect is to examine the distribution of measurements. From such plots it is clear that the effect also extends to F3.

4.4.7 Measurement Strategy

In the analyses reported so far, the measurements from each LPC order were examined together, and statistical measures were calculated for the results from each order. Whilst this is a useful way to summarise the influence of LPC order and the performance of the software, it does not reflect all of the ways in which the software could be used in practice. Considering all the results from a single LPC order as a set is the equivalent of making all measurements using the same LPC order. If the software is being used in an interactive way, i.e. by examining formant values overlaid on a spectrogram, then it is likely that the LPC order will be adjusted when necessary to obtain more accurate measurements for certain vowel tokens. Also, the analyses have followed the formant numbering system imposed by Praat, where the lowest frequency pole is F1, the second is F2 and so on. Again, this may not reflect how the tool is used by analysts and may reduce the potential accuracy of measurements by following this rule.

In order to investigate these issues, a series of measurement strategies were constructed which reflect the approaches that analysts might adopt when using Praat. The strategies involve selecting measurements which are closest to the specified values and constraining the choice in ways that reflect how an analyst might make decisions and use the software. The strategies are as follows:

1. Praat's formant numbering approach is followed and the LPC order is free to vary from token to token but the three formants must be measured at the same LPC order. For each token the LPC order is selected on the basis of the one which minimises the sum of the absolute error across the three formants.
2. Praat's formant numbering approach is followed and the LPC order is free to vary from token to token and from formant to formant, so the F1 measurement could be from LPC order 9, while the F2 measurement could be from order 10. The LPC order chosen is the one that produces the smallest error for each formant.
3. The selected formant is not restricted to Praat's numbering approach so the measured F2 value could have originally been labelled by Praat as F3. The LPC order can vary from token to token but must remain the same across the three formants. The LPC order and formants are selected on the basis of the ones that minimise the sum of the absolute errors across the three formants.

4. Praat's numbering approach is not used and the LPC order is free to vary across tokens and formants.

The strategies were applied to the measurements, and the mean absolute errors were calculated for the three formants. These are presented in Figure 4.18 together with the results from the approach previously adopted in this chapter, which is referred to as the 'default' strategy. When the formant measurements were made, only the first five formant values were logged. At LPC orders above 10, formant values would have been obtained by the software that were not logged. This means for the approaches where Praat's formant numbering scheme is not followed (i.e. strategies 3 and 4), the complete set of potential formants is not available, and so the analysis may not be a true reflection of the performance that could be achieved.

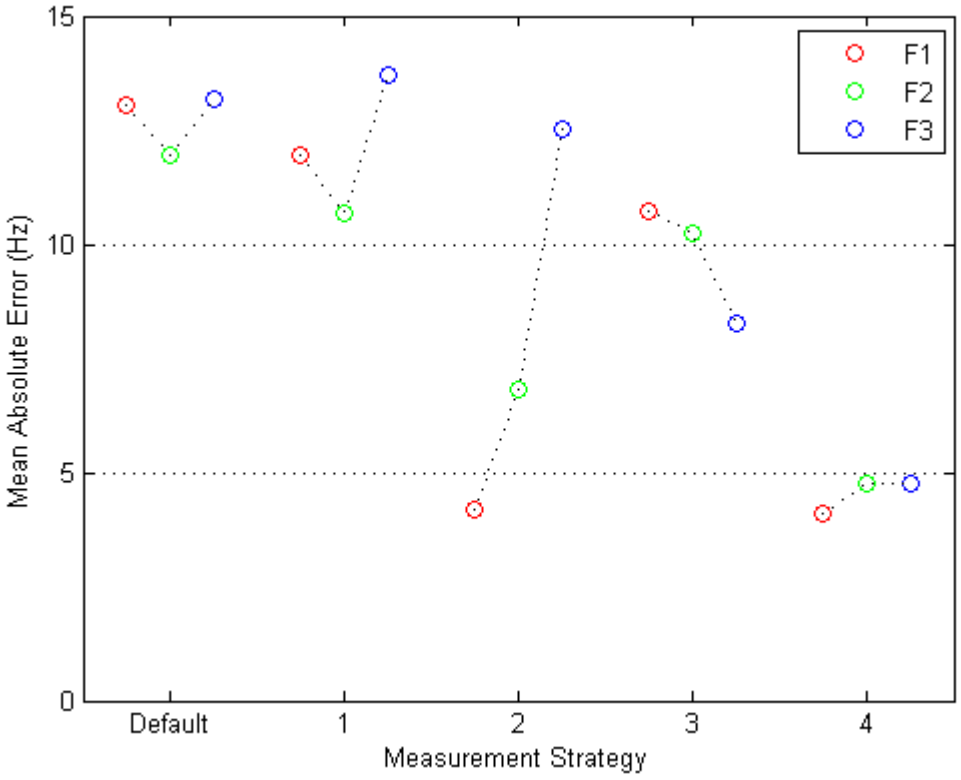

Figure 4.18 Mean absolute errors from synthetic speech for F1 (red), F2 (green) and F3 (blue) across all fundamental frequencies over the entire vowel space from four measurements strategies and the default approach.

The results in Figure 4.18 show that altering the measurement strategy does influence the magnitude of the measurement errors. Strategy 1 shows a slight improvement over the default approach for F1 and F2, but at the expense of the error for F3. Therefore simply changing LPC order on a token by token basis to obtain more accurate

measurements has a limited effect compared with keeping it constant across all tokens, as in the default approach. However, using different LPC orders for each formant does result in a marked improvement for F1 and F2 in Strategy 2. Abandoning Praat's formant numbering approach in Strategy 3 shows a moderate improvement in performance, which is greatest for F3. Finally, Strategy 4, the least constrained of the approaches, gives the best performance with mean absolute errors around 5 Hz for all three formants.

These results show that improvements in performance can be achieved by altering the LPC order from token to token, using different LPC orders for each formant and not being constrained by Praat's numbering approach. The topic of measurement strategies is returned to Section 6.3.2.

4.5 Summary

The main focus of this chapter addressed research question 2 concerning the effect of analysis parameters on formant measurement accuracy. The methodology employed considered the effects of LPC order. Using synthetic speech, rather than real speech, allowed the accuracy of the measurements to be quantified.

The analysis of the measurements demonstrated the variation in performance of Praat's formant measuring tool as LPC order was altered. The results showed that for the synthetic speaker LPC order 9 produced the most accurate measurements, with a mean absolute error of approximately 13 Hz for each formant. Below this LPC order the measurements tended to be overestimates, whereas above it the measurements were underestimates. Outside the range of orders 7 to 9 the magnitude of the errors was very large, especially for F2 and F3. Closer inspection of the results showed that the behaviour of the measurements for each of the three formants was different across the LPC orders.

Since formant measurements were obtained from the synthetic speech across a range of fundamental frequencies, some insight was gained that helps to answer the third research question, which asks how the accuracy of measurements can vary across speakers. Comparing the mean absolute error obtained across the fundamental frequencies of the synthesised speech showed a clear influence on the accuracy, with the most accurate measurements being made at the lower fundamentals and the least

accurate at the highest. The change in performance across the fundamentals appeared linear and this held for all three formants and the three LPC orders tested.

The analysis also revealed the influence of the harmonics of the fundamental frequency on the measurements. Since the harmonics of the fundamental are effectively sampling the frequency response of the vocal tract and concentrations of energy are present at these harmonics, it is perhaps unsurprising that the measured values will be drawn towards the harmonics. This behaviour resulted in the error surface plots over the vowel space having a cyclic form, with the period of oscillation of the surface being approximately equal to the fundamental.

The way in which the accuracy of the measurements was initially assessed across the LPC orders only gives an insight into the practical situation where all measurements are made using the same LPC order. Whilst this approach might be adopted where the measurement process is entirely automated, if an interactive approach is used then the analyst is likely to alter the LPC order in an attempt to obtain more accurate measurements. In order to replicate different ways in which an analyst might use the software a number of strategies were formulated and the performance was determined for each one. The strategies involved combinations of permitting the LPC order to vary across tokens and formants, as well as bypassing Praat's assignment of pole frequencies to formants. In the least constrained scenario, in which the LPC order could vary across formants and Praat's formant numbering approach was ignored, a large improvement in performance occurred with the mean absolute error being reduced to approximately 5 Hz for all three formants.

When interpreting the measures of performance and behaviour reported in this chapter it must be remembered that they have originated from synthetic speech, which conforms to the assumptions imposed by LPC analysis. Therefore, the results should be considered as a best-case scenario. Also, they only represent the performance obtained with one speaker. The following chapter addresses this issue, which is raised in the third research question, by examining the performance for multiple synthetic speakers. However, the results from this chapter should instil confidence in Praat's formant measuring tool as it is clearly capable of producing relatively accurate measurements, especially when used interactively.

The obvious guidance which stems from the analysis in this chapter is:

- To obtain the most accurate measurements the LPC order should be adjusted, where necessary, for each vowel token and formant. The numbering of formants employed by Praat can be ignored.
- If LPC order cannot be varied then care should be taken to ensure that a suitable LPC order is selected, since measurements obtained with an inappropriate order can lead to highly inaccurate results.

Chapter 5 Multiple Synthetic Speakers

5.1 Introduction

Chapter 4 examined formant measurement accuracy across the vowel space of a single synthetic speaker when fundamental frequency and LPC order were varied. Several trends and patterns were present in the data, but since they were derived from what is effectively a single speaker, it is not apparent how applicable they are to other speakers. The current chapter explores this issue by analysing and comparing the measurement errors from multiple synthetic speakers, which addresses the third research question:

RQ 3. To what extent does the accuracy of LPC formant measurements vary across speakers?

The methodology involves examining the accuracy of formant measurements from two groups of synthetic speakers, one which have different sets of specified F3 values and the other which employ more realistic glottal source signals. These parameters were chosen as variables since they are known to vary between individuals. The measurements were made across a range of LPC orders, and these results therefore provide insights into the second research question:

RQ 2. How does altering the LPC analysis parameters affect formant measurement accuracy?

The extent of variation in the synthesis parameters, and in the measurements, by no means covers the complete range of variability in real speech. However, the results provide an indication of the extent of variability in formant measurement errors that can exist between speakers. They also serve to reinforce the guidance offered in the previous chapter.

5.2 Alternative F3 Speakers

5.2.1 F3 Calculation

In Chapter 4 a bi-planar representation of F3 was used to calculate the F3 values for all F1~F2 combinations using equations and data from Nearey (1989). Whilst this representation of F3 is motivated by observations made by Broad and Wakita (1977), it is not the only method for describing the relationship between F3 and the first two

formants. An alternative approach is presented by Kasuya and Yoshizawa (1992) (cited in Kasuya et al., 1994) in which regression analysis is used to derive the coefficients for a quadratic representation of F3. This is shown in Equation 11, where a_0 to a_5 are coefficients.

$$[11]\ F_3 = a_0 + a_1F_1 + a_2F_2 + a_3F_1^2 + a_4F_2^2 + a_5F_1F_2$$

In Kasuya et al. (1994) coefficients were derived from five adult Japanese male speakers repeating the same short phrase /aoiue/ ("blue top" in English) in their normal speaking style, with two sets from one speaker (Speaker A and A') who also adopted a different prosodic style. The specified F3 values for the synthetic speakers analysed in this chapter were calculated using these coefficients and Equation 11 across the F1~F2 space defined in the previous chapter. Two sets of generated formants, those for speakers B and C, were found to contain values that were very close to or overlapped with either the F2 or F4 values in certain limited regions of the vowel space. These speakers were rejected from this study since the F1~F2 vowel space would have had to be modified to accommodate them and this would have resulted in non-directly comparable sets of measurements. The coefficients for the speakers that did produce acceptable F3 values are shown in Table 5.1.

Speaker	**a_0**	**a_1**	**a_2**	**a_3**	**a_4**	**a_5**
A	3570	0.12	-2.27	2680	1.05e-3	-1.54e-3
A'	4060	-2.61	-2.20	5270	0.90e-3	-0.77e-3
D	3970	-0.66	-1.55	1720	0.52e-3	-0.50e-3
E	3580	3.49	-1.99	-1250	0.73e-3	-1.44e-3

Table 5.1 Coefficients from Kasuya et al (1994) for predicting F3 values from F1 and F2 using a quadratic function for four speakers.

The specified F3 values generated for each speaker are summarised in Table 5.2. The values for the speaker generated and analysed in Chapter 4, referred to as the 'baseline', are also shown for comparison.

Speaker	**Mean (Hz)**	**SD (Hz)**	**Min (Hz)**	**Max (Hz)**	**Range (Hz)**
Baseline	2342	165	1994	2862	868
A	2269	164	2061	3068	1007
A`	2386	219	2098	3093	995
D	2647	117	2506	2958	452
E	2686	201	2366	3248	881

Table 5.2 Summary statistics for the specified F3 values for the Kasuya et al (1994) speakers and the baseline speaker from the previous chapter.

The summary statistics in Table 5.2 show that the specified F3 values generated for each speaker are similar. However, this is to be expected given the physiological constraints that are shared by the speakers and that the F3 values were generated using the same set of F1~F2 values. Figure 5.1 shows the specified F3 values as a surface for each of the four Kasuya et al (1994) speakers.

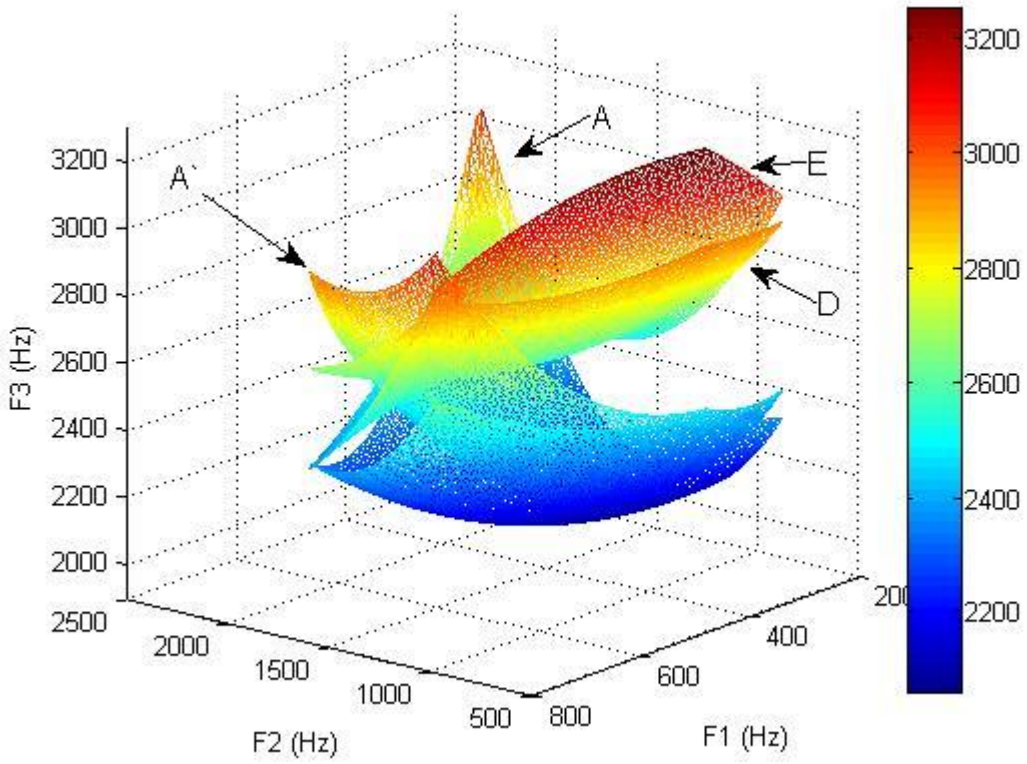

Figure 5.1 Three dimensional representation of the F1~F2~F3 synthetic vowel space for all four Kasuya et al (1994) speakers, with F3 represented by both height in the z axis and colour.

Overall, the shape of the F3 surfaces is different from that of the baseline speaker shown in Figure 4.3. This is mainly due to the two different mathematical approaches used to derive the values. Speakers A, D and A` all have bowl-like surfaces, whereas for Speaker E the main portion of the surface, rising up from the lowest region, has a convex shape. A and A` (recall these are in fact the same speaker) have similar F3 values for close vowels, but as F1 increases the values diverge. Speakers D and E are somewhat similar to each but different from A and A`. However, the overall shape of the surface for Speakers D and E is different and they tend to diverge at the edges of the vowel space.

5.2.2 Determination of Measurement Errors

To determine the measurement errors associated with each of the four Kasuya speakers, the specified formant values were used in the same processes described in Chapter 4. Since fundamental frequency was previously shown to produce small variation in overall performance relative to LPC order, it was was held constant at 100 Hz. Measurements were made at LPC orders from 6 to 20 and the vowels were synthesised with a sampling rate of 44.1 kHz.

5.2.3 Analysis of Measurement Errors

The formant measurements and resulting errors for the Kasuya speakers were subject to the analysis methods described in Chapter 4, namely the generation of error surface plots, the plotting of measurement errors against specified values, and the calculation and plotting of statistical summary data. The outcomes of these analyses were also compared with the results for the baseline speaker with a fundamental frequency of 100 Hz. The most relevant data and findings are presented below.

5.2.3.1 Statistical Summary Data

The statistical summary data showed the formant measurement errors from the Kasuya speakers to be very similar to those from the baseline speaker, particularly for the first two formants. LPC orders 7, 8 and 9 again produced accurate and relatively stable measurements, with order 9 being the most accurate overall for all speakers. Table 5.3 shows the mean absolute error for the first three formants, and all formants combined, from LPC order 9.

Speaker	**Mean Absolute Error (Hz)**			
	F1	**F2**	**F3**	**F123**
Baseline	8.82	8.04	10.70	9.19
A	8.80	7.87	9.09	8.58
D	9.25	8.53	15.55	11.11
E	9.31	8.58	16.32	11.41
A'	8.86	8.04	11.87	9.59

Table 5.3 Mean absolute error for Kasuya and baseline speakers with F0 of 100 Hz for the first three formants and all three formants combined at LPC order 9.

For F1 and F2 the mean absolute error in Table 5.3 shows very little difference across the speakers. Examination and comparison of the error surfaces for these formants at LPC orders 7, 8 and 9 showed them to be very similar at each order for all speakers in terms of structure and error values. Even at LPC orders 6 and 10 to 20, which produced

substantially inaccurate measurements, the statistical summary data showed the behaviour of the measurements to be closely aligned across speakers. The conclusion that can be drawn from these observations is that the measurement of F1 and F2 is minimally influenced by the specified F3 values.

Unlike F1 and F2, the F3 mean absolute error in Table 5.3 shows variability across the speakers. For speaker A the error is 9.09 Hz, which is less than the 10.70 Hz error for the baseline speaker, whereas for speaker E the error is considerably greater at 16.32 Hz. A similar pattern was observed for LPC orders 7 and 8, where the rank order of the speakers based on performance was the same as order 9, i.e. A, baseline, A', D, E. If the speakers are ranked according their mean specified F3 values, which are shown in Table 5.2, then the ordering is the same as for their performance. This suggests that the magnitude of the measurements errors is related to the specified values, an issue discussed further in Section 5.2.4.

5.2.3.2 F3 Error Surfaces

Comparison of the F3 error surfaces generated for the Kasuya speakers and the baseline speaker revealed differences between the speakers. For the baseline speaker, the error surface shown in Figure 4.6 has two distinct regions that correspond to the two planes on which the specified values lie. Within the two regions the cyclic peaks and troughs run parallel but their orientation and apparent period of repetition is different in each. For the Kasuya speakers the surfaces are again cyclic but the peaks and troughs are elliptical or curved rather than parallel. This can be seen clearly in Figure 5.2, which shows the F3 error surface for Speaker A at an LPC order of 8. For this speaker the peaks and troughs are elliptical and concentric.

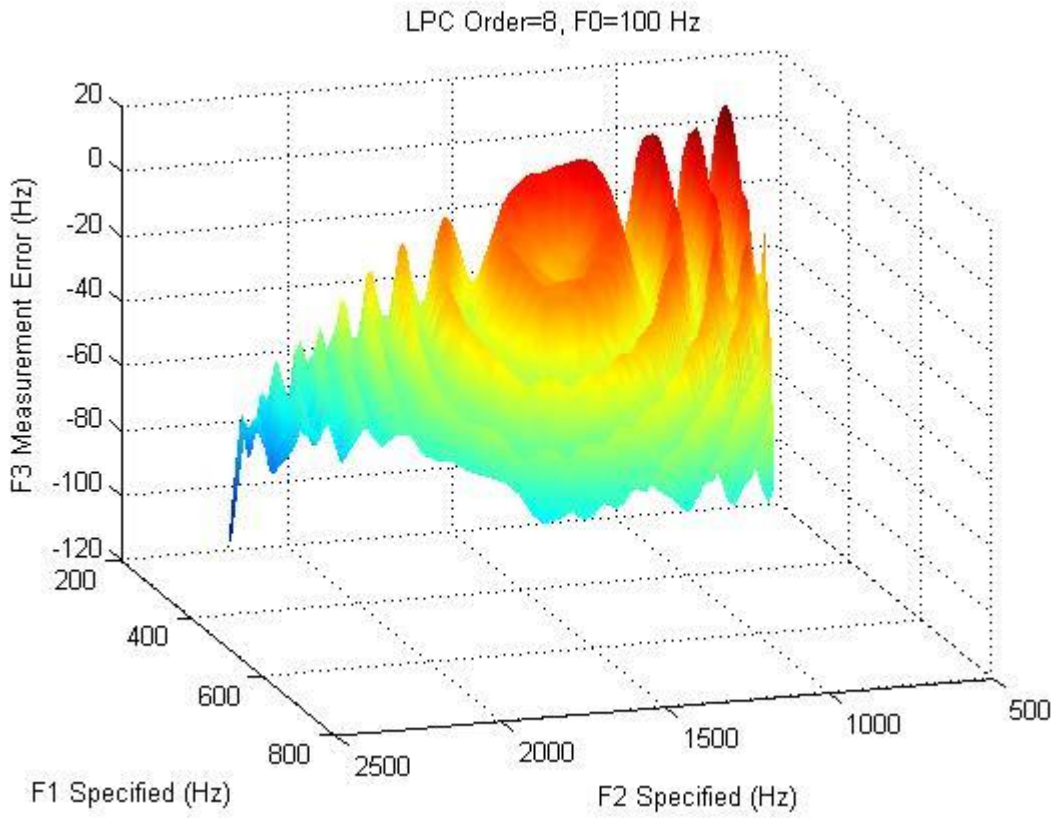

Figure 5.2 F3 measurement error surface for Kasuya Speaker A with F0 of 100 Hz at LPC order 8.

The differences in structure of the error surfaces are not surprising given the different forms of the specified F3 values over the F1~F2 vowel space. As discussed above, the peaks and troughs in the error surfaces, for all formants, correspond to regions where the specified formant value is the same. For the Kasuya speakers, the F3 values lie on quadratic surfaces and consequently the regions in which the specified formant values are the same are either elliptical or curved. Comparison of the specified F3 values with the F3 measurement errors over the F1~F2 vowel space for each of the Kasuya speakers confirmed that the peaks and troughs do correspond to regions with the same specified F3 values.

Even though the peaks and troughs correspond to vowels with the same specified F3 value, the measurement error associated with them varies over the vowel space. This shows that the F3 measurement error is not only dependent on the specified F3 value but also on the other formants, i.e. its position within the F1~F2 vowel space. The F3 error surface shown in Figure 5.2 for Speaker A at LPC order 8 shows a general trend for the magnitude of the errors to increase in a negative direction (underestimate the true value) towards the front open vowels, i.e. those with high F1 and F2 values. This is also

the case when the LPC order is 9, but at the lower LPC order of 7 the increase in error magnitude shifts to become a greater over-estimation towards the front open vowels. These overall trends in the error surfaces at different LPC orders are generally the same for all the Kasuya speakers as well as the baseline speaker.

5.2.4 Speakers With Constant F3

To investigate the dependence of the F3 errors on the F1 and F2 specified values, a further four synthetic speakers were generated with constant specified F3 values over the F1~F2 vowel space. Keeping the specified F3 values constant would eliminate their effect on the F3 errors. The F3 values used were 2500 Hz, 2750 Hz, 3000 Hz and 3250 Hz, as they spanned the range between the maximum F2 value at 2230 Hz and the constant F4 value at 3500 Hz. Again, the speakers were generated using the same method described in the previous chapter, except that the specified F3 values were held constant and fundamental frequency was not varied. The resulting synthesised vowel tokens were only analysed at LPC orders 7, 8 and 9 using the same methods already employed.

Figure 5.3 to Figure 5.5 show the F3 measurement error surfaces for each of the four constant F3 synthetic speakers at LPC orders 7, 8 and 9, with a fundamental frequency of 100 Hz.

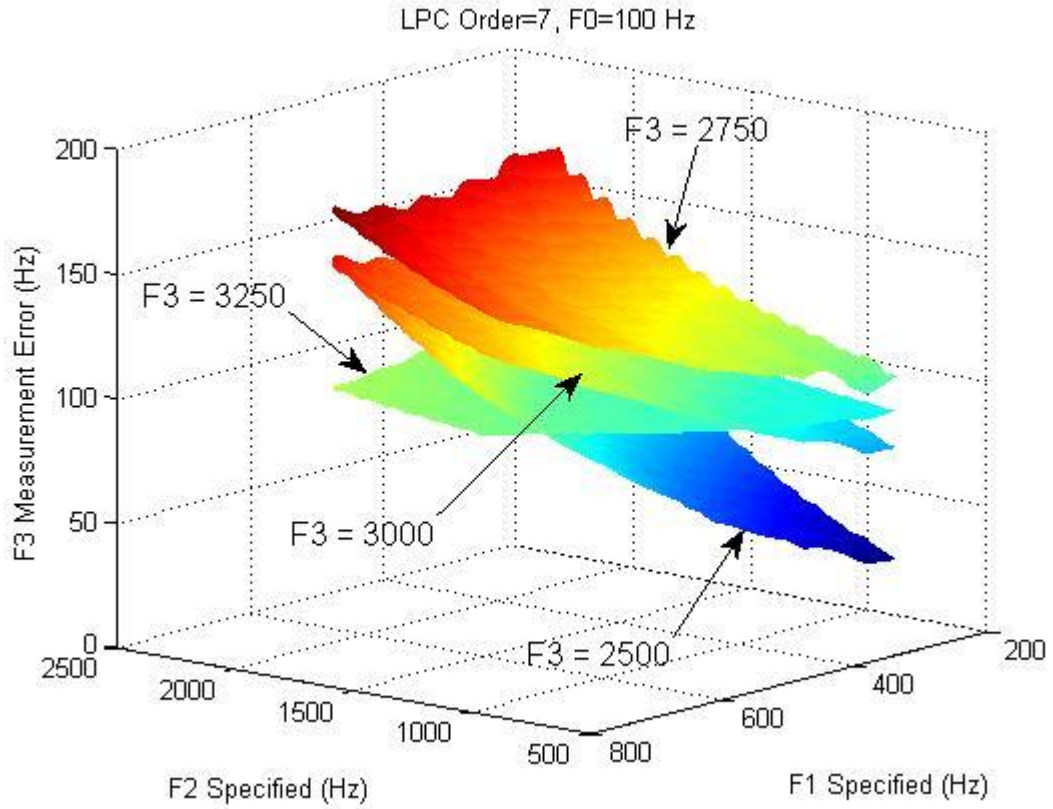

Figure 5.3 F3 measurement error surface from constant F3 synthetic speakers at LPC order 7 with fundamental frequency of 100 Hz.

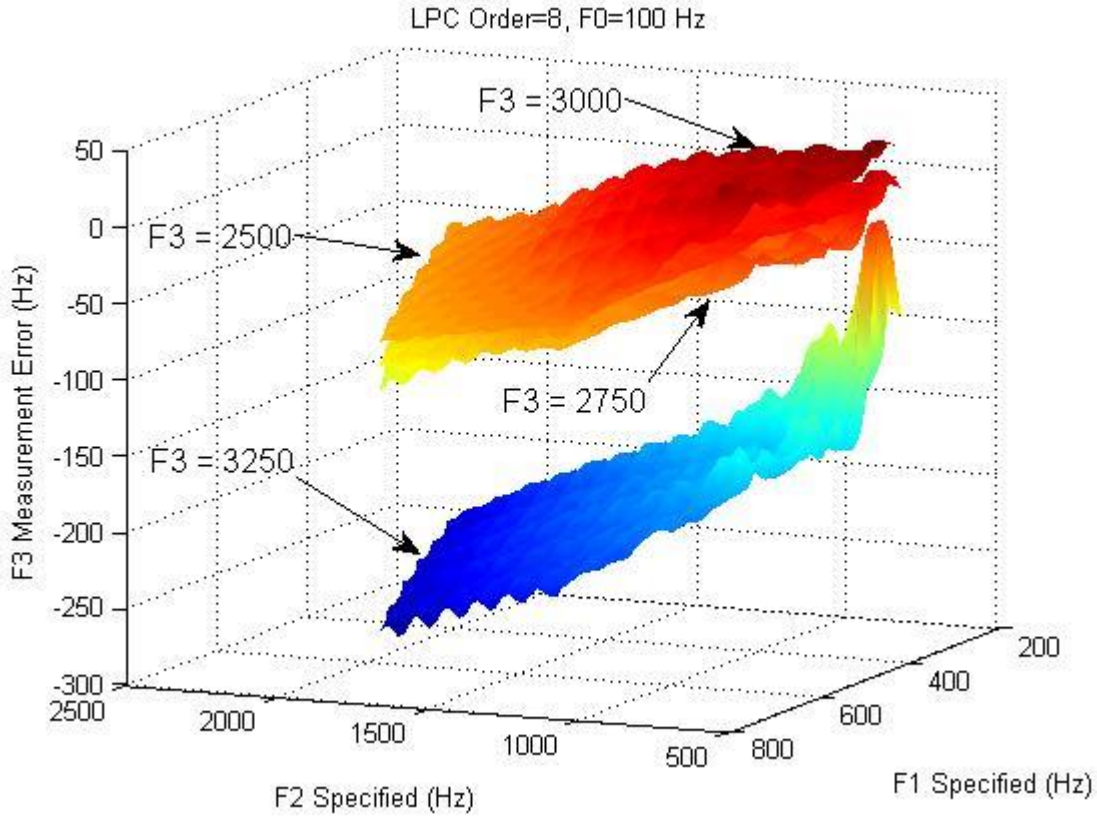

Figure 5.4 F3 measurement error surface from constant F3 synthetic speakers at LPC order 8 with fundamental frequency of 100 Hz.

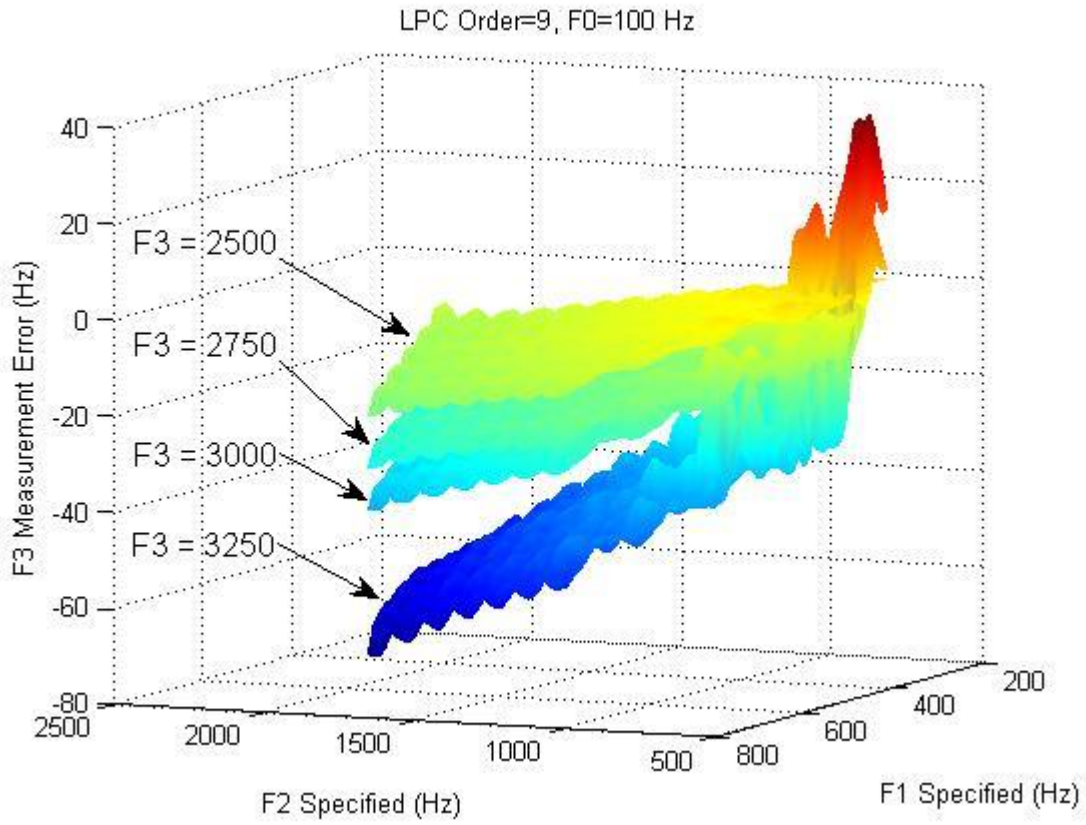

Figure 5.5 F3 measurement error surface from constant F3 synthetic speakers at LPC order 9 with fundamental frequency of 100 Hz.

All of the error surfaces are non-horizontal and exhibit different degrees of slope over the vowel space, which shows that the F3 measurement errors have a dependence on F1 and F2. There is also a dependence on the specified F3 values and the LPC order as the location and slope of all of the surfaces across these variables are different. The general direction of the slope of the surfaces at each of the LPC orders is the same as that described above for the Kasuya speakers and the baseline speaker. Figure 5.3 shows that at LPC order 7 the error surfaces tend to slope downwards from the open front vowels, where the largest errors occur, to close back vowels with the smallest errors. For LPC orders 8 and 9, shown in Figure 5.4 and Figure 5.5, the direction of slope is reversed, with the surfaces sloping downwards from close back vowels to open front vowels. However, the largest errors still occur with open front vowels but at these LPC orders they are negative, showing that the measurements are underestimates of the true formant values rather than overestimates, which occur at LPC order 7. The overall performance for the constant F3 speakers, represented by mean absolute error, is shown in Table 5.4.

Specified F3 (Hz)	Mean Absolute Error (Hz)		
	LPC Order 7	LPC Order 8	LPC Order 9
2500	95.70	37.58	9.55
2750	140.03	54.81	19.87
3000	121.07	36.47	24.11
3250	100.89	206.74	46.15

Table 5.4 Mean absolute measurement errors from synthetic speakers with constant specified F3 values with a fundamental frequency of 100 Hz.

Again, LPC order 9 produces the best performance for each speaker and the magnitude of the error increases as the specified F3 value increases. This is the same pattern that was seen with the Kasuya speakers. This trend is also relatively clear in the error surface plot in Figure 5.5. For LPC orders 7 and 8 the performance does not appear to be linked to the specified F3 values.

5.2.5 Summary of Results from Alternative F3 Speakers

The results presented above show the impact that changing the specified F3 values had on the measurement errors. For the F1 and F2 errors this was very small, with both the structure of the error surfaces and the values represented by them being very close to the baseline speaker. However, the effect on the F3 errors was more marked. LPC order was shown to have an impact on both the magnitude and the overall behaviour of the F3 measurement errors across all speakers, with LPC order 9 producing the most accurate measurements. At this order, the errors increased as the specified F3 values increased. The form of the error surfaces was strongly influenced by the structure of the specified F3 values over the vowel space. Generating speakers with constant F3 values not only demonstrated the dependence of F3 measurement errors on the vowel token's F1~F2 values but also the influence that the specified F3 values have. This was shown to vary across LPC orders.

5.3 Alternative Glottal Source Speakers

A simple pulse train signal was used as the glottal sound source for the synthetic speaker in Chapter 4, and for the alternative F3 speakers discussed above. Whilst intelligible speech can be produced using such a signal, other glottal signal representations have been developed, from which more accurate and realistic sounding speech can be generated (Fant et al. 1985). In the following sections one such representation is used to generate a further set of synthetic speakers that are subject to the same measurement and analysis procedures already employed. These results help

address the third research question, as they provide some indication of the potential extent of variation in measurement errors that can exist across speakers.

5.3.1 Approximation of the LF Model

The LF model is a four parameter glottal flow model (Fant et al. 1985). Whilst this model has been widely used both as the source for speech synthesisers and as a mathematical model for analysing real glottal source signals, it is computationally complex. A simplified approximation of the model has been proposed, which produces acceptably similar results (Qi & Bi 1994). This is the glottal source model that is used in the following sections of this chapter.

The simplified model, like the LF model, consists of two equations that define the derivative (the rate of change) of the glottal flow. The derivative of the glottal flow is generally used in speech synthesis applications since this form incorporates the radiation effect at the lips. The first equation is the same as that in the LF model (Qi & Bi 1994, equation 1). The second equation remains as an exponential function but is simpler in form than in the LF model (Qi & Bi 1994, equation 6). This allows the model parameters to be calculated easily without solving the roots of two non-linear equations, as the LF model requires.

The synthesis model parameters, α and ω_g can be calculated relatively easily (Qi & Bi 1994, equation 8). Two methods are provided in Qi and Bi (1994) to determine the value of the third parameter ε. The first, termed 'approximation I' requires root solving techniques, so the simpler 'approximation II' equation was used (Qi & Bi 1994, equation 11).

5.3.2 Generation of Glottal Waveforms

The number of combinations of parameters that will generate realistic glottal waveforms for LF-type models is large. To ensure that the parameter values used would produce viable waveforms with the simplified LF model, those presented by Qi and Bi (1994) were used. These values were as follows: the gain constant, E_0, was 1, whilst t_e and t_p were held constant at 60% and 45% of the fundamental period (t_c) respectively. In Qi and Bi (1994) the parameter t_a is varied between 1% and 20%, however, in the simplified model t_a and E_e are equivalent independent parameters, i.e. changing t_a or E_e has the same effect on the waveform. In the present study, E_e, the amplitude of

maximum negative excursion of the glottal derivative waveform, was chosen as the variable for computational simplicity. The range of values for E_e ranged from 1 to 10 in steps of 1, which are approximately equivalent to t_a values of 20% and 1% respectively. The fundamental frequency was held constant at 100 Hz.

To generate the glottal waveforms the equations from Qi and Bi (1994) were implemented in a Matlab script and the waveforms were generated at a sample rate of 44.1 kHz. The equations only generate a single period of the waveform, so each waveform was repeated in series to form a sound file with a duration of 300 ms that could be used by the Praat script to generate the synthetic vowel tokens. Figure 5.6 shows a single period of the 10 waveforms generated for each of the E_e values.

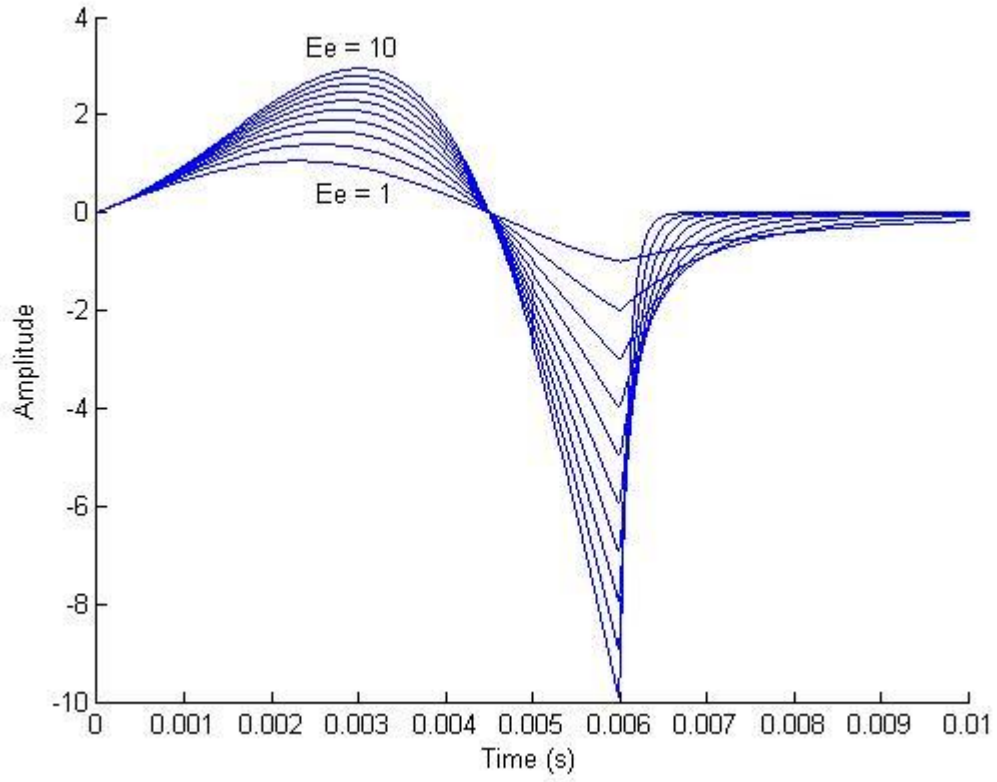

Figure 5.6 A single period of the ten waveforms generated using the simplified LF model with E_e varying from 1 to 10.

Figure 5.6 shows that as E_e increases from 1 to 10, the amplitude of the negative excursion relative to the positive excursion increases. Also, the speed of the transition from the negative excursion back to zero increases. This is modelling a more abrupt closure of the vocal folds. The resulting spectral characteristics of the waveforms are shown in Figure 5.7. These have been generated from the 0.3 second duration audio

files and smoothing has been applied to allow easier comparison of the spectra and to highlight their overall shape.

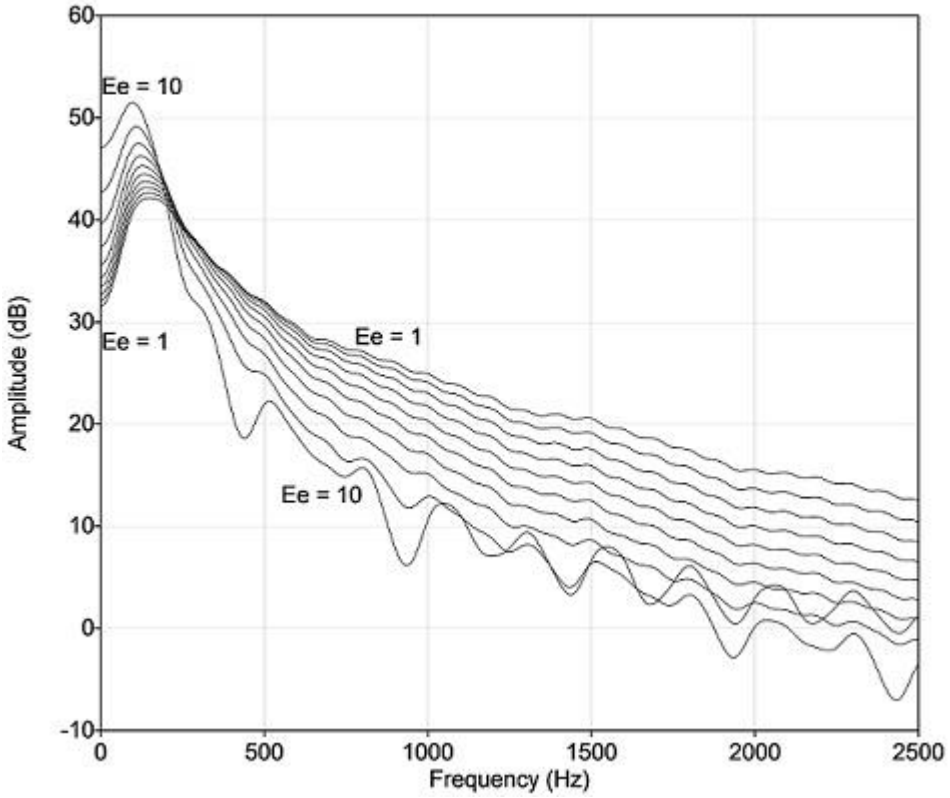

Figure 5.7 Smoothed frequency spectra of glottal waveforms generated using the simplified LF model with E_e varying from 1 to 10.

The spectra show that as E_e increases from 1 to 10 the difference in amplitude increases between the low frequency region around the fundamental and the higher frequency regions above. The spectral slope between 200 and 500 Hz also increases. At higher frequencies, above 1000 Hz, the spectral slope is relatively constant across all speakers at approximately –12 dB per octave.

Whilst the set of parameters used to generate the glottal waveforms have not been chosen to represent specific voice qualities, the spectra for E_e =1 could be considered as representing a modal voice, while at E_e = 10 the spectra is more like a breathy voice, but without the higher frequency aspiration noise (Fant et al. 1985). To represent specific voice qualities more accurately would require the adjustment of several glottal source parameters. Since the intention is to provide an indication of the potential degree of variation in formant measurement errors across speakers, and not to investigate the effects of voice quality directly, this was not done.

5.3.3 Determining Formant Measurement Errors

The ten glottal source signals were used in Praat to generate synthetic vowel tokens using the same method for the baseline speaker described in the previous chapter. The same measurement and error calculation process was used with LPC order ranging from 6 to 20. The measured formants and resulting errors were subject to the same analysis techniques previously used. The important results and findings are discussed below.

5.3.4 Analysis of Formant Measurement Errors

In general, the error surfaces generated from the variable glottal source speakers show less regularity in structure and greater variation than the baseline speaker across LPC orders 7 to 9. For example, Figure 5.8 shows the F1 error surface for LPC order 8 from the speaker with an E_e value of 2. The local variation in amplitude in terms of peak to peak (or trough to trough) differences is not consistent across the surface and is much greater than for the baseline speaker shown in Figure 4.4. This effect is present across the error surfaces for F1, F2 and F3 for LPC orders 7 to 9, and the degree of variation increases as the E_e value decreases.

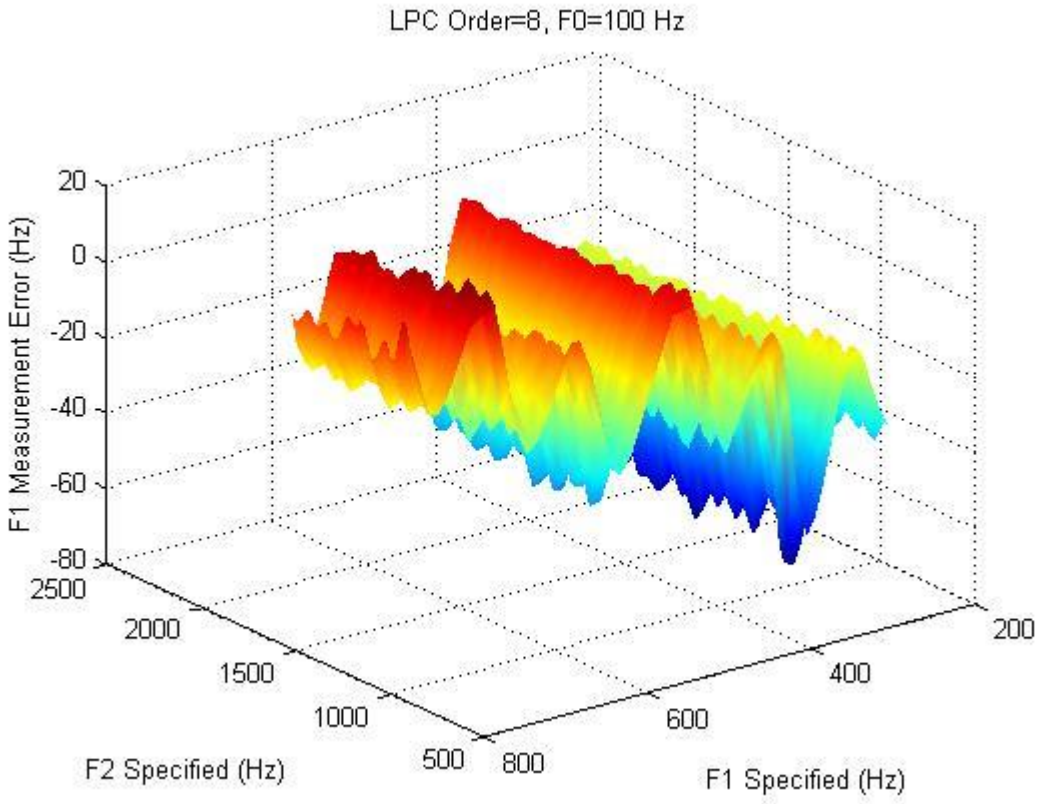

Figure 5.8 F1 error surface for LPC order 8 for synthetic speaker with glottal source E_e value of 2.

Another obvious feature in many of the F2 and F3 error surfaces is a switch in direction of slope at LPC order 8 when compared with the baseline speaker. For the baseline speaker at LPC orders 8 and 9 the direction of slope of the F2 and F3 error surfaces were the same. The F2 error surfaces sloped downwards from back vowels to front vowels, whilst the F3 error surfaces sloped downwards from close back vowels to open front vowels. At LPC order 7 the error surfaces exhibited slope in the opposite directions. For the variable pulse source speakers at LPC 8 the F2 and F3 error surfaces slope in the same direction as the surfaces for LPC 7 for both the baseline speaker and the variable pulse source speakers.

A feature that is present in many of the errors surfaces is localised regions with large error values relative to the rest of the surface. These regions occur either at the edge of the surface, as shown in Figure 5.9, or across several small relatively regularly spaced locations, as seen in Figure 5.10. In some instances they both occur in the same surface.

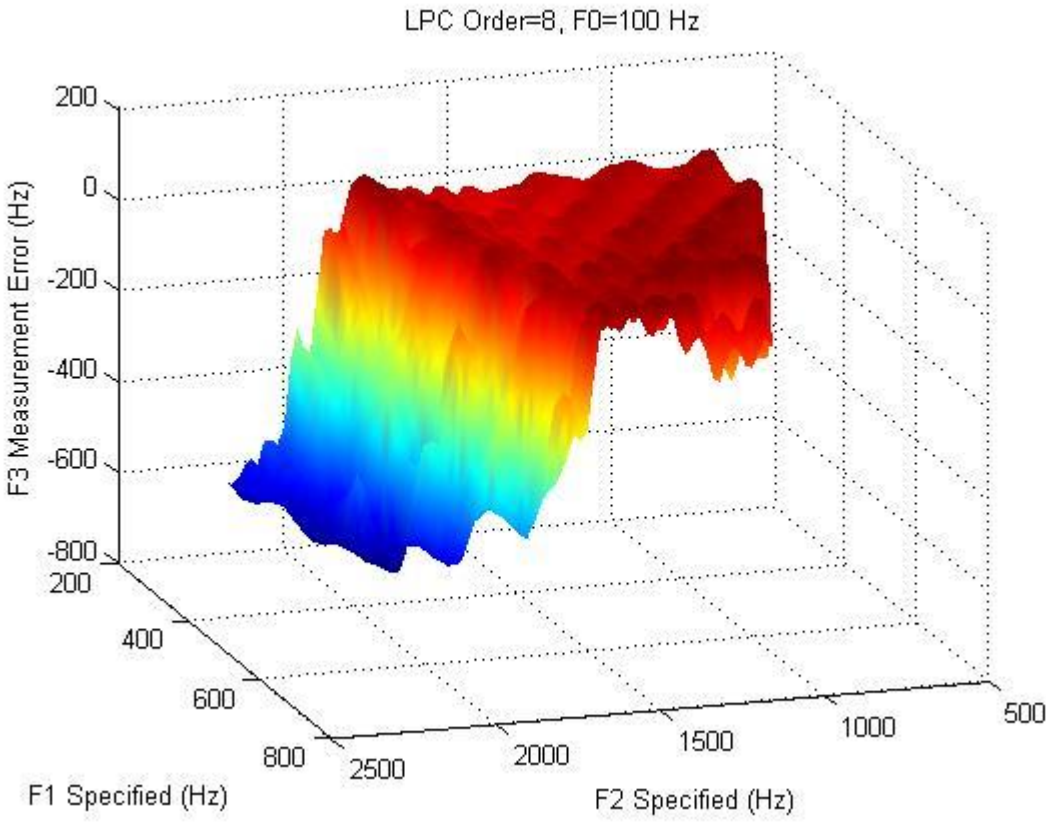

Figure 5.9 F3 error surface for synthetic speaker with glottal source E_e value of 8 at LPC order 8.

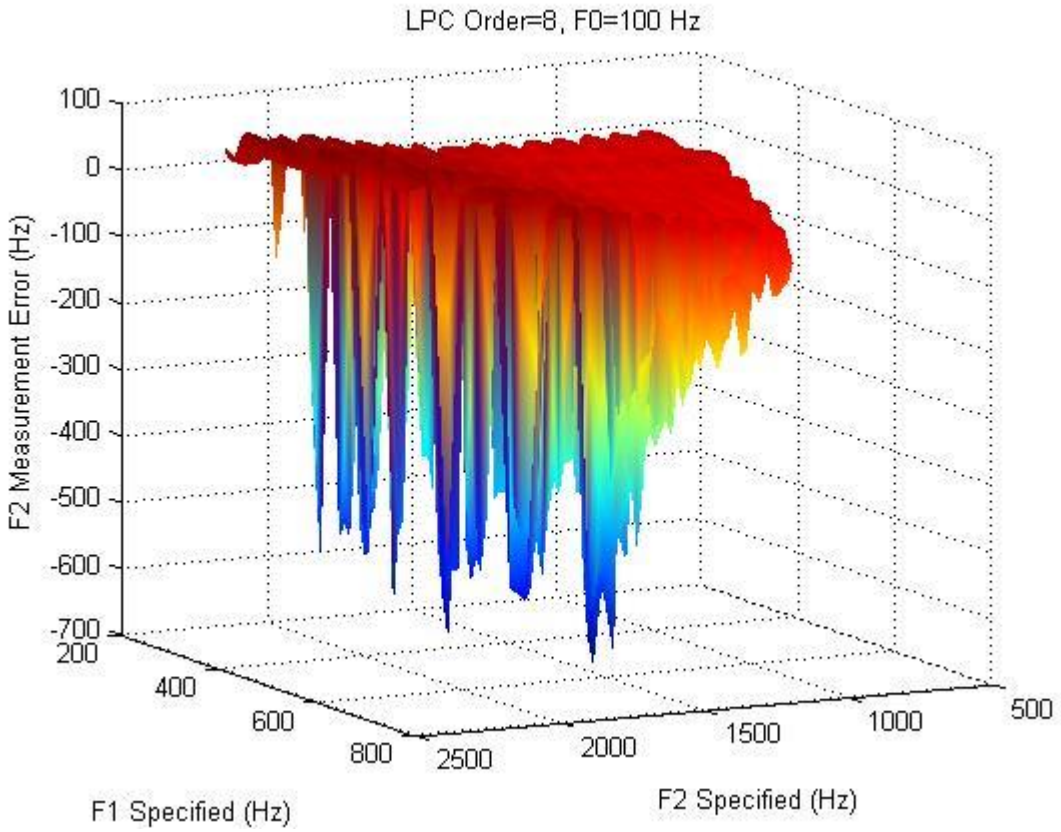

Figure 5.10 F2 error surface for synthetic speaker with glottal source E_e value of 6 at LPC order 8.

The F3 error surface for the speaker with an E_e value is 8 at LPC order 8, shown in Figure 5.9, exhibits very large errors for most front vowels. The error values across the rest of the surface are relatively small and comparable with those for other speakers and LPC orders. This type of localised divergence is not seen in all error surfaces, as it only occurs for certain combinations of LPC order, speaker and formant. However, these combinations do include all formants and LPC orders. Also, the divergent regions do not seem to be restricted to a particular edge of the vowel space and may occur along only a section of an edge. The same type of localised divergent errors were also present for some of the Kasuya and constant F3 speakers, shown in Figure 5.4 and Figure 5.5, however, the magnitude of the divergent errors was much smaller. The same feature was also observed for the baseline speaker but outside the range of LPC orders that produced the most stable results, i.e. LPC orders 7, 8 and 9.

In Figure 5.10, the F2 error surface for the speaker with an E_e value of 6 at LPC order 8, the divergent regions occur systematically across almost half of the vowel space. An alternative view of the surface, in which the error value is represented just by colour, is shown in Figure 5.11. In this plot, the spatial patterning of the divergent regions is much clearer than in Figure 5.10. It is also apparent that the degree of divergence, i.e. the

magnitude of the errors, as well as the size of the regions decreases from front to back vowels. The size also decreases from open to close vowels. In the specified F1 direction the divergent regions occur roughly every 100 Hz, the same as the fundamental frequency, but they appear centred approximately 25 Hz above the harmonics of the fundamental. The spacing in the specified F2 direction also appears to be every 100 Hz with the regions lying above the harmonics of the fundamental frequency.

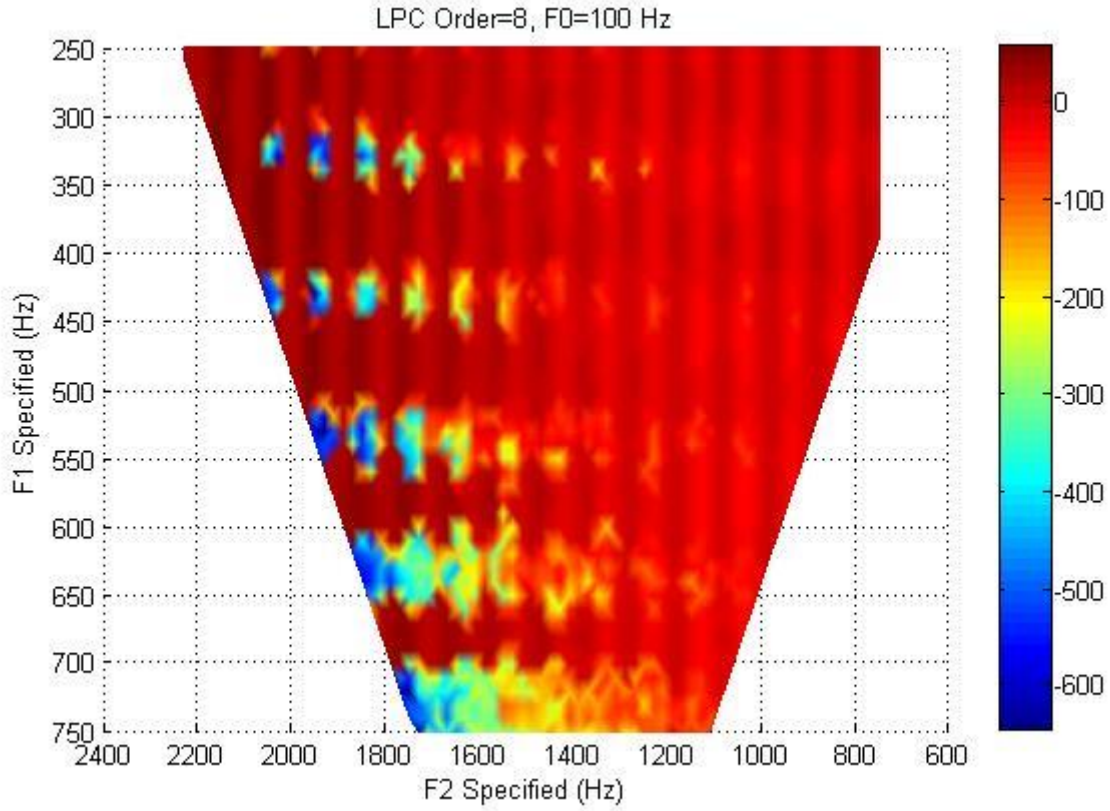

Figure 5.11 F2 error plot with error represented by colour only, for synthetic speaker with glottal source E_e value of 6 at LPC order 8.

Similar patterns of divergent error values were also observed for other speakers with different E_e values as well as the baseline speaker at LPC orders above 9. Again, they were only present for certain combinations of LPC order, speaker and formant. Even though these errors surfaces exhibited some patterning in the location of the divergent regions they were not all as structured as the example shown above.

5.3.5 Summary Data

The error surfaces discussed above demonstrate that altering the glottal source signal can have a marked impact on measurement performance. To gain an overall impression of the variation it is again helpful to summarise the results. Figure 5.12 shows the mean

absolute error for the ten variable glottal source speakers for LPC order 7, 8 and 9 for all three formants.

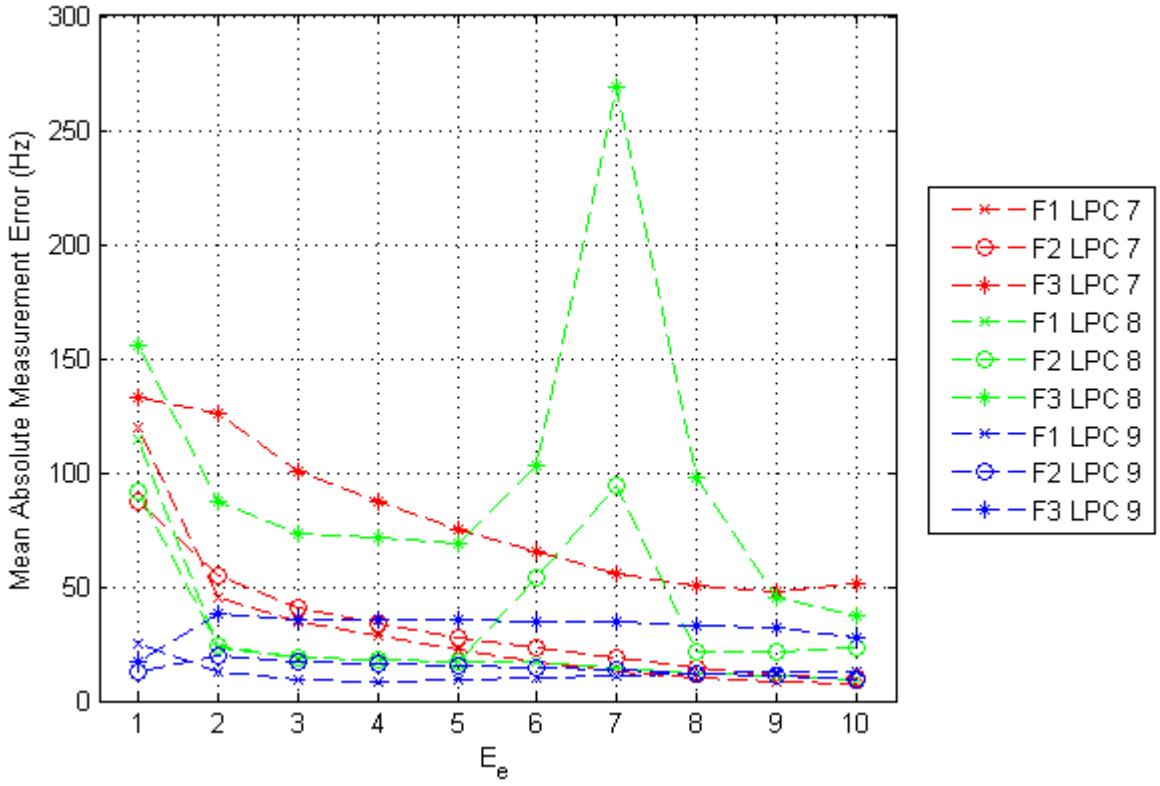

Figure 5.12 Mean absolute measurement error for speakers with varying glottal source with E_e values from 1 to 10 for LPC orders 7, 8 and 9.

At LPC order 7 (red points), the mean absolute error decreases as E_e increases for all formants, whereas at order 9 (blue points), the performance is relatively constant across E_e values. However, comparison across the formants at order 9 shows F3 to perform worse than F1 and F2, except when E_e is 1. This behaviour is different from the baseline speaker, whose results are shown in Table 5.3, where the performance across formants is very similar. The errors from LPC order 8 (green points) show much greater variation in performance across E_e values, with particularly poor performance for F2 and F3 when E_e is 7. This is a consequence of specific regions in the vowel space exhibiting very large negative errors, as shown in Figure 5.9 to Figure 5.11. These types of errors can be considered as formant numbering errors as they result from the method employed by the software to assign formant numbers to pole frequencies, which was previously discussed in Section 4.4.7. In these instances, where a large negative error occurs for the F2 measurement, a pole is present in the model between the true F1 and F2 values. Since this is the second pole, it is labelled as F2, and the third pole, which aligns with the true F2, is now labelled as F3, resulting in the measured F3 value also

being a large underestimate. The true F3 value is now modelled by the fourth pole. The extra pole causes both F2 and F3 to be mislabelled, so both formants show poor performance.

To determine the influence of the additional poles and the numbering errors, Praat's numbering approach was ignored and the measurement errors were recalculated. Table 5.5 shows an example of the effect of this approach on the mean absolute error for the three formants from the speaker with an E_e of 7 at LPC order 8.

Measurement Strategy	**Mean Absolute Error (Hz)**		
	F1	**F2**	**F3**
Default	14.10	94.28	268.80
Ignore Praat Numbering	14.10	37.62	60.23

Table 5.5 Mean absolute errors for speaker with E_e of 7 at LPC order 8 with Praat's default measurements strategy and Praat's formant numbered ignored.

Since no extra poles occurred below the true F1 values over the vowel space, the results for F1 remained unchanged. Whereas F2 and F3 show a very marked improvement in performance when the measured formant values are selected based on their proximity to the true formants. However, the mean absolute error values show that the performance is not as good as that obtained at LPC order 9 for the same speaker.

5.4 Summary

The testing and analysis conducted in this chapter focused on the issue of variation in the accuracy of formant measurements across speakers, which is raised in research question 3. Different synthetic speakers were created by altering the specified F3 values and the glottal source signal of the baseline synthetic speaker from the previous chapter. Formant measurements were made across a range of LPC orders and their accuracy was examined and compared.

The results demonstrate that modifying the baseline speaker to create different synthetic speakers did influence the accuracy of formant measurements. When the F3 values were altered, the greatest impact was on the magnitude of the F3 errors and the structure of the F3 error surfaces. The location and form of the cyclic regions was governed by the arrangement of the specified F3 values within the vowel space. Altering F3 had minimal influence on the performance in relation to F1 and F2. When constant specified F3 values were used it was apparent that the F3 measurement errors were also dependent

on the location of the vowel within the F1~F2 vowel space. Using constant F3 values also clearly demonstrated the dependence of the error behaviour on the specified F3 value as well as LPC order. Since real speakers show different patterns of F3 use (Peterson and Barney 1952, Kasuya et al. 1994) this variation in performance is to be expected across real speakers. Combining these effects with the influence from differences in use of the F1~F2 vowel space is likely to lead to greater performance variation in real speakers.

Changing the glottal source signal from a simple pulse train to a more realistic representation led to considerable variation across the different synthetic speakers both in terms of the structure of the error surfaces and the overall performance. These effects were apparent across all three formants. The main difference between these speakers, and the baseline speaker, was the appearance of localised regions of large errors either at the edge of the vowel space or systemically distributed across areas of the vowel space. Whilst such behaviour was observed at LPC orders above 9 for the baseline speaker, these features were present for the variable glottal source speakers at LPC orders 7, 8 and 9, which otherwise produced the most accurate measurements. Application of an alternative measuring strategy was shown to reduce the influence of these errors on the overall performance.

The changes made to the baseline speaker represent only a very small proportion of the possible ways in which real speakers vary. It would seem reasonable that the greater variation present in real speech would lead to even greater variation in formant measurement errors from real speech. The magnitude of the errors and the variation is also likely to be greater than that observed for the synthetic speakers, since they conform to the assumptions of the LPC model and therefore represent ideal speakers. The variation in performance from real speech is explored in Chapter 6 and Chapter 7.

From a practical perspective, the results from this chapter support the guidance offered in the previous chapter. Additionally, the error surface plots from the variable glottal source speakers demonstrate that for certain speakers, an LPC order that can produce relatively accurate measurements in certain regions of the vowel space can also lead to large errors in other regions. This finding again highlights the potential dangers of using a single LPC order for all vowel tokens or assuming that a single order is appropriate for several speakers, which is recommended by Rose (2002, p. 267). Reassuringly, these

large errors can be mitigated by the application of an alternative measurement approach. The results also show that the performance at different LPC orders will vary across speakers.

Chapter 6 Formant Measurement Accuracy from Real Speech

6.1 Introduction

Chapter 4 and Chapter 5 examined the performance of Praat's LPC formant measurement tool when analysing different synthetic speakers, since their speech could be controlled and analysed in ways that are not possible with real speech. However, it cannot be assumed that the same performance is achievable with real speech, due to the simple speech production model used. This chapter seeks to address the issue by analysing and comparing the behaviour of formant measurements from a large number of real speakers, which provides answers to the third research question:

RQ 3. To what extent does the accuracy of LPC formant measurements vary across speakers?

The methodology involves comparing formant measurements made using Praat's LPC tool across a range of LPC orders for a subset of the TIMIT speech corpus with a set of reference formant values (Deng et al 2006). The reference values allow the accuracy of the measurements to be determined. The measurements are also subjected to different analysis frameworks that replicate the decisions analysts might make when measuring formants interactively. This approach provides further insight to the second research question:

RQ 2. How does altering the LPC analysis parameters affect formant measurement accuracy?

The findings from these experiments demonstrate that whilst the general behaviour of the errors from the real speakers is comparable with synthetic speakers, the magnitude of the errors and range of variation seen across speakers is much greater. The results also highlight the improvements in performance that can be made by allowing LPC order to vary across speakers, tokens and formants. These findings justify the guidance that LPC order should be tailored as specifically as possible to obtain the best performance.

6.2 The VTR Database

The VTR database is a large set of 'ground truth' formant, or vocal tract resonance (VTR), values that have been compiled with the specific aim of facilitating objective testing of automatic formant estimation methods. The formant values were obtained from a subset of speech samples from the TIMIT speech corpus. A sophisticated formant estimation technique (Deng et al. 2004) was initially applied to the samples to produce 'crude estimates' that were subject to 'extensive manual correction' to 'provide accurate VTR' values (Deng et al 2006, p. 370). In this context the 'ground truth' concerns information derived from the speech signal, whereas in Chapter 4 and Chapter 5, the ground truth reference values were properties of the synthetic vocal tract filter.

The TIMIT database (Zue et al. 1990) consists of relatively high quality digitised microphone recordings, of 6300 read sentences spoken by 630 speakers, (438 male, 192 female), from eight major dialect regions of the United States of America. The audio files have a sample rate of 16 kHz. Ten sentences are spoken by each speaker, with a total of 2,342 distinct sentences from three different sets, designed to elicit dialectal differences, and cover an extensive range of phonetic pairs in varying contexts. Each recording is accompanied by a time aligned word transcription and phonetic transcription.

The VTR database contains formant values for a subset of 516 sentences[4] from the TIMIT corpus. For all sentences the first four formant centre frequency and bandwidth values are given at intervals of 10 ms across the entire recording, even for periods of silence, and for speech sounds for which formants would not normally be measured, such as consonants. The authors' motivation for including values for these segments is that resonances are a physical property of the vocal tract, not the speech signal, and they exist even if they are not excited. Of the VTRs provided only the first three formant centre frequency values have been hand corrected. The remaining values are the results from the algorithm (Deng et al. 2004) initially used to produce the 'crude estimates'.

[4] The documentation accompanying the VTR database states that it contains formant values for a total of 538 sentences. However, only data for 516 sentences were provided in the version of the database that was available to download from the Internet.

6.2.1 Limitations of the VTR Database

The VTR database is undoubtedly a valuable resource, but it has limitations that must be considered when using the data and interpreting results derived from it. The formant values cannot be considered as absolute ground truth values, since they are still measurements from the speech signal, which are subject to the same fundamental limitations inherent in all formant measurement techniques (discussed previously in Section 1.2). But, given the extent of checking and manual correction used, the values are likely to be approaching the limits of accuracy that are achievable from pre-recorded speech with the currently available measurement techniques.

Concerns about the 'ground truth' aspect of the VTR database are also raised by Gläser et al. (2010), who used the database in their evaluation of a novel formant tracking technique. They suggest that the formant tracker used to obtain the initial VTR values may benefit from incorporating additional information, such as the speaker's sex, and they state in relation to the manual corrections of values that 'in some cases even visual inspection may not provide means to identify real formant locations' (2010, p. 230). However, they do conclude that they 'nevertheless think that this database provides a reasonable basis for deriving quantitative performance measures' (2010, p. 230). Mehta et al (2012) who also use the database to test a formant tracking algorithm note that it 'only yields estimates of ground truth' (2012, p. 1737).

Another limitation of the database is that some vital information relating to the generation of the measured formant values is not provided in the documentation. This makes the measurement of formants for the purposes of comparison with the database somewhat problematic. The lack of information is surprising given that the purpose of making the database publically available was to allow the comparison of other formant measurement techniques with the 'ground truth' values provided. These issues are discussed in more detail in Sections 6.2.4, 6.2.5 and 7.3.1.

6.2.2 Speech Material Examined

The synthetic speech generated for the previous chapters was highly controlled and was restricted to monotone, stable monophthongs, with limited variation in the glottal source, and with vocal tracts constructed to reflect typical male speakers. The 516 TIMIT sentences used by the VTR database vary much more. The sentences selected were chosen to form a balanced set of speakers, dialects, genders and phonemes. They

are spoken by 186 different speakers, of which 113 are male and 73 are female. For 24 speakers (16 male, 8 female) there are 8 sentences, and for the remaining 162 speakers (97 male, 65 female) there are only 2 sentences. By using such a varied dataset, relative to the previous chapters, the findings from the analyses would be well suited to answer RQ3, concerning variation in measurement accuracy across speakers.

Since the focus of this study is vowel formant measurements, the analysis below is limited to the portions of speech segmented and labelled as vowels within the TIMIT phonetic transcriptions. A total of twenty different vowel categories are used within the transcripts, of which 15 are monophthongs and 5 are diphthongs. Within the 516 sentences there are a total of 6,601 vowel tokens with an average of almost 13 per sentence, of which 5,528 are monophthongs (an average of just less than 11 per sentence) and 1,073 are diphthongs (an average of 2 per sentence).

6.2.3 Determining Formant Measurement Errors in Praat

The general approach used for determining formant measurement errors for synthetic speech was also applied to the real speech. The measurements from the TIMIT samples produced by Praat were compared with the reference values in the VTR database to determine the measurement errors. Again, the measurements were obtained from the 'Sound: To Formant (burg)…' function. However, notable differences exist between the two sets of speech material that influenced the specific measurement and analysis processes used. One of the most significant is that the specified formant values for the synthetic speech were time invariant, whilst the real speech is dynamic and the formant values change across time. It was therefore critical that the measurements made in Praat were compared with values in the VTR database that had originated from the same short section of speech. The way in which this was addressed is discussed in the following sections.

6.2.4 Comparable Measurements – Time Step & Window Length

The analysis parameters in Praat that determine the amount of material contained in each analysis frame and their relative spacing, namely the time step and window length, were selected to be the same as those used to generate the VTR database measurements. The time step, or frame advance, value of 10 ms was provided in the documentation accompanying the database, while the window length, or frame width, value of 25 ms was obtained via a private communication with one of the database's authors (Deng,

2011). The function in Praat used to measure the formants ('To Formant (burg)…') does not provide the option to select the windowing function applied to the analysis frames and information about the function used for the VTR database was not provided in the documentation. The settings used for the other analysis parameters are discussed in Section 6.2.6.

6.2.5 Comparable Measurements – Time Alignment

Even with the same time step and window length settings it was essential that the measurement frames were aligned in time. The crucial information to ensure this occurred were the timings associated with the first measurement frames for the VTR database and the measurements generated in Praat. If the first frame from each set of measurements is aligned then the remainder of the frames will also be aligned since the time step and window length settings are the same. The timing of the first frame is easily obtained for the measurements made within Praat through the execution of a query within the software. However, the equivalent information is not provided for the VTR database in the accompanying documentation and the authors of the database were unable to provide it (Deng 2011). The lack of this critical information is somewhat surprising given that the main purpose of releasing the database is to allow the data to be used for comparative testing.

Several attempts were made to determine the correct alignment of the formant values by both numerical and visual comparison of the VTR measurements with equivalent measurements from Praat at different timing offsets. Unfortunately, none of these approaches provided a satisfactory alignment across multiple utterances. Further attempts to solve this problem were undertaken by Dr Frantz Clermont who was also unable to achieve a satisfactory logically motivated alignment.

No mention of this problem is made by Gläser et al. (2010), but Rudoy et al. (2007) state that their analysis frames were 'left-aligned with the first sample of each TIMIT utterance' (2007, p. 527). This was further checked and confirmed through personal communications between Dr Clermont and the authors (Mehta, 2011). Given that the approach of aligning the left hand edge of the first analysis frame with the start of each recording appeared to have provided a satisfactory alignment for Rudoy et al. (2007) it was adopted for the analyses described below. See Section 7.3.1 for a further discussion of this issue and the approach adopted in Chapter 7.

6.2.6 Other Analysis Settings

Another important difference between the synthetic speech and the TIMIT recordings is that the synthetic speech was generated to represent an average male speaker, whereas the TIMIT recordings are of both male and female speakers. Whilst this has implications for the comparison of the measurement errors across the real and synthetic speakers, this also has an influence on the analysis settings used in Praat. The 'Maximum Formant' setting determines the upper frequency limit of the formant analysis, and for male speakers, including the synthetic speakers previously tested, is normally set at 5,000 Hz. Since female speakers tend to have shorter vocal tracts, and consequently higher formant values, a setting of 5,500 Hz is recommended in the Praat manual for female speakers (Boersma, 2010). Therefore, values of 5,000 Hz and 5,500 Hz were used for the 'Maximum Formant' setting for the male and female speakers respectively.

The only analysis parameter that was a variable was LPC order. As with the previous analyses of the synthetic speech this was varied from 6 to 20 in steps of 1. Varying LPC order meant the data could be considered in relation to RQ2, to further understand the influence of analysis parameters on measurement accuracy. The analysis parameter, 'pre-emphasis (from frequency)', was held constant at the default of 50 Hz.

6.2.7 Implementation

A script was used in Praat to load each TIMIT recording, perform the formant analysis and save the resulting formant measurements in separate log files for each recording and LPC order. Praat exhibits a peculiarity whereby the timing of the initial formant analysis frame is dependent on the duration of the material being analysed. It was therefore necessary to append a specific period of silence to the end of each recording before the formant measurements were made. A further complicating factor is that within Praat the specified window length is the effective duration rather than the true duration, which is twice the specified value due to the Gaussian-like windowing function applied to each analysis frame (Boersma, 2010). To ensure the correct alignment of the analysis frames it was therefore necessary to add a period of silence to the start of each file to compensate for the doubling of the frame width. However, before the formant measurements were saved to log files the timings associated with them were adjusted to reflect the true timings within the original audio files.

For the sake of simplicity of the Praat script and to allow flexibility in the analysis of the resulting data, formant measurements were made for all frames across the recordings. Since the Praat formant measurement process used does not take into account any frame to frame information, each measurement is independent of those from frames surrounding it, so there was no possibility of influence on the measurements from non-vocalic segments. The determination of which measurements originated from vowel tokens was made during the analysis of the results within Matlab.

The log files generated in Praat were loaded into Matlab together with the data from the VTR database and the TIMIT time-aligned phonetic transcripts. The phonetic transcripts were used to determine which analysis frames related to vowel tokens. The transcript files contain a start sample number, an end sample number and a phonetic label for each segment. It was necessary to convert the start and end sample values to corresponding analysis frames. Owing to the length and overlap of the analysis frames, each sample occurs within two or three adjacent analysis frames. The rule applied to determine which one of the two or three frames should be selected as the start or end frame was to choose the one whose centre was closest to the sample. The only exception to this rule was for adjacent vowel tokens. If the rule had been applied in these circumstances then the same frame would have been assigned as the final frame of the first token and the first frame of the second token. This would have resulted in that frame being included twice in the analyses. To avoid this occurring, the following frame was selected as the start frame for the second token.

Once it had been determined which frames corresponded to vowel tokens, the measurement errors were calculated across all the LPC orders. The errors were calculated on a frame by frame basis by subtracting the measured F1, F2 and F3 values from the reference values from the VTR database from the corresponding frame. During the calculation of the errors information concerning the speaker, vowel and frame was retained to allow subsequent analysis of subsets of the data.

6.3 Analysis of Measurement Accuracy

The set of data obtained from the processes described above is both large and relatively complex. The nature of the data means that there are many ways in which it can be analysed, summarised and presented. The following sections consider the results in three ways. The first approach begins by summarising how the entire set of

measurement errors vary across LPC order and provides analysis results that are equivalent to those derived for the synthetic speakers. These results are most applicable to the second research question. The analysis also shows the differences between the results from the male and female speakers. In addition, consideration is also given to how the measurement errors change when different analysis frameworks are applied to the results. For example, one framework requires the LPC order to be fixed across all tokens, whereas another allows it to vary across tokens. A set of 'benchmark' framework results are obtained which represent the best possible performance that can be obtained with the measurement method employed. The other results are then compared with this benchmark set. This again addresses RQ2.

The second approach examines how the measurement errors vary over the vowel space. The results are considered from the different analysis frameworks, as well as how the LPC orders used to obtain them are distributed over the vowel space. The third and final section examines how the performance of the different analysis frameworks varies across speakers. The results show the range of variation found both within and across speakers as well as examining the relationship with factors such as the mean fundamental frequency of the speaker and their location within the vowel space. The range of LPC orders used by speakers is also considered. These analyses are focused on RQ3.

The analyses only consider the errors from the first three formants, F1 to F3, since these were the values that were subject to hand correction within the VTR database. Also, these are the formants most often examined by forensic speech scientist, and phoneticians more generally.

6.3.1 Influence of LPC Order

6.3.1.1 Distribution of Errors

To obtain an overall impression of the influence of LPC order on the behaviour of the errors, and allow them to be compared with the results presented in Section 4.4.2 for the synthetic speech, boxplots were generated showing the distributions of the measurement errors for each formant across LPC order. These are shown in Figure 6.1 to Figure 6.3. To recap, the horizontal red line represents the median value, the lower and upper edges of the blue box are the 25th and 75th percentile, and the black whiskers extend to the limits of the data that are not considered outliers. The outliers are values that fall outside

a range defined as being 1.5 times the interquartile range above the 75th percentile and 1.5 times the interquartile range below the 25th percentile, and are shown as red crosses. Note that the range and scale of the vertical axes are different across the three plots.

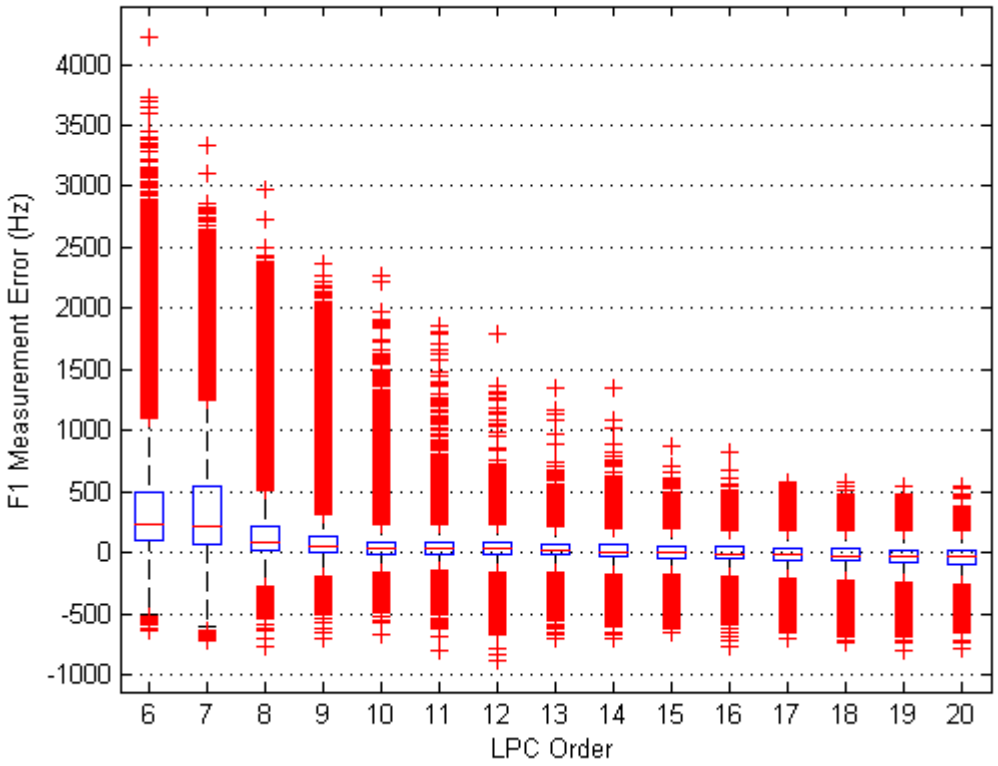

Figure 6.1 Boxplot showing the distribution and variation of F1 measurement errors for all frames from the VTR database across LPC order with Praat's normal measuring tool.

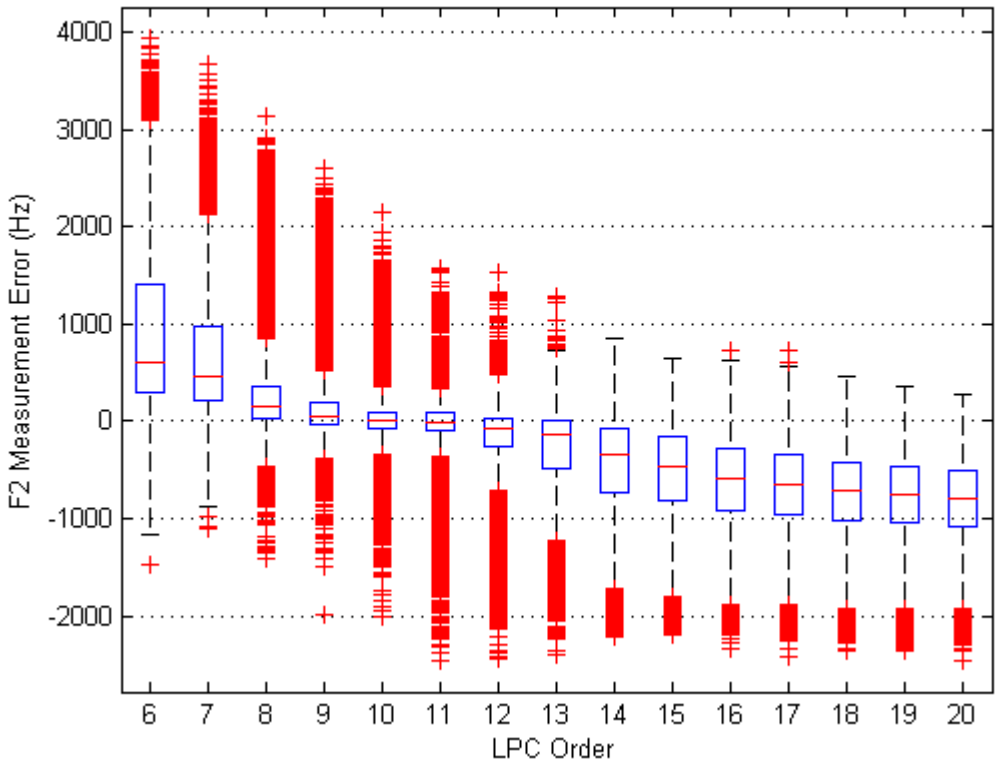

Figure 6.2 Boxplot showing the distribution and variation of F2 measurement errors for all frames from the VTR database across LPC order with Praat's normal measuring tool.

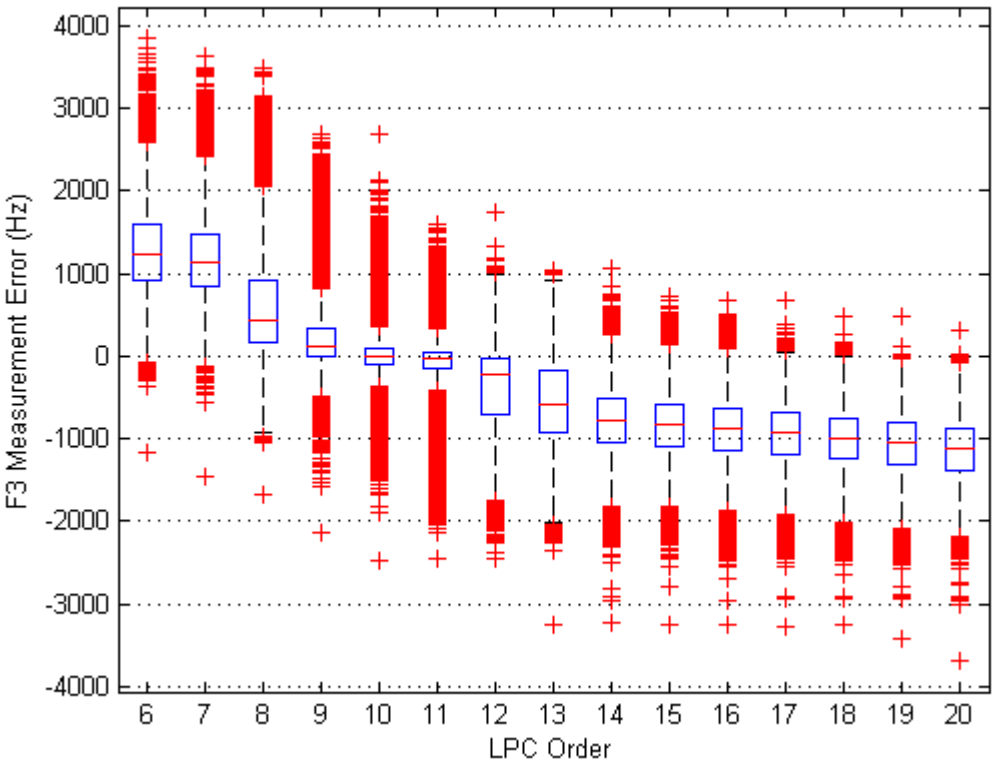

Figure 6.3 Boxplot showing the distribution and variation of F3 measurement errors for all frames from the VTR database across LPC order with Praat's normal measuring tool.

In general terms, the overall behaviour of the errors is the same as that observed for the synthetic speakers. The main difference between the two sets of results is that the errors from the synthetic speakers are smaller. At the lower LPC orders the mean errors are positive meaning that the measurements tend to be overestimates, whilst at the higher LPC orders the mean errors are negative showing that the measurements are generally underestimates. This behaviour is most marked for F2 and F3.

Histograms were also generated to show the distribution of the errors for each formant at each LPC order. These tended to confirm the impression of the distributions that was provided by the boxplots. The distributions at the LPC orders which gave the best results were generally symmetric so the mean absolute error and standard deviation were judged to be suitable measures of performance. The mean absolute errors, which ignore the sign of the error, were found to decrease as LPC order increases, reach a minimum and then increase. This was most apparent for F2 and F3.

For F2 and F3 the smallest mean absolute error occurs at an LPC order of 10. Whilst for F1 the lowest mean absolute error is at LPC order 15. The error values at these LPC

orders are shown in Table 6.1 under the 'All' column, together with the standard deviation and their percentage equivalents.

6.3.1.2 Differences between Male & Female Results

The measurement errors were grouped and analysed according to the sex of the speakers. In general terms, the errors from both the male and female speakers exhibit the same behaviour across LPC order that was described above. The differences between the two groups lie in the magnitude of the errors and the LPC orders at which the smallest errors occur. Table 6.1 shows the results across all frames for male speakers, female speakers, and both sexes combined.

	All		**Male**		**Female**	
F1 LPC Order	15		16		14	
F1 Mean Absolute Error (Hz)	63.68	13.27%	61.37	13.79%	63.70	11.60%
F1 Standard Deviation (Hz)	88.51	20.08%	82.33	20.39%	93.05	18.20%
F2 LPC Order	10		11		10	
F2 Mean Absolute Error (Hz)	125.67	8.28%	113.61	7.86%	139.71	8.23%
F2 Standard Deviation (Hz)	188.57	13.00%	169.06	11.45%	205.54	12.29%
F3 LPC Order	10		10		10	
F3 Mean Absolute Error (Hz)	144.12	5.91%	137.63	5.95%	154.67	5.83%
F3 Standard Deviation (Hz)	228.94	9.69%	220.20	9.79%	242.49	9.53%

Table 6.1 Summary statistical data and percentage equivalents at LPC order with lowest absolute mean errors from VTR database for male speakers, female speakers and all speakers.

The data in Table 6.1 reveals some differences between the male and female speakers. In terms of the mean absolute error and standard deviation across all formants, the values for the male speakers are consistently lower than those from the combined data set, whilst the values from the female speakers are consistently higher than the combined set. This result is to be expected since the male speakers tend to have lower formant values than the female speakers (see Table 6.2), which is a consequence of the shorter female vocal tract. However, in percentage terms the situation is reversed for F1 and F3 where the smallest absolute errors and standard deviations occur for the female speakers rather than the male speakers. For F1 and F2 the LPC order at which the

smallest absolute error occurs is lower for the female speakers than for the male speakers, whilst for F3 it is the same.

The mean values for the reference formant values from the VTR database are shown in Table 6.2 for all speakers, male speakers and female speakers. The percentage difference from the combined set is also given for the male and female group.

	All	**Male**	**Female**
F1 (Hz)	527	493 (-6.5 %)	583 (10.6 %)
F2 (Hz)	1593	1487 (-6.7 %)	1765 (10.8 %)
F3 (Hz)	2520	2384 (-5.4 %)	2743 (8.8 %)

Table 6.2 Mean reference formant values from the VTR database for all speakers, and male and female speakers separately, with percentage differences between male and female speakers and the entire set.

Another way to summarise the results, which makes them easier to compare with the results from other frameworks in the following sections, is to combine the errors from all three formants and calculate the absolute mean and standard deviation in absolute and percentage terms. These results are shown in Table 6.3 for all the results as well as for male and female speakers.

	All		**Male**		**Female**	
F123 Mean Absolute Error (Hz)	111.15	9.15 %	105.63	9.49 %	120.14	8.61 %
F123 SD (Hz)	178.71	15.00 %	169.99	15.66 %	191.62	13.67 %

Table 6.3 Mean absolute error and standard deviation for combined errors from all formants for the VTR database with Praat's normal tool shown for all speakers, and male and female separately, with LPC orders shown in Table 6.1.

Combining the measurement errors from all three formants does not alter the relative performance between the sexes.

6.3.1.3 Overall Performance

The results above consider the performance when the three formants are considered separately and show which LPC orders give the smallest errors for each formant. However, it is also possible to consider the three formants in combination and determine which LPC order provides the smallest errors overall. A criterion needs to be established in order to determine the best overall performance. The two most straightforward are the minimum mean combined absolute error across the three formants and the minimum mean combined absolute percentage error across the formants.

The minimum mean absolute combined error is determined by summing the absolute error for F1, F2 and F3 for each frame at each LPC order. The mean combined absolute error across all the frames is then calculating for each LPC order. The best performance is achieved at the LPC order which has the lowest mean combined absolute error. The same approach is applied to the minimum mean combined absolute percentage error, except that the absolute percentage errors are summed rather than the absolute errors.

For both criteria the smallest overall error occurs at an LPC order of 10. This is the same LPC order that gave the smallest errors for F2 and F3 in isolation. The absolute mean and standard deviations, as well as their percentage equivalents are shown in the table for all three formants at LPC order 10. The mean combined error across the three formants is also shown. This value allows the overall performance of this approach to be compared with the other frameworks discussed below.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	75.61	16.70 %	125.67	8.28 %	144.12	5.91 %	115.13	10.29 %
Standard Deviation (Hz)	144.71	29.72 %	188.57	13.00 %	228.94	9.69 %	184.69	20.32 %

Table 6.4 Mean absolute and standard deviation values at LPC order 10 across all VTR database frames from Praat's normal tool for individual formants and all formants combined.

6.3.2 Analysis Frameworks

The analysis of the measurement errors presented above provides an overall picture of the behaviour of the results across different LPC orders and demonstrates the influence that LPC order has on measurement accuracy. However, the relevance of these summarised results to real world measurement scenarios is somewhat limited as it represents a measurement strategy that involves keeping the LPC order constant across all speakers and vowel tokens. Whilst this approach might be applied when making a large set of automated measurements, it is not to be recommended given that previous work (Harrison, 2004 and Vallabher and Tuller, 2004, 2006) clearly shows that performance is speaker and vowel category dependent, and that errors can be reduced by using different LPC orders.

It is therefore desirable to analyse the data in ways that more accurately reflect realistic analysis strategies, such as modifying the LPC order on a token by token basis or using different LPC orders for different formants of the same token. The behaviour of measurement errors in more realistic circumstances is presented in the following

sections by considering subsets of the measurements which are the equivalent of employing different analysis frameworks. Before examining the analysis frameworks, the results for a benchmark condition are established that considers the best performance that can be obtained from the measurement process applied.

6.3.2.1 Benchmark Performance

The measurement process described above produced formant measurements, and their associated errors, across a range of LPC orders for all vowel frames. Comparison of these results with the reference values from the VTR database makes it relatively straightforward to determine the LPC order at which the smallest errors occur for a given formant, frame, token or speaker. Determining these minimum errors provides a benchmark of the best achievable performance with the measurement process used. Whilst this benchmark performance is not achievable with a realistic analysis framework, the magnitude of the errors and the behaviour of the LPC orders that led to them provides useful information when assessing and comparing the performance of the other measurement strategies introduced below.

To conduct this analysis as part of a manual process would be the equivalent of allowing a different LPC order for each formant within an analysis frame and allowing the LPC orders to vary from frame to frame. Such an approach would be very time consuming and difficult to apply in a normal measurement scenario. It should be noted that within this approach, and those that follow, it is still assumed that the lowest estimated formant frequency corresponds to the first formant, the next lowest to the second formant and so on. In Sections 4.4.7 and 5.3.5 analysis strategies were applied to the measurements which ignored the formant numbering imposed by Praat's tool. It was not possible to apply this approach to the results in this chapter as only the first three formant values were logged when the measurements were made.

6.3.2.2 Benchmark Performance Results

The results below have been obtained by determining the smallest absolute error obtained for each formant for each frame regardless of the LPC order. The same previously used statistical measures have then been calculated across this set of minimum errors. Again, percentage as well as absolute values have been calculated.

	F1		F2		F3	
Mean Absolute Error (Hz)	25.79	5.98 %	58.00	3.74 %	71.09	2.91 %
Standard Deviation (Hz)	46.21	12.43 %	91.21	5.85 %	111.08	4.61 %

Table 6.5 Mean absolute error and standard deviation for measurements from VTR database with Praat's normal tool when LPC order is free to vary across frames and formants - the benchmark condition.

Comparison of the benchmark minimum possible error statistical results in Table 6.5 with those from Table 6.4, where the LPC order is 10 for all frames and formants, shows a very large change in performance. In terms of the mean absolute error, they have reduced by approximately half between the two situations. For F1 the mean absolute error has decreased from 75.61 Hz (16.70 %) to 25.79 Hz (5.98 %), for F2 from 125.67 Hz (8.28 %) to 58.00 Hz (3.74 %) and for F3 from 144.12 Hz (5.91 %) to 71.09 Hz (2.91 %). The variability of errors, measured as standard deviation, also shows a reduction of approximately a half.

Comparison of the results from the combination of all three formants with those from a fixed LPC order of 10 shows the same trends. All the measures shown for the benchmark case are less than half for the LPC order 10 case.

	F123	
Mean Absolute Error (Hz)	51.63	4.21 %
SD (Hz)	87.29	8.54 %

Table 6.6 Combined absolute mean error and standard deviation across all three formants for benchmark case.

When determining the minimum error for each formant in each frame a record was retained of the LPC order that had produced each minimum error. This was done to enable an analysis of the LPC orders that resulted in the minimum errors. A summary, in terms of the median, mode and interquartile range for LPC order are shown in Table 6.7.

	F1	F2	F3
LPC Order Median	15	10	10
LPC Order Mode	20	10	10
LPC Order Interquartile Range	9	3	2

Table 6.7 Summary statistics of LPC orders resulting in the minimum formant errors for the benchmark case.

The results in Table 6.7 reflect the earlier findings in Section 6.3.1 that the best performance for F2 and F3 occurs at LPC order 10. The median value of 15 for F1 is

also the same as the single order that produced the overall best performance in Section 6.3.1. However, the mode shows that order 20 was encountered most frequently. Order 20 was used more than twice the number times of the second most frequent order, which was 19. The distribution of the orders for F1 was relatively uniform apart from a peak at 20. For F2 and F3 the distributions were much narrower, as reflected by the interquartile range, and roughly symmetric, with a slight positive skew. For each formant, the full range of LPC orders from 6 to 20 was encountered.

When interpreting this data it must be remembered that the reference values used to obtain these results are still only estimates themselves. Therefore, what is most important is the change in performance across different analysis frameworks rather than the magnitude of the errors.

6.3.2.3 Other Analysis Frameworks

The previous sections 6.3.1.3 and 6.3.2.2 have considered the measurements in terms of two extreme analysis frameworks, the first with the LPC order restricted to a single value across all frames and formants, and the second with no constraint on LPC order. The following sections consider intermediate frameworks with differing constraints, some of which are equivalent to realistic approaches that could be applied by human analysts.

The first approach allows the LPC order to be different across each of the three formants but requires that it remains constant for each formant within each vowel token. The second approach allows the LPC order to vary within a token, i.e. from frame to frame but it must be the same across the three formants within each frame. The third approach combines the previous two so that the LPC order is constant across the three formants within a vowel token. Whilst the first framework could be easily adopted when manually measuring formants, the second approach may be difficult to achieve in the real world. The combined third framework also represents a realistic approach to measuring formants and is perhaps the one most often adopted by analysts when making computer assisted measurements.

6.3.2.4 LPC Order Fixed within Tokens, Variable Across Formants

In this first approach each of the three formants is considered in isolation since the framework allows a different LPC order for each formant. However, for each formant

the LPC order must remain the same within individual vowel tokens. In order to determine which LPC order produces the smallest overall error within a token, the combined errors from each frame of that token must be considered. As discussed previously in section 6.3.2.1, this requires a criterion to determine which LPC order produces the smallest errors. The same criteria of smallest mean absolute error and smallest mean absolute percentage error are adopted for this, and the following two frameworks. However, unlike the earlier applications of the criteria, this framework requires the mean absolute error is calculated by combining the measurements for each frame across a token, rather than across the three formants, since the LPC order constraint is across the token not the formants.

Applying this framework to the entire set of measurements results in a subset of measurements and associated LPC orders that represent the minimum errors achievable under these conditions. The summary statistics for the set of measurements found with the minimum mean absolute error criterion are presented in Table 6.8.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	43.11	9.26 %	96.07	6.29 %	102.70	4.21 %	80.63	6.58 %
Standard Deviation (Hz)	63.60	15.43 %	139.02	9.20 %	152.08	6.31 %	124.56	11.13 %

Table 6.8 Mean absolute error and standard deviation for measurements from VTR database with Praat's normal tool when LPC order was fixed across individual tokens but varied across formants with minimum summed absolute error criterion.

A statistical summary of the LPC orders that gave rise to these results is shown in Table 6.9.

	F1	F2	F3
LPC Order Median	15	10	10
LPC Order Mode	20	10	11
LPC Order Interquartile Range	8	3	1

Table 6.9 Summary statistics of LPC orders that gave rise to minimum errors when LPC order was fixed across individual tokens but varied across formants, with mean absolute error as minimum criterion.

The results show that the performance in terms of mean absolute error for each formant lies roughly halfway between that obtained for the benchmark case and that where the LPC order is held constant across all analysis frames. The percentage mean absolute error and standard deviation results are also similarly located approximately centrally

between the results from the two extreme frameworks. The combined formant errors are also positioned between the two sets of results from the other frameworks. The LPC orders, in terms of their distributions and summary statistics for each formant, are very close to those that produced the benchmark results. This is perhaps to be expected since each formant is considered independently of the other two as is the case in the benchmark framework. However, the interquartile ranges are slightly reduced in the current framework for F1 and F3, meaning that the constraint of the LPC order within a token has resulted in slightly less variation in LPC order.

The results from using the summed percentage error criterion for the current framework where the LPC order is held constant across a token is shown in Table 6.10.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	43.27	9.22 %	96.30	6.28 %	102.73	4.20 %	80.77	6.57 %
Standard Deviation (Hz)	63.96	15.20 %	140.27	9.14 %	152.23	6.31 %	125.13	10.99 %

Table 6.10 Mean absolute error and standard deviation for measurements from VTR database with Praat's normal tool when LPC order was fixed across individual tokens but varied across formants with minimum summed percentage error criterion.

The summary statistics for the LPC orders that produced these results are identical to those in Table 6.9 for the same LPC order constraint with the minimum absolute error criterion. The distributions of the orders for each formant were very similar across the two conditions.

Comparison of Table 6.8 with Table 6.10 shows that the error results are also very similar. The similarity is not surprising given that the LPC order constraint only applies within a formant across a frame, not across all three formants. The consequence of this is that the summed absolute and summed percentage errors for an individual formant across a frame will tend to track each other as the LPC order is changed. A different outcome is seen below in the following frameworks where the LPC order constraint is applied across the formants.

6.3.2.5 LPC Order Fixed Across Formants, Variable Across Frames

For the second approach, the LPC order is considered as being fixed across the formants for the frame being considered but the order can change from frame to frame. Since three measurements for each frame are being used to determine the LPC order at which the minimum error occurs, a criterion must be applied that specifies what constitutes the minimum error. Again, the sum of the absolute errors across the three formants in a frame and the sum of the absolute percentage errors across the three formants are used.

Applying this framework and the summed absolute error criterion to the entire set of results gives a subset of measurements where the associated errors are summarised by the figures shown in Table 6.11.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	69.09	15.06 %	97.53	6.46 %	81.36	3.34 %	82.66	8.29 %
Standard Deviation (Hz)	83.8	20.86 %	137.66	9.27 %	124.70	5.17 %	119.59	14.53 %

Table 6.11 Mean absolute error and standard deviation for measurements from VTR database with Praat's normal tool when LPC order was fixed across formants but varied across frames, minimum criterion was summed absolute error.

Unlike the previous framework, where a different LPC order could be used for each of the three formants, the current framework applies the same LPC order to all three formants. The median and the mode of the LPC orders that gave rise to the errors above are both 10 and the interquartile range is 2. The distribution was roughly symmetric, with a slight positive skew, and the full range of LPC orders from 6 to 20 was encountered. This behaviour is very similar to that found for F2 and F3 in the other frameworks presented above, showing that the errors for F2 and F3 have the greatest influence on the determination of the LPC order used. This is to be expected as the F1 errors are less variable across LPC order.

Examination of the error results from the three formants reveals that unlike the previous frameworks, F3 has a mean absolute error that is smaller than F2's. This is a consequence of the minimum absolute error criterion. Since F3 tends to have the largest errors the criterion effectively reduces the size of the error associated with F3 measurements. Whilst it might be expected this would result in the F3 errors being

smaller than for the previous framework, it is surprising that overall the F3 errors are actually less than those for F2.

Comparison of the F2 errors with the previous framework shows them to be similar, whilst the F1 errors are higher. Since the F3 errors are smaller it then becomes harder to assess which approach gives the best overall performance based on the results from individual formants. The combined F1, F2 and F3 errors allows the overall performance to be assessed. The current framework has a combined mean absolute error of 82.66 Hz or 8.29%, whilst the previous framework has slightly better performance with 80.63 Hz or 6.58%. However, the standard deviation for the previous framework is higher than for the current one.

Application of the minimum sum of absolute percentage errors to the current framework leads to a subset of formant measurement errors with the summary results shown below.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	60.98	13.24 %	97.13	6.34 %	105.95	4.31 %	88.02	7.97 %
Standard Deviation (Hz)	75.00	18.31 %	138.87	8.98 %	172.32	6.96 %	137.10	13.36 %

Table 6.12 Mean absolute error and standard deviation for measurements from VTR database with Praat's normal tool when LPC order was fixed across formants but varied across frames, minimum criterion was summed percentage error.

The LPC orders that produced these results have identical summary statistics to those described above for the same LPC order constraint with the minimum absolute error criterion, i.e. a median and mode of 10, and an interquartile range of 2. The distributions of the orders for each formant were very similar across the two conditions. Comparison of the error results with those from the absolute mean criterion show very similar results for F2, whilst the F1 values have decreased and the F3 values have increased and risen above those for F2. The reduction in the F1 error is to be expected since overall, F1 tends to produce the largest percentage errors and the criterion is minimising this measure. Therefore it will have a larger impact on the results of F1. The combined F123 absolute mean error is higher for the percentage criterion, but the percentage mean error is smaller, again a consequence of the criterion minimising the percentage error.

6.3.2.6 LPC Order Fixed within Tokens and Across Formants

The final measurement framework requires the LPC order to remain constant within each token and be the same across the three formants. The error results from this framework are shown in Table 6.13. Again, in the first instance the criterion for determining the minimum error is to sum all the absolute errors across each formant for the entire token.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	70.84	15.44 %	108.76	7.16 %	107.51	4.40 %	95.70	9.00%
Standard Deviation (Hz)	90.57	22.52 %	152.95	10.24 %	158.94	6.57 %	139.33	15.72%

Table 6.13 Mean absolute error and standard deviation for measurements from VTR database with Praat's normal tool when LPC order was fixed across individual tokens and fixed across formants, minimum criterion was summed absolute error.

These results were obtained with LPC orders that had a median of 10, a mode of 11 and an interquartile range of 1. Again, the distribution of orders was approximately symmetric with a slight positive skew, although it was narrower than for the previous condition, which is reflected by the smaller IQR. Also, the range of orders encountered was reduced with a minimum order of 7, and a maximum of 16. This is in contrast to all the other previous frameworks where the full range of LPC orders from 6 to 20 was encountered, apart from for F3 under the constant LPC order within a token situation where the maximum LPC order was 18.

Unsurprisingly, the combination of the two previous frameworks has resulted in measurement errors that are greater than either of those produced by the frameworks when applied individually. Again, the absolute F3 error is smaller than the F2 error, but by only 1 Hz, rather than the difference of 16 Hz seen in the previous framework under the absolute criterion.

Application of the minimum summed percentage error criterion to the combined frameworks leads to the summary statistics shown in Table 6.14.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	66.29	14.43 %	110.06	7.19 %	118.72	4.83 %	98.36	8.82 %
Standard Deviation (Hz)	83.05	20.47 %	155.29	10.17 %	178.20	7.24 %	146.58	14.74 %

Table 6.14 Mean absolute error and standard deviation for measurements from VTR database with Praat's normal tool when LPC order was fixed across individual tokens and fixed across formants, minimum criterion was summed percentage error.

The summary statistics and distribution for the LPC orders remains similar to those from the absolute error criterion with a median of 10, but the mode increased to 11. The IQR was again 1, but the range increased as the highest LPC order encountered was 20.

Overall, the results in absolute terms from the minimum percentage criterion are worse than those for the absolute criterion, but the situation is reversed when the percentage results are considered. This again repeats the patterns seen for the two frameworks in isolation, albeit with very small differences between the two minimum error criteria for the first framework.

6.3.2.7 Summary of Results From Different Frameworks

The clear pattern that emerges from the results above is that the greater the restriction on the variation of the LPC order, the greater the size of the errors. The smallest errors occur when the LPC order is free to change across formants and frames (i.e. the benchmark condition), whilst the largest are when the LPC order remains constant across all formants and analysis frames. In terms of the intermediate analysis frameworks, constraining the LPC order across the frames of a token results in smaller errors than constraining the LPC order across the three formants. The combination of these constraints produces a further increase in the magnitude of the errors. The criterion used to determine which measurements at which LPC order constitutes the best measurement, or minimum error, also has an impact on the results. For all the frameworks, the absolute minimum mean error criterion results in an absolute mean error that is less than when the absolute minimum percentage mean error is used. The situation is reversed for the percentage mean errors. However, the results from the different criterion for a given framework are relatively close.

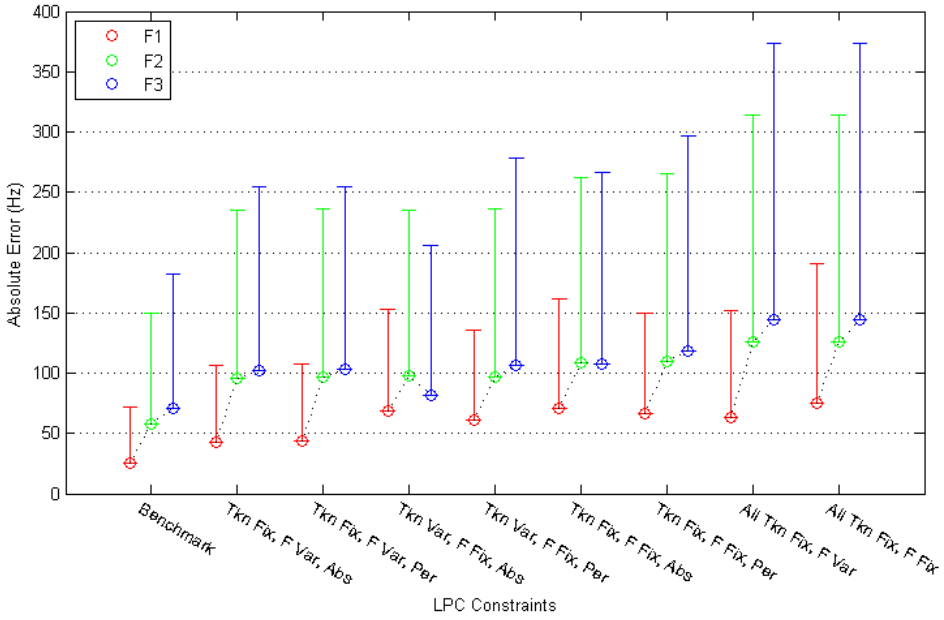

Figure 6.4 Mean absolute error (circles) and standard deviation (line extending 1 SD above mean) across 9 LPC variation conditions. (Key to conditions: Tkn = Token, F = Frame, Fix = Fixed, Var = Variable, Abs = Absolute Error Criterion, Per = Percentage Error Criterion).

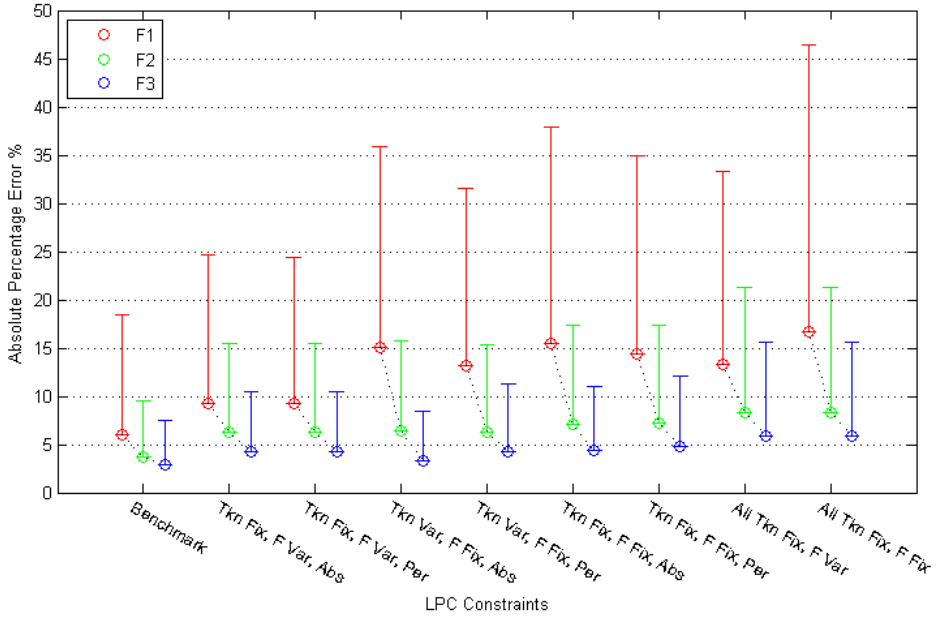

Figure 6.5 Percentage mean absolute error (circles) and standard deviation (line extending 1 SD above mean) across 9 LPC variation conditions. (Key to conditions: Tkn = Token, F = Frame, Fix = Fixed, Var = Variable, Abs = Absolute Error Criterion, Per = Percentage Error Criterion).

Figure 6.4 and Figure 6.5 show the results from each of the nine measurement frameworks previously examined in terms of mean absolute error and mean absolute

percentage error. The results are ordered according to increasing combined error across the three formants for the absolute error results. The circles represent the mean absolute error for each formant whilst the vertical bars extend one standard deviation above the mean. The naming convention for the LPC order constraints is Tkn = Token, F = Formant, Fix = LPC order fixed, Var = LPC order variable, Abs = Absolute minimum error criterion, Per = Percentage minimum error criterion.

The clear difference between the two sets of results is that in the absolute error case for each set of results the F1 errors are always smaller than those for F2, which in most cases are smaller than those for F3. By contrast, for percentage error the F1 results are always larger than the F2 errors, which are in turn always larger than the F3 errors.

6.3.3 Distribution of Errors Across the Vowel Space

The analysis of the synthetic data in the earlier chapters has shown that one of the factors that influences the errors associated with LPC derived formant measurements is the vowel category or location of the vowel token within the vowel space. The following sections examine how the errors and associated LPC orders derived from the various frameworks behave within the vowel space. The section begins by examining the distribution of the reference formant measurements within the vowel space. This is followed by the results from the various analysis frameworks considered over the vowel space. The results are presented in the same order that they were above. Only an illustrative subset of the generated plots has been included.

6.3.3.1 Distribution of Vowels Within the VTR Database

The subset of the TIMIT corpus used for the VTR database was specifically selected to contain a balanced representation of speakers, dialects, genders and phonemes (Deng et al. 2006, p. 370). Therefore, it is to be expected that the vowel space will be well represented. The vowel tokens within the TIMIT corpus have been labelled with vowel categories, but these classifications will not be used in the analyses presented below. This is for two reasons. Firstly, vowel categories are a linguistic construct motivated by the perceptual and phonological properties of vowels. Whilst the categories are clearly linked to the acoustic properties of vowels, namely the formant frequencies, they are not defined by them. Secondly, there are only twenty categories used within the corpus, of which fifteen are monophthongs and five are diphthongs. There is overlap between the categories in terms of F1~F2 values, and the amount of the vowel space covered by the

categories is also different. This makes it problematic to compare performance across the categories. Also, the categories are potentially too broad to provide the resolution necessary to observe patterns or tendencies within the results. Therefore, as for the previous analyses of the synthetic data, the location of the vowels within the vowel space will be defined by their reference F1 and F2 values, and in some instances their F3 values.

For the purposes of these analyses each speech frame is considered independently, rather than within the context of the vowel token that it is a part of. There are a total of 67,424 analysis frames. The F1, F2 and F3 values for these frames are summarised in Table 6.15. This shows the extent of the range of the reference formant values within the database.

	F1	**F2**	**F3**
Mean (Hz)	527.45	1593.00	2520.34
SD (Hz)	132.78	384.98	358.66
Min (Hz)	113.74	638.53	1218.41
Max (Hz)	1131.77	3048.02	4030.12[5]

Table 6.15 Summary statistics for reference formant values in the VTR database relating to vowels.

The overall vowel space represented by these values relate to many speakers, both male and female, so it is obviously much larger than that which would be expected for any single speaker. To be able to determine statistical measures for the errors across the vowel space it is necessary to divide it into small regions and then analyse the frames located in each region.

Before analysing the measurement errors the distribution of the specified formant values across the vowel space was determined. This is shown in Figure 6.6 as a surface plot. The F1 values were divided into bins between 100 Hz and 1,150 Hz. with a width of 50 Hz, creating 21 bins. The F2 values were divided in to 100 Hz wide bins between 575 Hz and 3,075 Hz, resulting in 25 bins.

[5] The maximum reference F3 value within the database is 5206 Hz. This is an erroneous value which has been excluded from the summary statistics in Table 6.15.

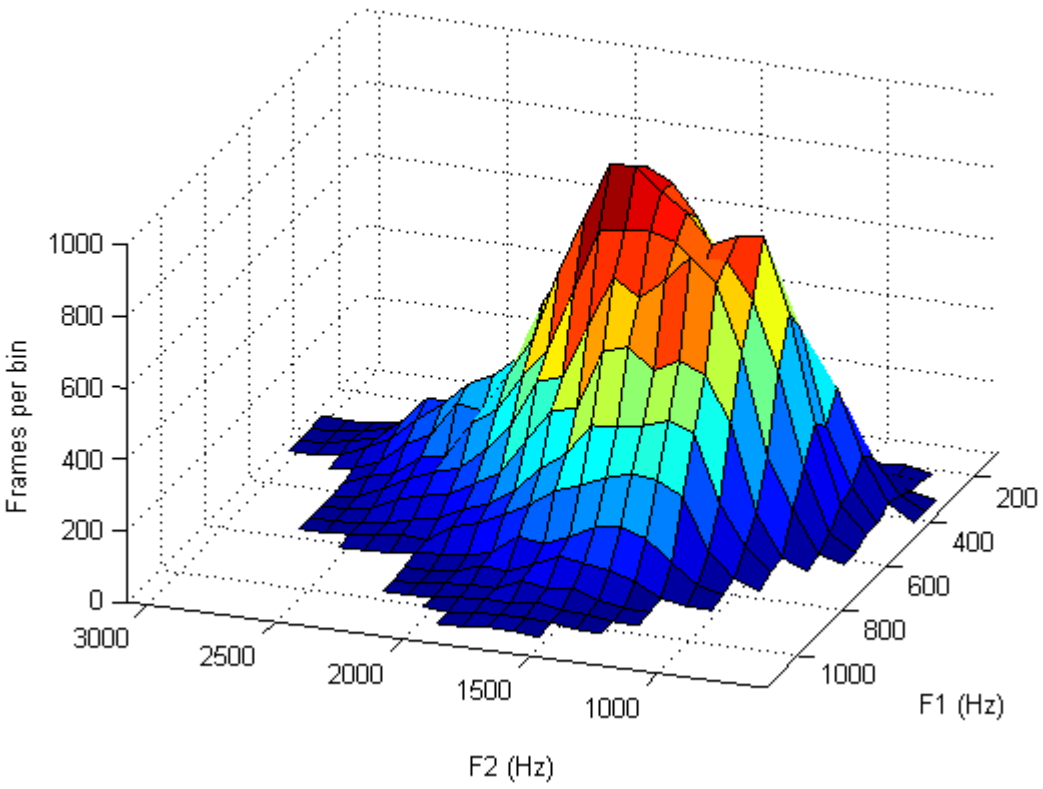

Figure 6.6 Distribution of F1 and F2 reference values from the VTR database shown across the F1~F2 vowel space.

It is clear from the plot that there is a central tendency in the data where the greatest number of frames occurs. The central peak in the plot is around an F1 value of approximately 525 Hz and an F2 value of 1,800 Hz.

6.3.3.2 Distribution of Errors With Constant LPC Orders

Following the same order of presentation of the analyses of the data considered above, the first situation examined is where the LPC order is held constant for the analysis of all frames of data but the results for each formant are considered at different LPC orders.

The mean error values for F1 at an LPC order of 15 are shown in Figure 6.7. This is the LPC order at which the smallest average absolute error of 63.68 Hz occurred across all frames.

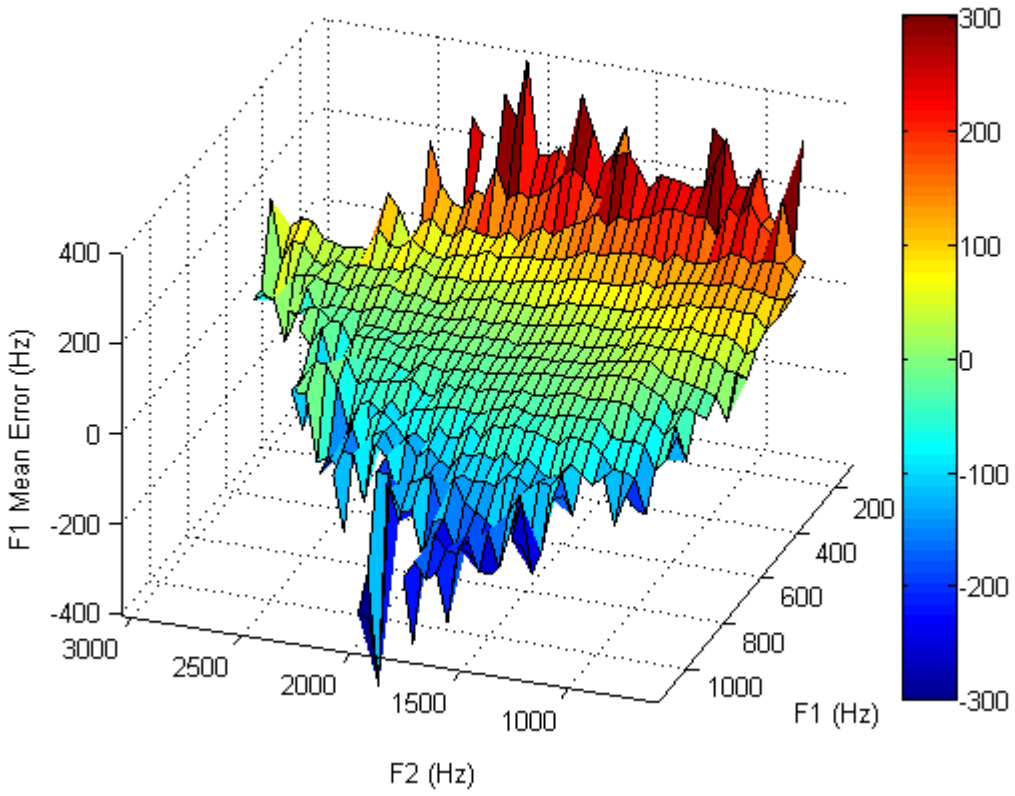

Figure 6.7 F1 mean error surface over the F1~F2 vowel space for all the VTR database vowel frames measured using Praat's normal tool with an LPC order of 15.

The plot clearly shows that the mean measurement error is dependent on the location of the vowel within the F1~F2 vowel space. The dependency on the F1 value is much greater than it is for F2. The vowels with lower specified F1 values tend to produce errors greater than 0, i.e. overestimates, whilst the higher specified F1 values produce negative error values, i.e. underestimates. The F1 values around which the errors cross from positive to negative are in the region of 400 to 600 Hz. The equivalent absolute error surface plot has a concave or U shape, most pronounced in the F1 direction. This is a consequence of the underestimates found at the higher specified F1 values becoming positive errors when the absolute value is determined. Also, the entire surface is shifted upwards in the positive direction with the lowest point of the surface being around 50 Hz. This is because the absolute mean values remove the effect of the underestimates and overestimates cancelling each other out. The distribution of the standard deviation values across the F1~F2 space is very similar in structure to the absolute mean surface, as it shows the greatest variation in the means towards the extremes of the F1 range, more so at the higher specified values, with minimal influence from F2 across the space.

The error surfaces from LPC orders above and below 15 are not dramatically different, apart from those at the lower orders of 6 and 7. The magnitude of the errors at these LPC orders is so large that the results are not really meaningful in terms of their distribution across the vowel space. The general consistency in the results across the LPC orders is to be expected given the relatively stable performance of the F1 measurements shown in Figure 6.1.

The surfaces produced from the errors expressed as percentages are different in one significant respect from those described above. In the error surface plots of the numeric error values, the magnitude of the errors at the extremes of the reference F1 values is approximately the same. However, for the percentage errors the excursion at the lower F1 values is much greater than at the higher F1 values. This is simply a consequence of the results being expressed as percentages. At the lower reference F1 values a given error is much larger in percentage terms than the same error at a higher reference F1 value. The same pattern is also observed for the standard deviation results expressed in percentage terms.

The error surface for F2 at an LPC order of 10, which produced the smallest average absolute error of 125.67 Hz, also shows variation across the F1~F2 vowel space (Figure 6.8). In contrast with the error surface for the F1 error values, in Figure 6.7, the F2 error surface shows dependency of the errors on both F1 and F2. The largest errors occur in the region with the lowest F1 and F2 values (close back vowels) whilst the largest negative errors occur with the highest F1 and F2 values (open front vowels). The region in which the errors change from being over estimates to underestimates is across a band that runs from low F1 values and high F2 values (close front vowels) to high F1 values and low F2 values (open back vowels).

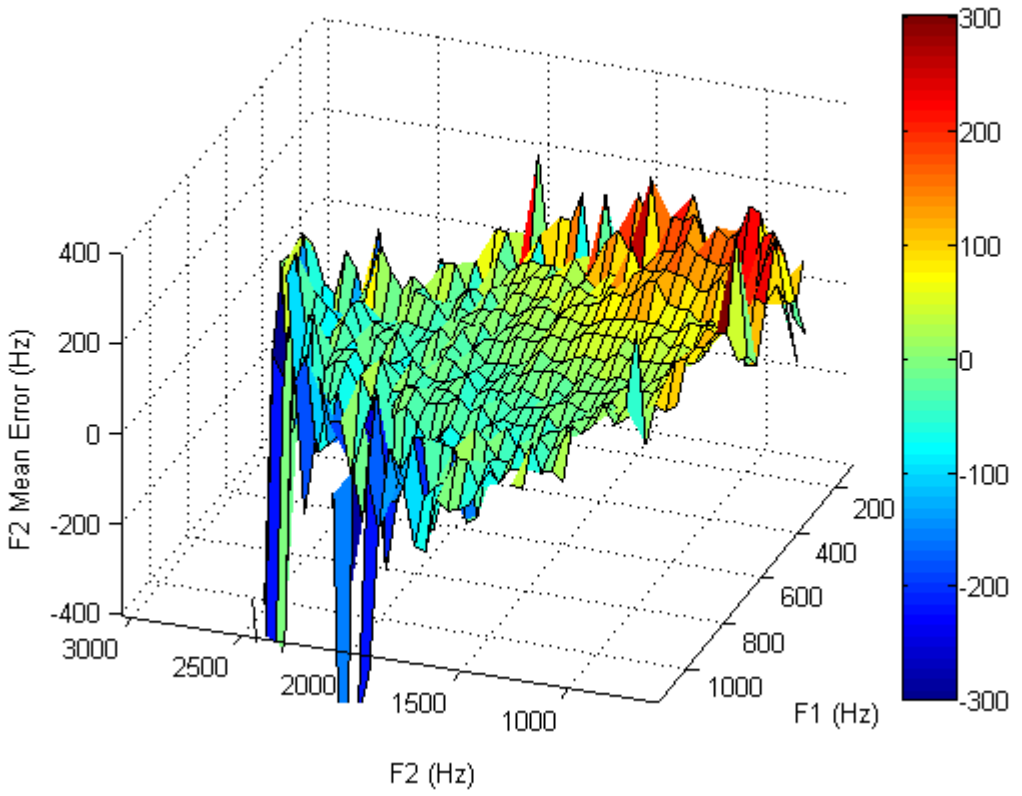

Figure 6.8 F2 mean error surface over the F1~F2 vowel space for all the VTR database vowel frames measured using Praat's normal tool with an LPC order of 10.

The F2 absolute error surface is again a concave shape or U shape with the highest errors occurring at the low and high F1~F2 extremes. The shape of the standard deviation surface is again the same as that for the absolute error surface. When considering the errors in percentage terms, the same difference found for the F1 errors is again observed. The magnitude of the percentage errors at the higher F1 and F2 values are less than at the lower F1 and F2 values.

At the lower LPC orders of 6 and 7 the errors are so large that their distribution is not really relevant. At LPC order 8 the region with the higher specified F1 and F2 values produces errors in the region of 100 Hz, whilst for the lower specified F1 and F2 values the errors are considerably higher. At LPC order 9 the surface is similar to that at order 10 but the low F1~F2 region with the higher errors covers a larger proportion of the surface. As the order increases above 10 the region with the higher F1~F2 specified values produces the largest negative errors and the size of this area increases through the orders. At order 13 the majority of the surface has very large negative errors with only a small section with low specified F1 and F2 values producing relatively small errors.

The error surface for F3, again at LPC order 10, shown in Figure 6.9, which produced the smallest mean absolute error of 144.12 Hz, is very similar in structure to the F2 error surface. The absolute error and standard deviation plots are also similar with the greatest errors and deviations being found in the regions with the highest and lowest F1 and F2 reference values. Altering the LPC order has the same effect on the error surfaces described for F2 above. When examining the F3 error results in percentage terms, the same differences described for the F1 and F2 percentage results are also observed.

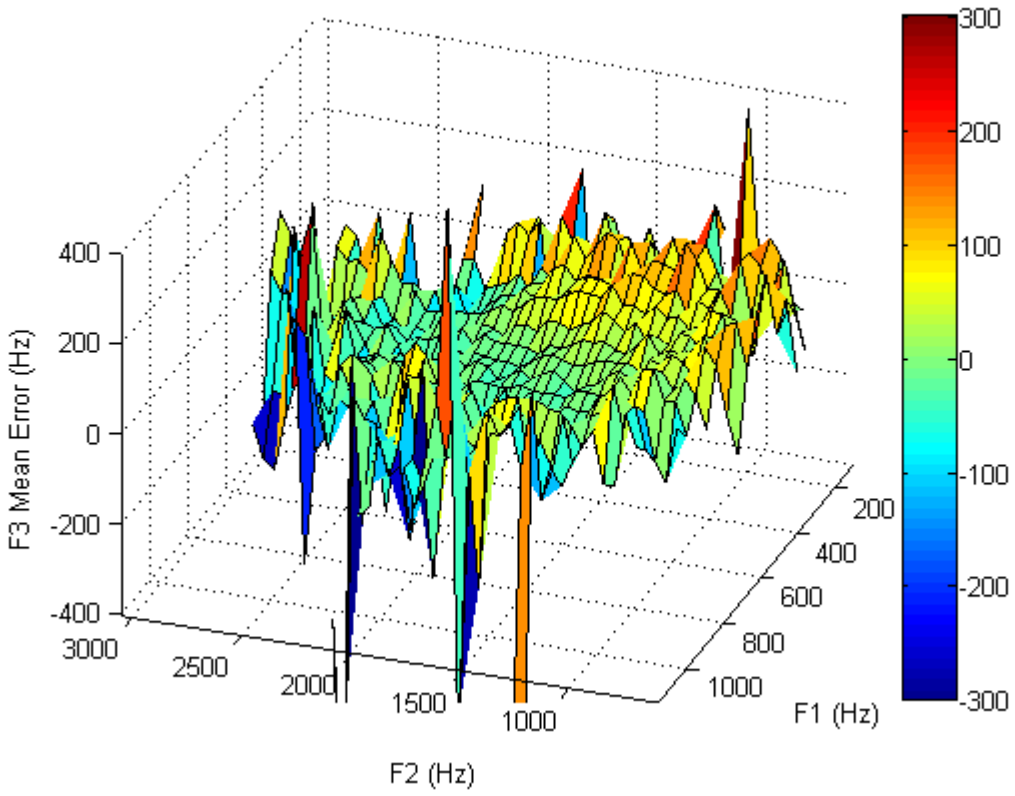

Figure 6.9 F3 mean error surface over the F1~F2 vowel space for all the VTR database vowel frames measured using Praat's normal tool with an LPC order of 10.

Whilst it is useful to examine how the F3 errors behave across the F1~F2 vowel space it is not possible to determine from such a plot if the F3 errors are dependent on the reference F3 values. Since it is clear that F1 and F2 errors are dependent on their specified values it is worthwhile considering the F3 errors against the specified F3 values. When the F3 errors are plotted on an F2~F3 vowel space then a clear dependency is visible between the specified F3 values that is similar in nature to that exhibited for F1 and F2 and their respective specified values. At the lower F3 values the measurements are overestimates, i.e. the errors are positive, whilst at higher specified

F3 values the measurements are underestimates and the errors are negative. Over the F2~F3 vowel space the mean absolute errors and the standard deviation are the highest at the edges of the space where the number of analysed frames is also at its smallest.

In general, these results show that there is a clear dependency of the measurement errors on the specified reference formant values. For the LPC orders that resulted in the smallest errors, the lower specified values tend to have overestimated measurements, i.e. positive errors, whilst the higher specified values have underestimated measurements, i.e. negative errors. This pattern is clear across all three formants. A consequence of this is that the smallest errors occur in the central region of the vowel space where there is a relatively large and even spacing between the formants. The largest errors tend to occur towards the extremes of the vowel space. These are also the areas where the least number of frames exist. The tendency towards the central area is also a consequence of the greater number of frames within that area which biases the selection of the LPC order at which optimum performance occurs.

6.3.3.3 Distribution of Errors Over Vowel Space for Benchmark Case

Having examined how the errors are distributed across the vowel space for the situation where the LPC order is held constant across all frames, the following section considers the distribution of errors where the minimum possible error for each frame and formant is determined, i.e. the benchmark scenario presented in Section 6.3.2.1. Figure 6.10 shows the distribution of F1 mean errors across the F1~F2 vowel space for the benchmark case.

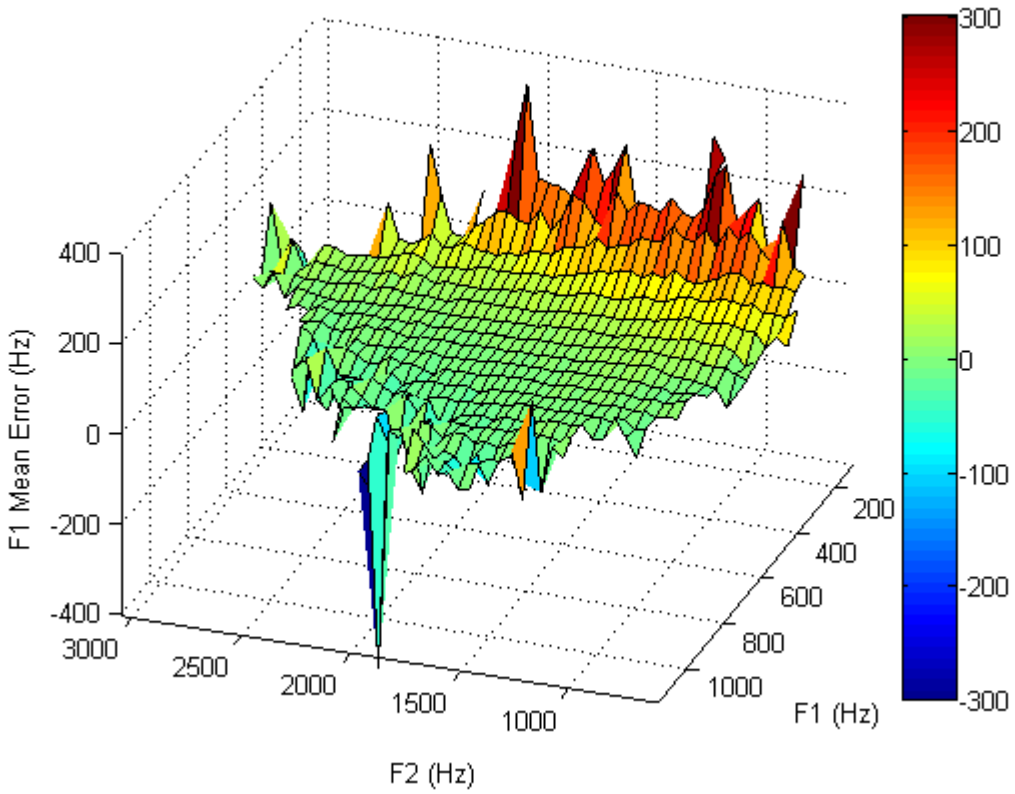

Figure 6.10 F1 mean error surface over the F1~F2 vowel space for all the VTR database vowel frames measured using Praat's normal tool for the benchmark LPC order variation case.

Comparison of Figure 6.10 with Figure 6.7 reveals that for the benchmark situation there is still a similar dependency on the F1 values, but the magnitude of the errors at the lower specified F1 values is less than when the LPC order is fixed at 15. Also, the negative errors at the higher F1 values seen in the fixed LPC condition are almost non-existent in the benchmark situation. Comparison of the absolute errors reveals a less marked U shape with the only significant excursions occurring at the lower specified F1 values and a large area of stable errors within the centre of the vowel space. The results expressed as percentages reveal the same patterns.

Figure 6.11 shows the mean F2 errors across the F1~F2 vowel space. Comparison with the equivalent plot when the LPC order is fixed at 10 (Figure 6.8) shows that, as expected, the magnitude of the errors is much less for the benchmark results, the dependency on the specified F1 values is no longer apparent and the direction of the dependency on the specified F2 values has changed. In the benchmark situation the lower F2 values are resulting in underestimates, whilst the higher F2 values are leading to measurements that are overestimates. Plotting the absolute mean F2 error reveals a surface very similar to the mean F2 surface.

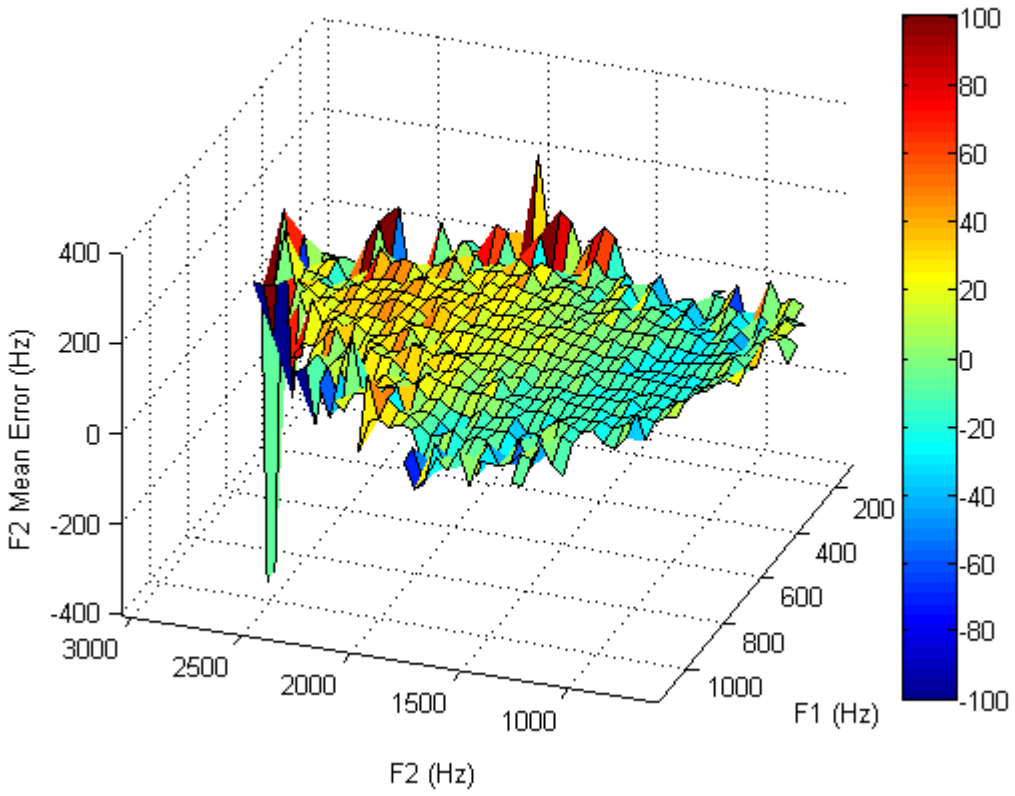

Figure 6.11 F2 mean error surface over the F1~F2 vowel space for all the VTR database vowel frames measured using Praat's normal tool for the benchmark LPC order variation case.

The F3 mean errors over the F1~F2 vowel space do not show the same dependency as the F3 mean errors with a fixed LPC order of 10 (Figure 6.9). Rather, for the benchmark case the surface is relatively flat. Plotting the F3 mean errors over the F2~F3 space reveals a dependency on the specified F3 values. However, like the F2 errors, the direction of the dependency has changed from the fixed LPC order condition so that the underestimated negative errors occur at the lower F3 values whilst the overestimated positive errors occur at the higher F3 values.

In general, for the numeric values and percentage representations, the distributions of mean error, absolute error and standard deviation for the three formants across the F1~F2 and F2~F3 vowel space for the benchmark case are much more stable and have lower values than those seen above for the constant LPC order cases. This is to be expected given that the measurements have the minimum possible errors. Whilst there is a dependency of the errors on the reference values, for F2 and F3 the direction of this dependency has switched between the fixed LPC order case and the benchmark case.

What is also of interest is the distribution over the vowel space of the LPC orders that have resulted in the minimum benchmark errors. The following three figures show the distribution of the median LPC order for F1 to F3 over the F1~F2 vowel space. In these plots the LPC order is represented only by colour, rather than as a surface, since it is a discrete variable. The same bin sizes that were used to calculate the error distributions over the vowel space have also been used for the calculation of the summary LPC order statistics.

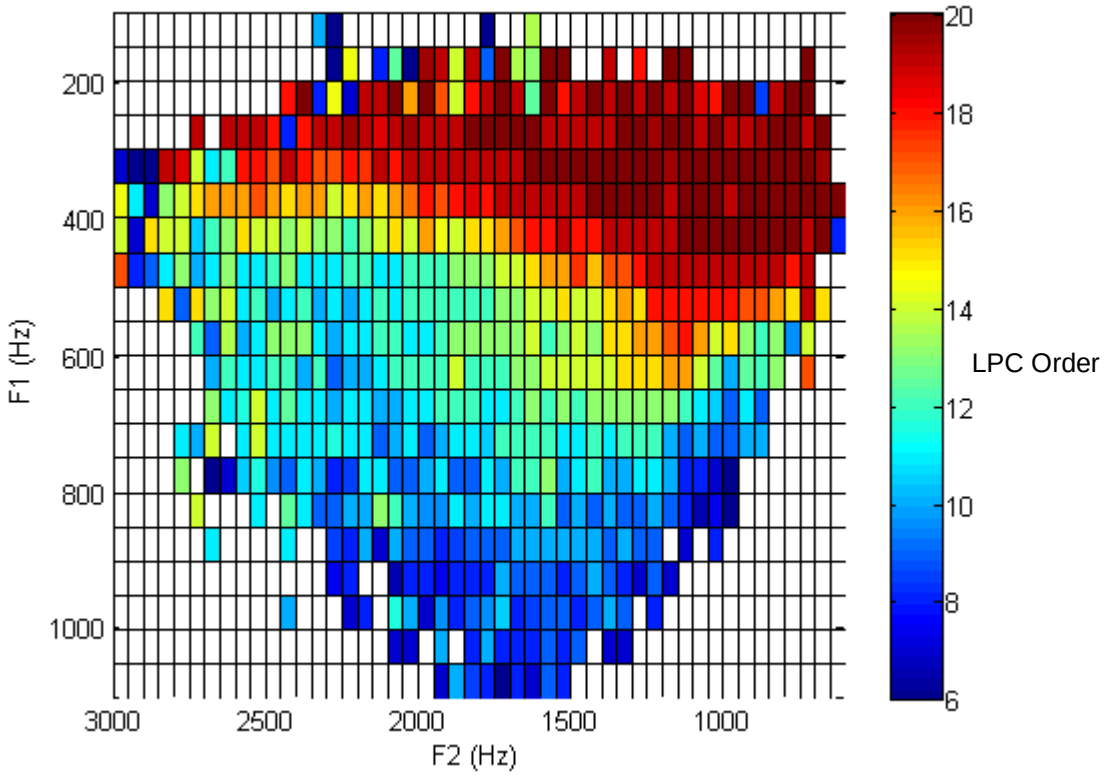

Figure 6.12 Median LPC order across the F1~F2 vowel space which produced the F1 errors in the LPC variation benchmark case.

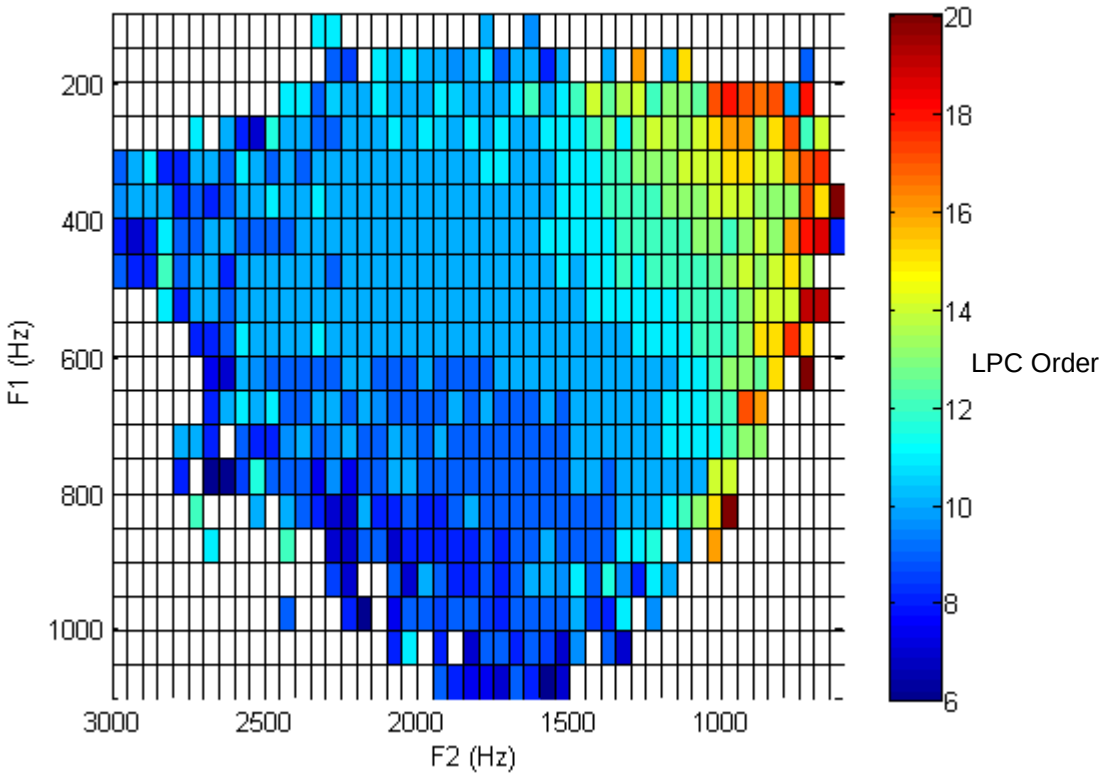

Figure 6.13 Median LPC order across the F1~F2 vowel space which produced the F2 errors in the LPC variation benchmark case.

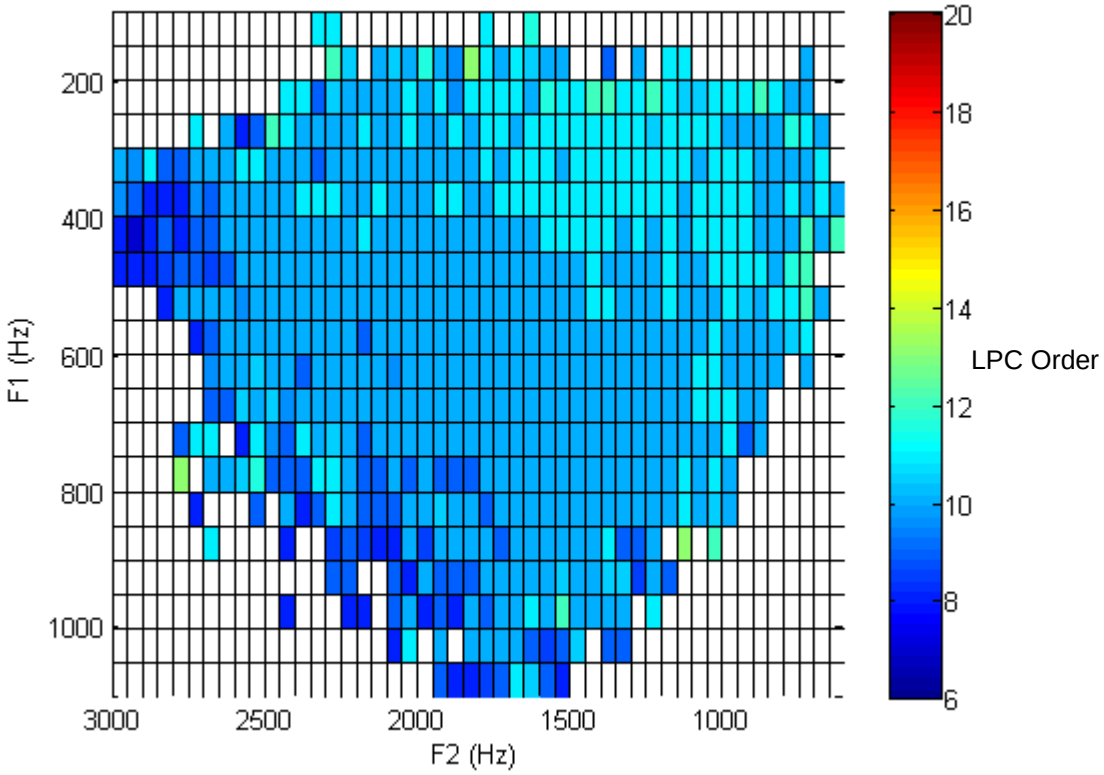

Figure 6.14 Median LPC order across the F1~F2 vowel space which produced the F3 errors in the LPC variation benchmark case.

Examination of Figure 6.12 to Figure 6.14 shows that the median of the LPC orders giving rise to the minimum benchmark errors are dependent on the location within the vowel space. In the case of the F1 benchmark errors (Figure 6.12) the dependency on F1 is the most pronounced. At the lower specified F1 values the LPC orders are the highest, around 19 and 20, whilst at the higher specified F1 values the LPC orders are the lowest, at 6 and 7. There is also a smaller dependency on F2 resulting in the highest LPC orders occurring for low F1 and low F2 values i.e. close back vowels. The general tendency for the F2 influence is the same as for F1 in that for the higher specified F2 values the LPC orders are lower.

A similar dependency on location within the vowel space in seen in Figure 6.13 for the LPC orders for the benchmark F2 errors. Again, the highest LPC orders are seen at the lower specified F2 values, while the lower LPC orders are seen at the higher F2 values. Also, a small influence is seen in the F1 direction. In comparison with the LPC order surface for the F1 errors, there is a relatively stable region in the surface where the LPC order remains between LPC orders 8 and 10. This accounts for approximately two thirds of the surface.

In contrast with the LPC order surfaces for F1 and F2 errors, the F3 surface shows much less variation across the F1~F2 vowel space. The mean LPC order values all lie relatively close to 10. However, there is a slight dependence on F1 and F2, again, in the same direction, with lower LPC orders at higher specified values. These data were also considered over the F2~F3 vowel space and a clear dependency on the specified F3 values was apparent. However, the overall range was less than that seen for F1 and F2 and generally occurred between LPC orders 8 to 12.

It is clear from the distribution of LPC orders over the vowel space that for the smallest errors to be generated when measuring formants, the LPC order must change. The greatest variation in LPC order is seen for F1, followed by F2 and then F3. In view of this finding it is then perhaps obvious that when the LPC order is fixed across all frames, as in Section 6.3.1.3, the overall performance is worse than the benchmark case where the LPC order is free to change.

A further finding highlighted by these results is that for a given region of the vowel space, the three formants show a tendency towards different LPC orders. This is most

marked between F1 and the other two formants, which tend to be relatively similar, apart from the region with the lowest F2 values. This has implications for the analysis frameworks where the same LPC order must be applied to each of the three formants within an analysis frame. The impact this has on the results in terms of the distribution of errors and LPC orders over the vowel space is considered in the following sections where the results from the other analysis frameworks are presented.

6.3.3.4 LPC Order Fixed within Tokens, Variable Across Formants

The analysis framework that produced results closest to the benchmark condition was the situation where the LPC order was fixed across the frames within a token for a given formant, but the LPC order could be different across the three formants. Consideration of the error results across the vowel space in both numeric and percentage terms and with both minimum error criteria (minimum absolute error and minimum absolute percentage error) reveals results very similar in structure to the benchmark condition. For the F1 errors, the surfaces only show overestimates at the lower F1 values, with no significant regions of underestimates at the higher F1 values. The surfaces for both F2 and F3 errors in numeric terms show no apparent dependency across the vowel space. However, when the F2 absolute errors are considered as percentages there is dependency on the F1~F2 values with the larger errors occurring in the lower F1~F2 region (close back vowels) and the smallest errors in the higher F1~F2 region (open front vowels).

The distributions of the LPC orders that produced these results are very similar to those for the benchmark condition. The only obvious difference is that the range of LPC orders used for F2 is somewhat reduced in the low F1~F2 region (close back vowels).

6.3.3.5 LPC Order Fixed Across Formants, Variable Across Frames

The next analysis framework considered was where the LPC order was fixed across the three formants and was free to vary from frame to frame. This produced overall performance results that were the next closest to the benchmark set, after the framework in the previous section. Examination of the distribution of the errors over the vowel space for the current framework reveals a set similar to those already described but with the following points of note. The errors for F1 are more similar to the benchmark results that the fixed LPC order case, but show slightly more underestimates at the higher F1 values than the benchmark case. The errors for F2 show the same tendency as for the

fixed LPC order 10 results, i.e. the overestimates occur at lower F1 and F2 values with the underestimates at the higher F1~F2 values, but the dependency is less marked. The F3 results show a dependency not previously encountered, where the negative errors occur at the lower F1~F2 values (close back vowels) and the positive errors occur at the higher F1~F2 values (open front vowels). The variation across the vowel space is less marked than for the other formants, but when the minimum error criterion is the summed percentage error then a small region of negative errors in the lower F1~F2 area becomes quite apparent. This is also clear on the F2~F3 vowel space plots. These also reveal that positive F3 errors occur at the higher F3 values.

The current framework requires the same LPC order across all three formants for each frame so there is only a single distribution of LPC orders for all three formants. The distribution is almost identical to the distribution of LPC orders for F3 in the benchmark condition shown in Figure 6.14. There is slight variation across the surface with the median LPC orders around 11 in the low F1~F2 region (close back vowels) dropping over the surface to 9 in the high F1~F2 region (open front vowels). Considering the LPC orders in relation to the F2~F3 vowel space produces a distribution almost identical to that discussed above for F3 in the benchmark condition. Again, the higher median LPC orders occur at the lower F3 values. When the minimum error criterion is the mean absolute percentage error, the LPC orders in the lower F1~F2 region are slightly higher than for the minimum absolute error criterion. The same is also true for the distribution over the F2~F3 vowel space at the lower F3 values.

6.3.3.6 LPC Order Fixed within Tokens and Across Formants

The final framework examined was a combination of the previous two, so the LPC order must remain the same across all three formants and within each vowel token, but is free to change from one token to the next. The distribution of errors across the vowel space from this framework are very similar to those from the previous framework where the LPC order was fixed across the three formants but free to change from frame to frame. Also, the distribution of LPC orders leading to these results is again very similar to those from the previous framework. This suggests that the constraint across the three formants has a greater influence on the LPC order and resulting formant measurements than the constraint across the frames.

6.3.3.7 Summary of Errors & LPC Orders Over the Vowel Space

The results presented in the sections above make it clear that formant measurement errors are dependent on the location of the vowel within the vowel space. Not only is the magnitude of the errors affected by the vowel's location, but also the direction of the error, i.e. whether they are over or underestimates of the true value. Also, the nature of the dependence over the vowel space is different across the three formants examined. Furthermore, the dependencies can change according to the analysis framework adopted. The greater the constraints on the measurement process, the larger the errors and the greater the variation in the errors across the vowel space.

For analysis frameworks where the LPC order can vary, patterns also emerge over the vowel space for the LPC orders used. Where a different LPC order can be adopted for each formant then the higher LPC orders tend to occur at the lower formant values and the lower LPC orders at the higher formant values for a given formant. The greatest variation in LPC orders is seen for F1. For the frameworks where the LPC order must be the same across the three formants, the variation in the use of LPC orders is dramatically reduced.

6.3.4 Variation of Performance Across Speakers

Previous studies have found that the performance of formant analysis techniques and the behaviour of resulting errors are to some extent dependent on the speakers (Künzel (2001), Byrne and Foulkes (2004), Vallabha and Tuller (2004) and Duckworth et al. (2011)). The large number of speakers in the VTR database makes this an ideal data set within which to further explore the variation in performance across speakers and address RQ3. However, a limitation is that for most of the speakers, 162 out of 186, there are only 2 sentences of speech. This final section begins with a description of the speakers within the dataset, followed by an analysis of the results already presented in this chapter with the speaker considered as a factor.

6.3.4.1 Description of the Speakers

The VTR database contains reference formant measurements for 186 different speakers, 113 male and 73 female. For 24 of the speakers (16 male and 8 female) there are 8 sentences each and for the remaining 162 speakers (97 male and 65 female) there are 2 sentences each. The average number of vocalic frames for the 8 sentence speakers is

997 across an average of 98 tokens, whilst for the 2 sentence speakers there are an average of 267 frames across an average of 26 tokens.

6.3.4.2 Speakers' Reference Formant Values

The distribution of reference formant values within the entirety of the dataset has already been described in Section 6.3.3.1. In order to examine the distribution of formant values for individual speakers, the mean values were calculated for the first three formants for each speaker. Figure 6.15 shows the mean reference F1 values plotted against the mean F2 values for each speaker. The axes have been oriented to align with the representation of the F1~F2 vowel space used in the rest of the thesis. To differentiate between the two sexes, the values for the male speakers are represented with a blue circle, and the female speakers are shown as red circles.

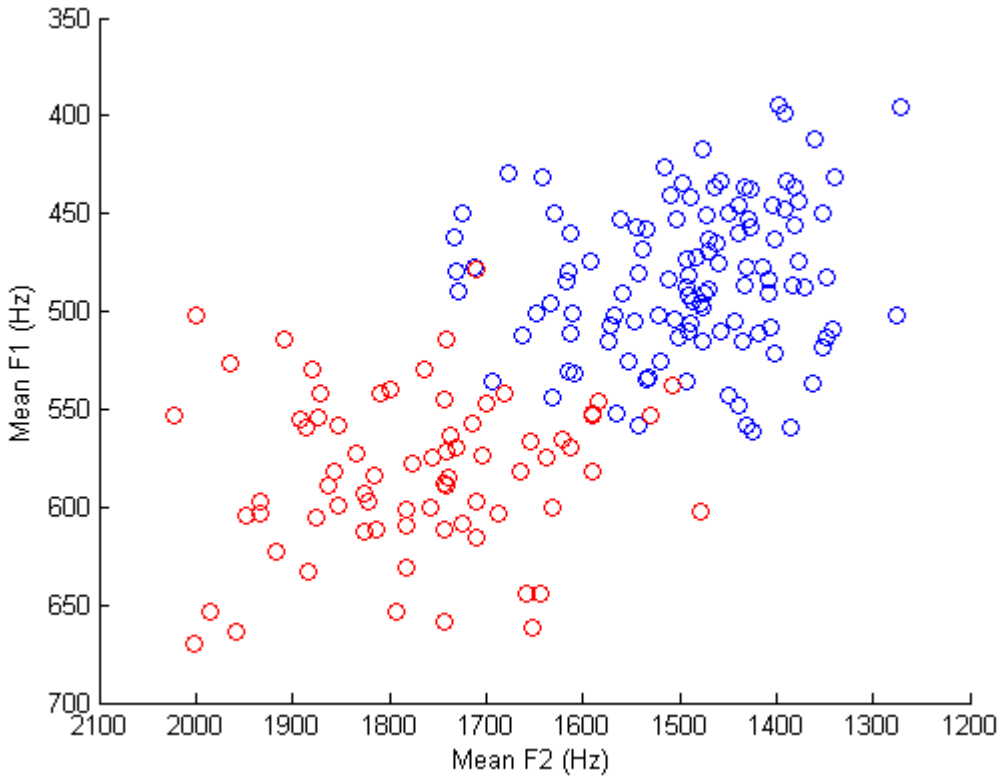

Figure 6.15 Plot of the mean F1 reference value against the mean F2 reference value for each of the 186 speakers in the VTR database. Male speakers are shown as blue circles, female speakers are red circles.

The plot shows the expected tendency for the mean F2 values to increase as the mean F1 values increase. The plot also shows that there is very limited overlap between the mean values for male and female speakers. However, there is a large degree of overlap between the sexes in terms of the individual vowel tokens. For F3, the mean values

ranged from 1,998 Hz to 2,703 Hz for the men and 2,369 Hz to 3,112 Hz for the women.

To provide a further indication of the variability that exists within the reference formants across the speakers, the standard deviation values for all speakers were calculated. In general, male speakers exhibited lower mean formant values than the females, and showed smaller variability. The variability of the F3 values was smaller than for the F2 values.

It should be noted that the mean values do not necessarily represent the mid-point of a speaker's normal vowel space or a measure that is directly comparable across speakers, since the distribution of vowel tokens for each individual was not controlled for in the original TIMIT corpus or in the selection of speakers for the VTR database.

6.3.4.3 Speakers' Fundamental Frequency

As well as differences in the region and range of the vowel space used by individual speakers, differences are also found in their fundamental frequencies. The mean and standard deviation of fundamental frequency were calculated for each speaker within Praat using the autocorrelation method. The values were calculated across all the speech material for each speaker. Figure 6.16 shows the distribution of the measured mean fundamental frequency for each speaker in the form of a histogram with a bin width of 10 Hz. The results have been separated for the male and female speakers, with the blue bars showing the results for the male speakers and the red bars for the female speakers. The range and overall mean for each sex are what one would expect for normal male and female speakers of American English (Fitch and Holbrook 1970, Baken and Orlikoff 2000, p. 175-176 Table 6-2). There is some overlap in the distributions between 150 and 180 Hz.

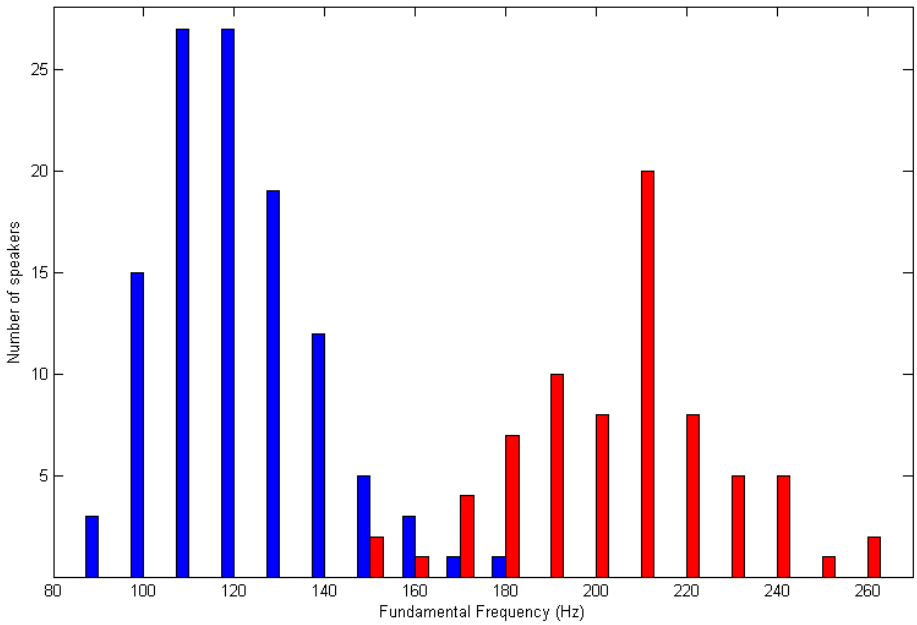

Figure 6.16 Histogram showing the distribution of speakers' mean fundamental frequency from the sentences used in the VTR database. Male speakers are blue and female speakers are red.

6.3.4.4 Analysis of Speakers' Results

The following sections consider the results of the analysis framework already described in this chapter in terms of the performance of individual speakers. Section 6.3.4.5 summarises the mean absolute errors for the speakers across the frameworks already used in this chapter. This is followed by Sections 6.3.4.6 to 6.3.4.11, which consider how these results vary for individual speakers both across and within the frameworks, and they examine the relationships between the errors and speaker properties, such as fundamental frequency and location within the vowel space. Finally, Sections 6.3.4.12 to 6.3.4.18 examine the behaviour of the LPC orders for speakers across the analysis frameworks.

6.3.4.5 Analysis of Mean Speaker Errors Across Frameworks

This first section summarises the performance or the magnitude of the errors for speakers across the different analysis frameworks. The first stage was to determine the mean absolute and mean absolute percentage errors for all three formants, across all the frames for each speaker. The mean, standard deviation, minimum and maximum values of the speakers' mean absolute error were then calculated. These results are shown in Figure 6.17 for the mean absolute errors and Figure 6.18 for the mean absolute percentage errors. The plots are similar in form to Figure 6.4 and Figure 6.5, which are

used to summarise the overall performance across each of the analysis frameworks. Again, the same ordering of frameworks has been used and the circles represent the mean of the speaker means for each formant with the vertical bars extending one standard deviation above the mean. The minimum and maximum speaker mean values have also been included as upward and downward pointing triangles respectively, to show the range of speaker means encountered.

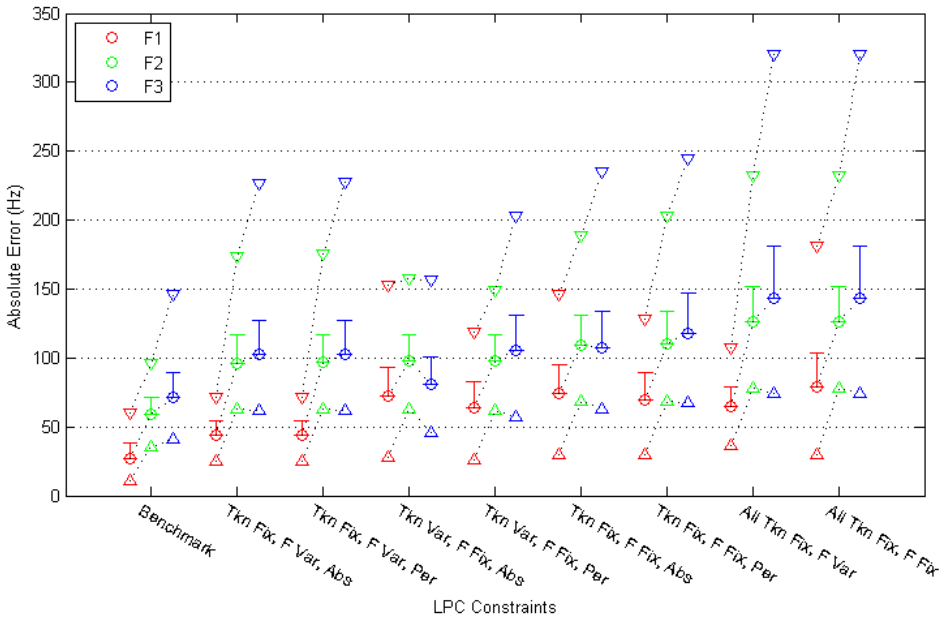

Figure 6.17 Mean (circles), standard deviation (bar = 1 SD), minimum (upward triangles) and maximum (downward triangles) of individual speakers' absolute mean error across analysis frameworks for F1 (red), F2 (green) and F3 (blue). (Key to conditions: Tkn = Token, F = Frame, Fix = Fixed, Var = Variable, Abs = Absolute Error Criterion, Per = Percentage Error Criterion).

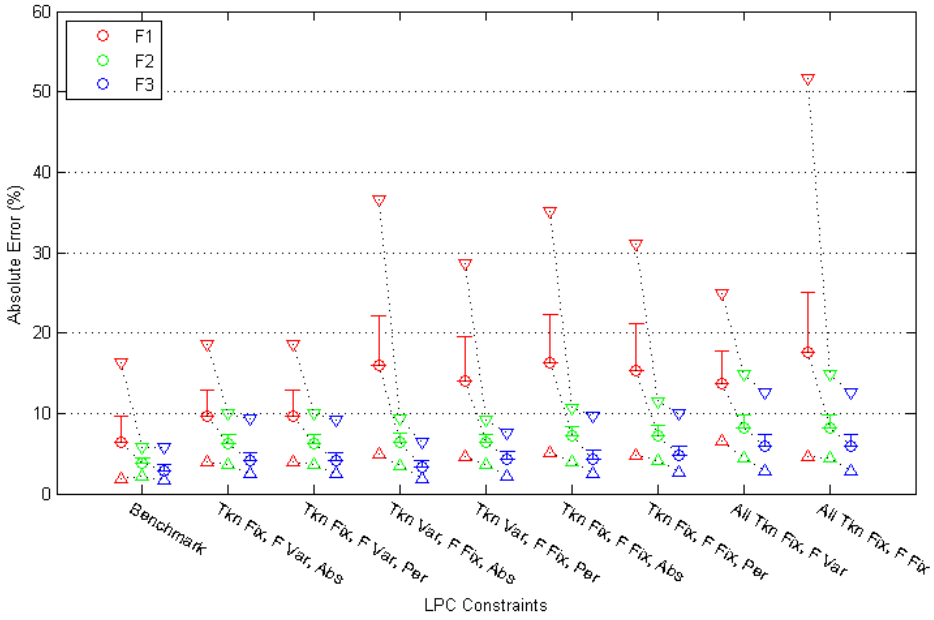

Figure 6.18 Mean (circles), standard deviation (bar = 1 SD), minimum (upward triangles) and maximum (downward triangles) of individual speakers' percentage absolute mean error across analysis frameworks for F1 (red), F2 (green) and F3 (blue). (Key to conditions: Tkn = Token, F = Frame, Fix = Fixed, Var = Variable, Abs = Absolute Error Criterion, Per = Percentage Error Criterion).

As expected, the means of the speaker means (both in numeric and percentage terms) are very close to the overall means from each of the frameworks. However, the standard deviations of the speaker means are much less than the overall standard deviations of the errors. The differences across the numeric and percentage representations are again present, namely that for the numeric values the F1 means and standard deviations are less than those for F2 and F3 whilst the reverse is true for the percentage error results. Across the frameworks, as the LPC order constraints become more restrictive, the magnitude of the mean speaker errors increases.

6.3.4.6 Analysis of Mean Errors Within & Across Frameworks

The previous section provides an overall summary of the individual speaker means and shows how they vary across the analysis frameworks, but it does not consider the behaviour of individual speakers. Figure 6.19 shows the mean absolute F1, F2 and F3 errors in numeric terms for each speaker in the framework where the LPC order is fixed both within individual tokens and across formants, with the absolute minimum error criterion (sixth framework from the left in Figure 6.17 and Figure 6.18). The red, green and blue vertical bars represent the F1, F2 and F3 mean absolute error respectively for

each speaker. The speakers have been ordered according to increasing combined error, i.e. the sum of the mean error from F1, F2 and F3.

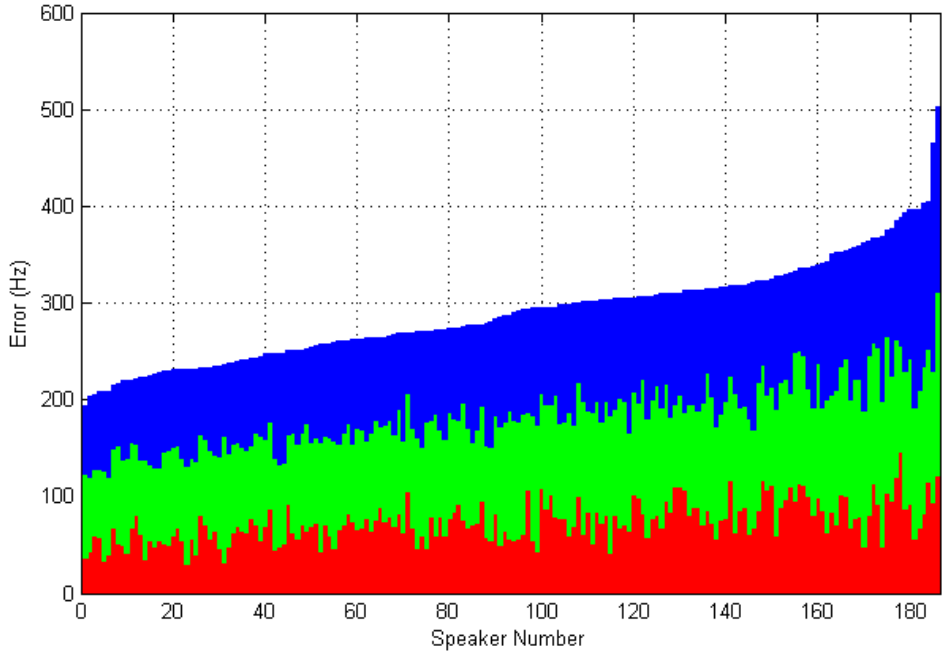

Figure 6.19 Mean absolute errors for F1 (red), F2 (green) and F3 (blue) for 186 speakers in the VTR database from the analysis framework where the LPC order is fixed both within individual tokens and across formants, with the absolute minimum error criterion.

The plot shows the range of combined errors extends from just less than 200 Hz for the speaker on the far left to just over 500 Hz for the speaker on the far right. In general, the errors for F1, represented by the red bars, are less than those for F2 and F3, represented by the green and blue bars, which overall are roughly equal. One obvious feature of the plot is that the errors for the individual formants do not appear to increase proportionally as the combined error increases. This feature is examined in more detail below.

Figure 6.19 only concerns the results from one framework. Examination of the same type of plot for the results from the other frameworks reveals the same lack of proportionality between the individual mean formant errors as the combined error increases. The only difference of note across the frameworks is the range and magnitude of the errors, which are represented in Figure 6.17 and Figure 6.18.

Comparison of the plots with those generated from the mean absolute percentage errors shows the same overall structure with a non-proportional increase in the errors for the

three formants. The only significant difference across the two sets of plots is that the F1 errors are larger than the errors from F2 and F3 in the percentage error plots, whilst the opposite is true in the numeric error plots. This is to be expected given the differences seen in the summary plots at Figure 6.17 and Figure 6.18.

6.3.4.7 Relationship Across Formants

The apparent lack of proportionality between the errors from the three formants for the individual speakers suggests that there is not an obvious relationship between them. In order to comment further on this, scatter plots of the mean speaker errors for F1 against F2, F1 against F3 and F2 against F3 for all the frameworks were generated. Examples are shown in Figure 6.20 to Figure 6.22 for the mean values presented in Figure 6.19.

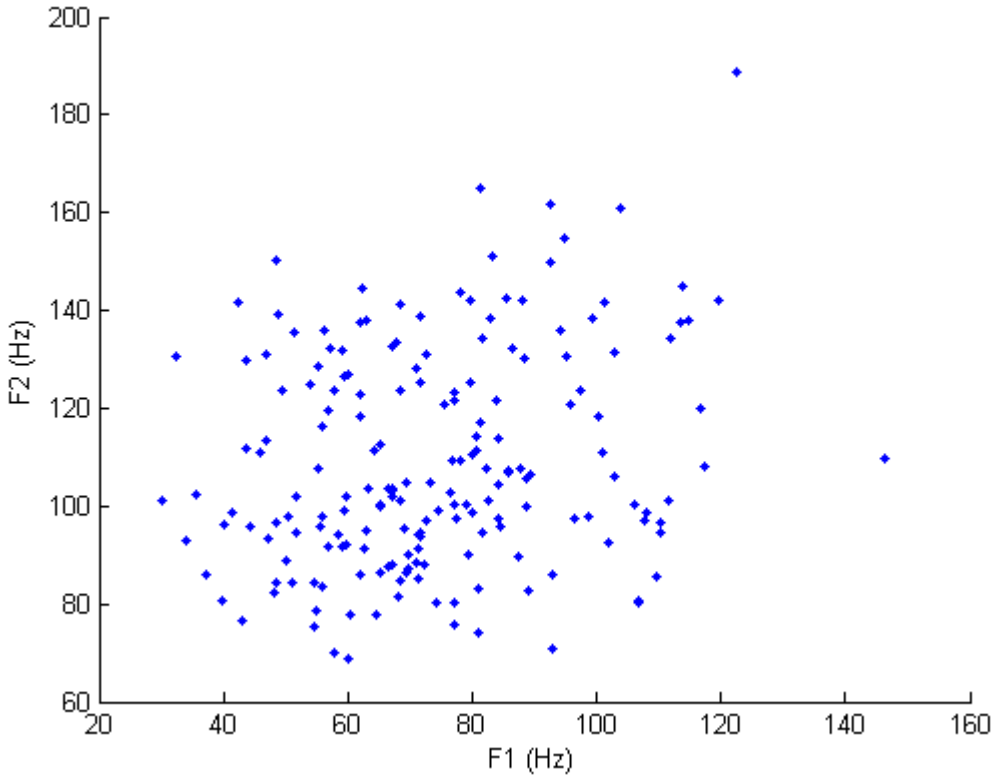

Figure 6.20 Scatter plot of mean absolute F1 error vs mean absolute F2 error for 186 VTR database speakers from the analysis framework where LPC order is fixed within the token and across formants, with the absolute error criterion.

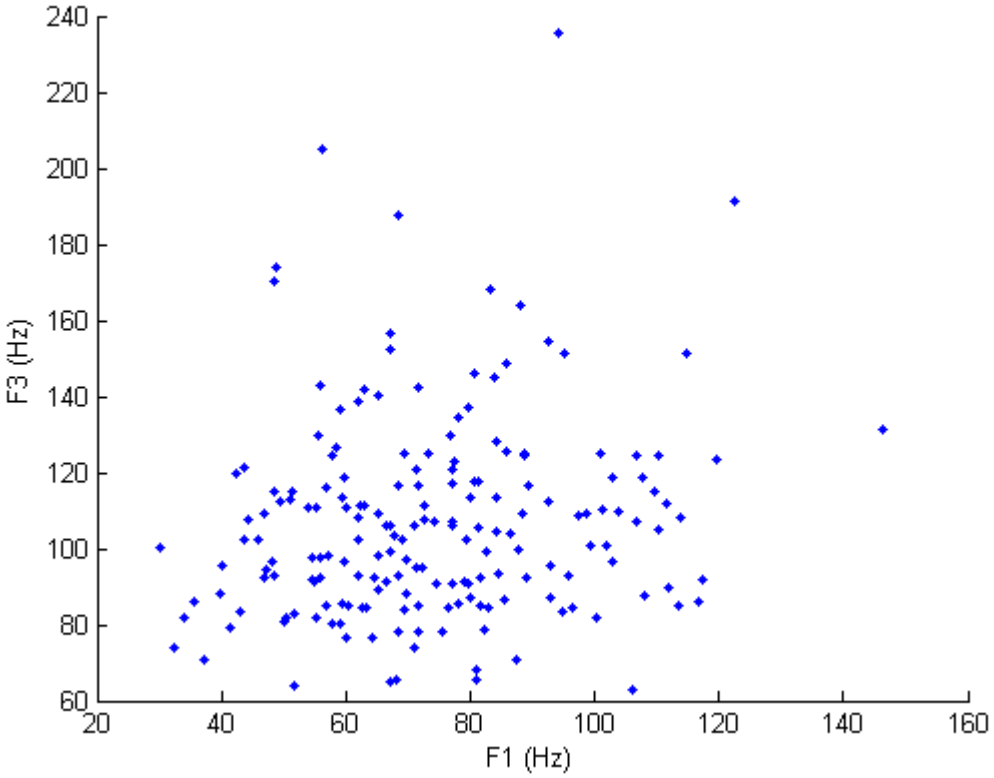

Figure 6.21 Scatter plot of mean absolute F1 error vs mean absolute F3 error for 186 VTR database speakers from the analysis framework where LPC order is fixed within the token and across formants, with the absolute error criterion.

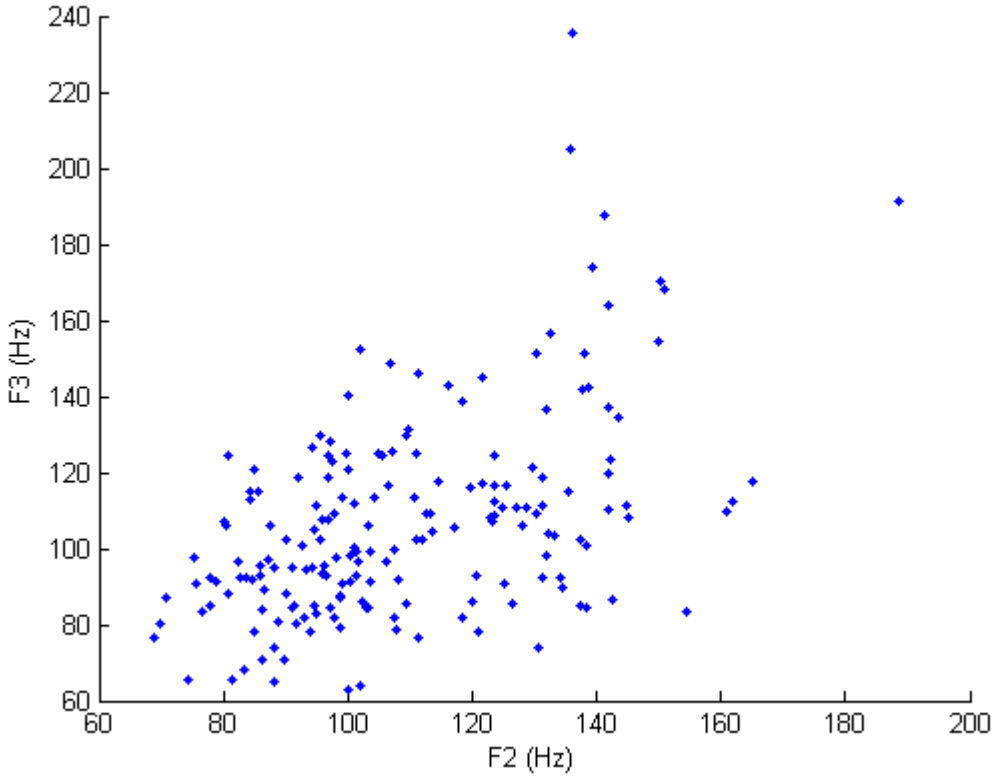

Figure 6.22 Scatter plot of mean absolute F2 error vs mean absolute F3 error for 186 VTR database speakers from the analysis framework where LPC order is fixed within the token and across formants, with the absolute error criterion.

The scatter plots of the speaker means for F1 vs F2 and F1 vs F3 show no apparent trends in the data. However, the scatter plot for F2 vs F3 does show a positive linear tendency with the data points lying in a general diagonal orientation from the bottom left to the top right of the plot. However, they are not tightly grouped. Similar patterning is seen across the scatter plots for the other analysis frameworks for both the numeric and percentage representations of the speakers' mean errors. Overall, these results suggest that there is no apparent relationship between the mean absolute errors of F1 and F2, and between F1 and F3, but that a slight dependence exists between F2 and F3.

To allow a numeric assessment and comparison of the cross formant relationships, Pearson's correlation coefficients were calculated for the three formant comparisons across the frameworks. The correlation coefficients support the observation that the relationship between F2 and F3 is stronger than that for both F1 and F2, and F1 and F3. Whilst the majority of the coefficients were significant, they nevertheless showed that the relationships between the formants are reasonably weak. This is apparent from the degree of dispersion in the scatter plots.

6.3.4.8 Relationship Across Frameworks

It is clear from the results presented in the sections above that for all of the analysis frameworks the performance of individual speakers occurs over a range. What is not apparent from those results is whether the speakers who achieve a high performance for one analysis framework also do so for the other frameworks, or whether speakers that perform well in one framework perform poorly in others. The results presented below address this issue.

The first approach taken was to generate a series of scatter plots, examining the relationships between the speaker means for each formant across the different analysis frameworks. Again, this was done for both the numeric and percentage representations of the mean absolute error values for each speaker. As well as considering the individual formants, plots were also generated for the combined means across the three formants. Two example scatter plots are shown in Figure 6.23 and Figure 6.24. The first shows F1 errors for the framework with LPC order fixed across individual tokens and formants with the absolute minimum error criterion plotted against F1 errors from the benchmark condition. The second shows F3 errors for the same framework in the first plot against the F3 errors from the framework with LPC order fixed across individual tokens but variable across formants with the absolute minimum error criterion.

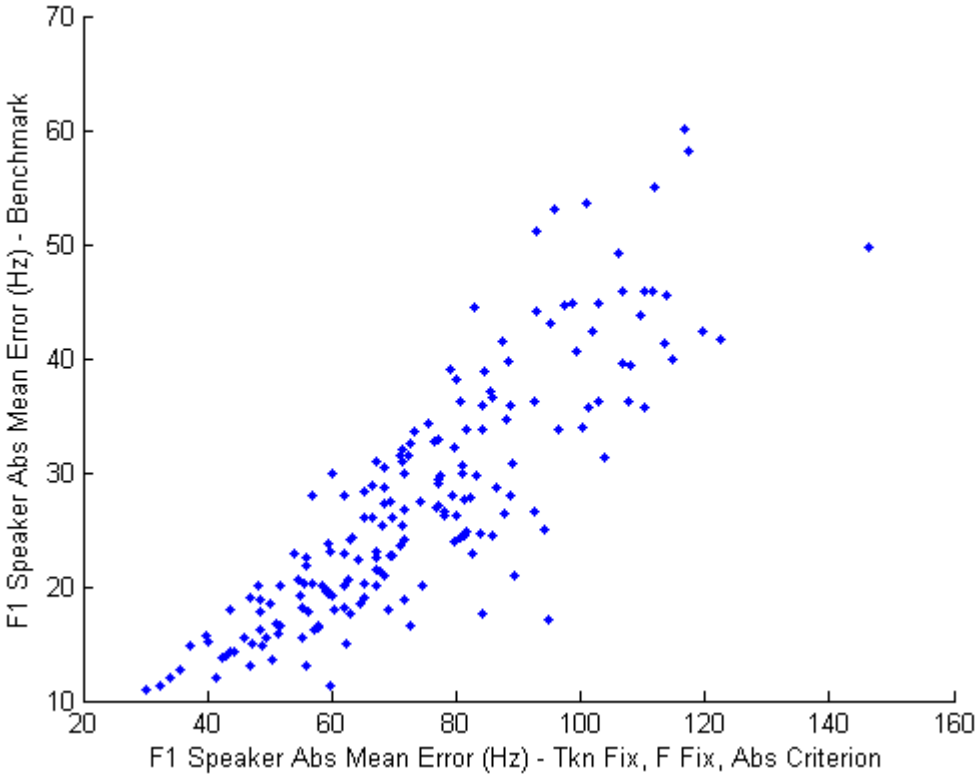

Figure 6.23 Scatter plot of mean absolute F1 error from analysis framework where LPC order is fixed within the token and across formants, with the absolute error criterion vs mean absolute F1 error from the benchmark framework for 186 VTR database speakers.

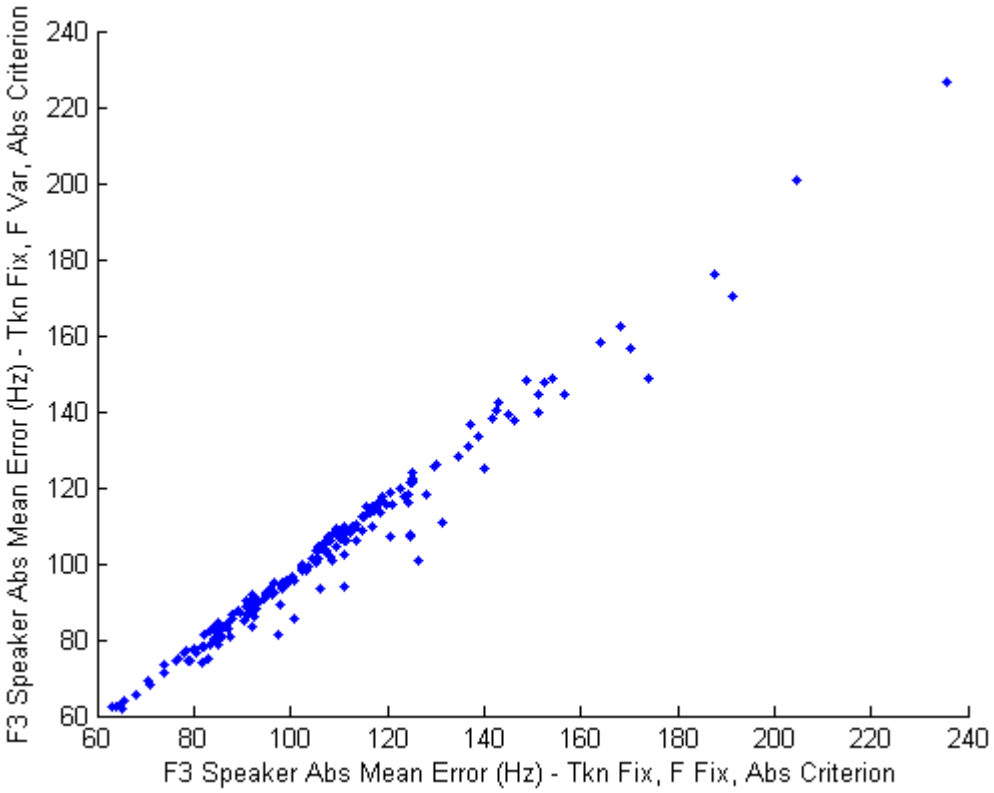

Figure 6.24 Scatter plot of mean absolute F3 error from analysis framework where LPC order is fixed within the token and across formants, with the absolute error criterion vs mean absolute F3 error from the analysis framework where LPC order is fixed within the token and variable across formants, with the absolute error criterion for 186 VTR database speakers.

Figure 6.23 shows a relatively strong correlation between the two sets of speakers' mean absolute errors, whilst Figure 6.24 shows an even stronger correlation. Pearson's r correlation coefficients for the data in the two plots are 0.8542 and 0.9877 respectively. Correlations of the strength seen in these two plots were found across the entire set of plots generated. Overall they show much tighter groupings that those seen above for the comparison of speakers' mean absolute errors across formants. Correlation coefficients were also calculated for all combinations and they were all significant at the .01 level (two tailed). The correlation coefficients for the percentage means tended to be even higher.

These results show that for the errors from both the individual formants and all three formants combined, the change in performance across frameworks is relatively consistent within the group of speakers. Even though the magnitude of the errors is different across frameworks, speakers that perform well in one, relative to the rest of the group, perform well in the others, and those that perform badly, relative to the other speakers, do so across all the frameworks. This suggests that there is some feature or

features of the speaker that influences or determines the level of performance achieved. A number of the speakers' attributes were compared with the error results to determine if they were related to the performance of the speakers. The results from these comparisons are presented below.

6.3.4.9 Performance of Male & Female Speakers

The first factor considered was the sex of the speakers, and whether it was related to performance. In terms of the mean absolute errors presented in Table 6.1 and Table 6.3, the errors represented numerically were greater for female speakers than male speakers, whilst the sexes were reversed for the percentage errors as a consequence of female speakers having, on average, higher formant values. In order to examine the relationship between speaker sex and performance, the speakers were ranked according to the mean combined error across the three formants for each of the nine frameworks. The speaker with rank number 1 had the smallest mean combined error and the speaker with rank number 186 had the largest mean combined error. The average rank position for each speaker across the frameworks was then calculated. This was done for both the numeric and percentage representations of the mean. Since the results in the previous section showed that within the group each speaker had a similar relative performance across the frameworks, the approach of averaging the rank positions across the frameworks was justified.

To determine the distribution of the two sexes within the ranked speakers, histograms were generated for both sets of rankings from the numeric and percentage errors. These are shown in Figure 6.25 for the rankings derived from the absolute numeric errors, and in Figure 6.26 for the rankings from the percentage errors across the analysis frameworks. In both plots, the number of male speakers is represented by the blue bars, and the female speakers are represented by the red bars. The bin width for the histograms is 12. When interpreting these plots it should be remembered that out of the 186 speakers only 73 are female and 113 are male.

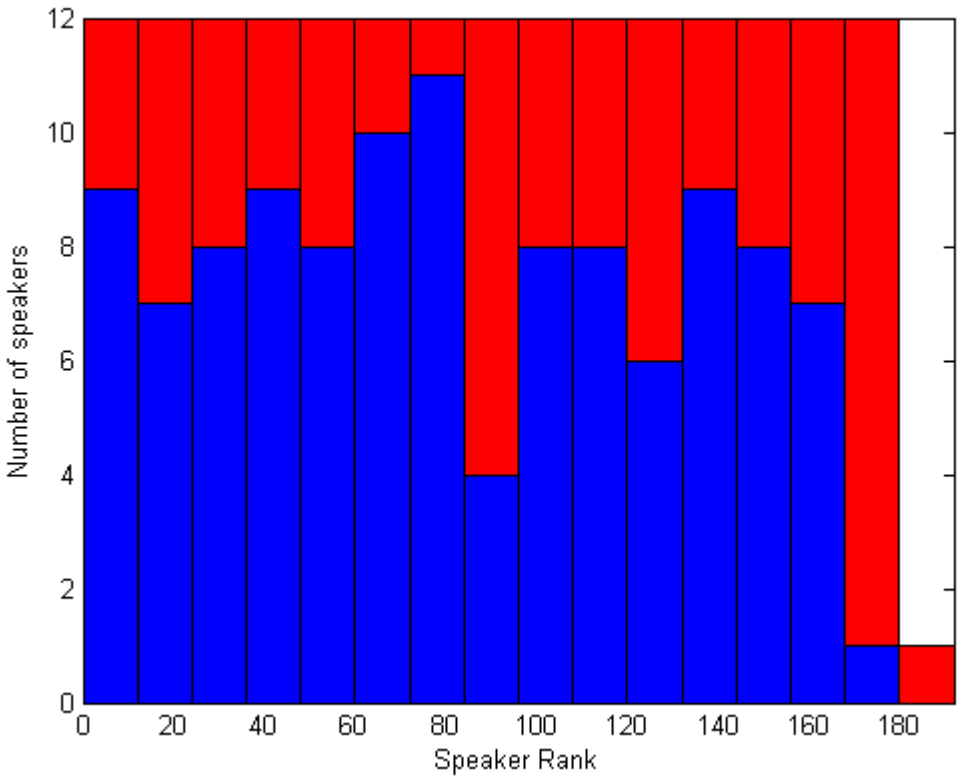

Figure 6.25 Distribution of speaker sex (male = blue, female = red) by mean rank position based on mean combined errors across frameworks, grouped in 12 speaker blocks.

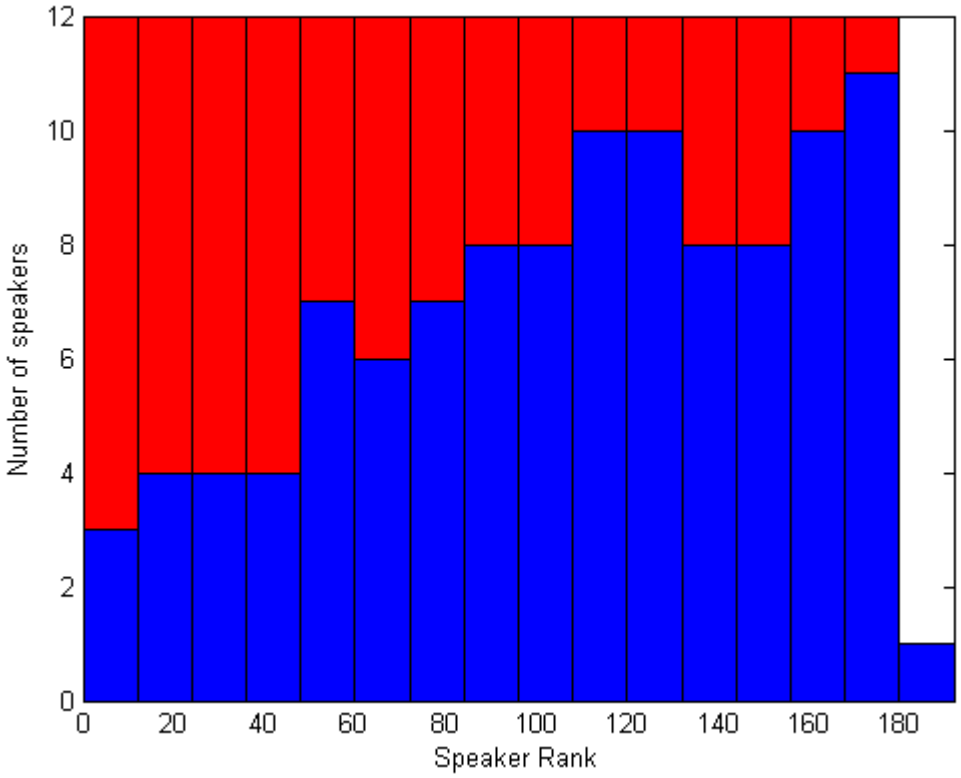

Figure 6.26 Distribution of speaker sex (male = blue, female = red) by mean rank position based on mean combined percentage errors across frameworks, grouped in 12 speaker blocks.

In Figure 6.25 the distribution of male and female speakers according to their ranking from the numerically expressed errors does not show any particular patterning in comparison to Figure 6.26 where the number of female speakers per interval decreases in a relatively systematic way as the number of male speakers increases across the ranks. In the case of the numeric errors the number of female speakers per interval varies across the intervals but shows no particular patterning across the range, apart from the final complete interval where all the speakers, except for one, are female. However, the speaker with the lowest average rank position, i.e. the best performing speaker, was a woman, speaker 'fjen0'. Of the female speakers in the first interval, they occupied positions 1, 3 and 7. At the other end of the performance range female speakers also occupied the worst 16 positions. The histogram derived from the percentage results shows that expressing performance in this way does reveal an overall tendency between the performance of the speakers and the sex of the speakers. However, there is still significant overlap between the sexes.

6.3.4.10 Fundamental Frequency

The next factor considered that might be related to performance was fundamental frequency. The mean fundamental frequency values for the speakers presented in Section 6.3.4.3 fall into two relatively distinct sex groups, with minimal overlap. The results in the section above show that both sexes display a wide range of performance in terms of mean absolute error which are almost in complete overlap. Given the two different groupings of the sexes across fundamental frequency and performance, it seems unlikely that any strong relationship will exist between a speaker's fundamental frequency and their performance. To confirm whether or not this was true, two scatter plots were generated for the fundamental frequency against mean rank position across frameworks, derived from numeric mean formant errors, and against mean rank position derived from the percentage errors. The first of these two scatter plots is shown in Figure 6.27. Again, male and female speakers have been distinguished by the colour of the data points.

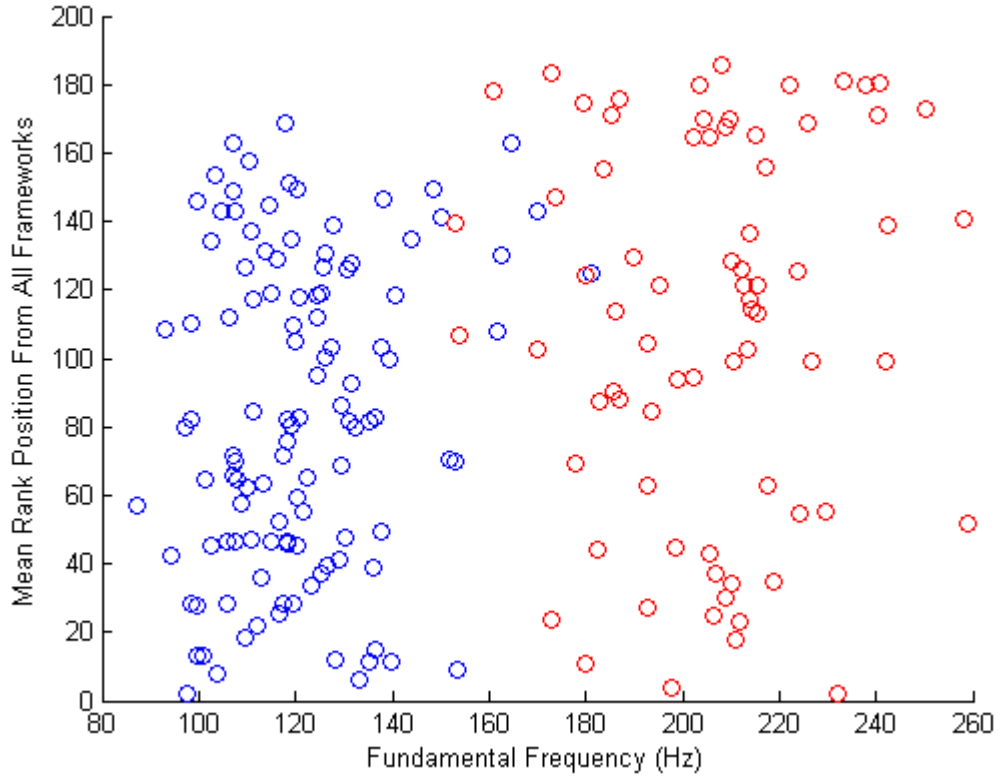

Figure 6.27 Scatter plot of speakers' mean fundamental frequency against mean rank position across frameworks derived from numeric errors. Male speaker are blue, female speaker as red.

Figure 6.27 shows the separation between the male and female speakers in terms of their mean fundamental frequencies, with the female speakers occupying the right hand

side of the plot and the male speakers on the left. However, as observed above, there is no such separation in terms of the performance. Overall, this leads to a dispersed set of data points. Had a strong relationship existed between fundamental frequency and performance a tighter grouping of the data points would be expected. Even within the sexes there is no apparent dependency no fundamental frequency. The scatter plot derived from the percentage data had a very similar overall structure to Figure 6.27 and revealed no apparent relationship between the measures.

Pearson's correlation coefficients were also calculated for the mean rank positions derived from both the numeric and percentage against fundamental frequency for the group as a whole and for the two sexes. The coefficients are shown in Table 6.16. Those correlation coefficients that are significant at the 0.01 level (two tailed) are marked with a double asterisk, whilst those that are significant at the 0.05 level (two tailed) are marked with a single asterisk.

Performance	**All n = 186**	**Male n = 113**	**Female n = 73**
Numeric	0.2718**	0.1786*	0.0616
Percentage	-0.3282**	0.0812	0.0214

Table 6.16 Pearson's correlation coefficients for comparison of mean speaker rank across frameworks determined from performance expressed numerically and in percentage terms with mean speaker fundamental frequency. ** = significant at 0.01 level (two tailed), * = significant at 0.05 level (two tailed).

The weak correlation coefficients confirm the absence of a strong linear relationship between fundamental frequency and performance. The difference in direction of the correlation between the mean rank positions derived from numeric and percentage error values aligns with the findings from the previous section whereby the distribution of the sexes was different across the two performance measures. The differences in magnitude of the correlation coefficients between the individual sexes and the entire set of speakers suggests that the correlation seen for the group is a consequence of combining the two sexes rather than it being an extension of a relationship that exists within the individual sexes.

A number of scatter plots were also generated to compare the mean absolute errors for individual speakers for different frameworks with their mean fundamental frequency values. They were created to check that the process of determining the mean rank position across frameworks had not weakened or obscured any relationships that may have been apparent for individual formants for specific frameworks. The plots revealed

very similar results to those seen for the mean rank based plots, which could be considered as a validation of the selection of the mean ranks as a representative performance measure.

6.3.4.11 Location Within Vowel Space

Another factor that was considered to be potentially related to the performance of speakers was their location within the vowel space. The results in section 6.3.3 show that the localised mean errors from the entire data set exhibit various trends and tendencies over the vowel space. To discover if any relationships existed between the performance of speakers and their location within the vowel space the mean rank positions of the speakers were compared with their mean reference formant values. The mean reference formant values are presented in Section 6.3.4.2 and show that there are differences across the sexes. Again, as with fundamental frequency, the values tend to fall into one of two groups according to the speaker's sex. However, the amount of overlap is greater for the formants than for fundamental frequency. Given this similar behaviour it is again expected that there will be no strong relationship between the performance, expressed as mean rank position, and the location of the speaker within the vowel space.

Scatter plots were produced to show the mean rank position against mean reference formant values and the results from each sex were colour coded. The plots again showed the clear grouping of data points from the two sexes across the reference formant values, with the female speakers generally having higher values than the male speakers. However, in terms of the performance, the plots showed no strong relationships between the two parameters with a large degree of dispersion in the data points similar to that seen for the fundamental frequency plots above. In order to assess whether any underlying tendencies were present the Pearson correlation coefficients were calculated. These are shown in Table 6.17. Correlation coefficients are marked with a double asterisk if they are significant at the 0.01 level (two tailed) and those that are significant at the 0.05 level (two tailed) are marked with a single asterisk.

Performance	Formant	All n = 186	Male n = 113	Female n = 73
Numeric	**F1**	0.0704	-0.2409*	-0.1543
	F2	0.2853**	0.1377	0.1711
	F3	0.3008**	0.2374*	0.1123
Percentage	**F1**	-0.5794**	-0.5724**	-0.3494**
	F2	-0.3497**	-0.1355*	-0.0172
	F3	-0.3478**	-0.0204	-0.1211

Table 6.17 Pearson's correlation coefficients between mean rank position determined both numerically and in percentages terms, and mean reference formant values for each formant. ** = significant at 0.01 level (two tailed), * = significant at 0.05 level (two tailed).

On the whole, the correlation coefficients show a weak relationship between the performance of the speakers, expressed as mean rank position, and mean reference formant values. For the mean rank positions derived from the numeric errors there is a slight tendency for the lower performers to have higher reference formants, at least for F2 and F3. For the rank positions derived from the percentage errors the tendency is reversed and the effect is strongest for F1. Even though some of the results for the sexes are significant, the coefficients only indicate a weak relationship between the parameters.

6.3.4.12 LPC Order Variation Across Speakers

This final section of the analysis considers the usage of LPC orders by the speakers across the analysis frameworks and whether this is related to factors previously examined, such as the speaker's sex, mean fundamental frequency and location within the vowel space. Such information could prove useful in helping to determine suitable LPC orders for speaker based on these attributes. The results presented in Sections 6.3.3.2 to 6.3.3.6 for the entire set of results shows that there is a range of different behaviours for the LPC orders for different analysis frameworks and across the three formants. The analysis of the behaviour across the speakers begins by considering the LPC orders used by speakers within the different analysis frameworks.

6.3.4.13 LPC Use Within Frameworks by Speakers

To summarise the use of LPC orders by speakers, the median, minimum and maximum LPC orders were determined across all of the analysis frames for each speaker for each analysis framework. As noted in Section 6.3.2, for every analysis frame in each of the analysis frameworks, the LPC order that resulted in the minimum error for each formant was recorded. It is these values that were used to determine the summary measures of LPC order for each speaker. For the analysis frameworks where the LPC order was

permitted to be different across the three formants, the summary values were calculated separately for each formant. For the frameworks were the LPC order was fixed across the formants only a single set of summary values was calculated. The frameworks where the LPC order was fixed across all tokens for all speakers have not been considered in this part of the analysis.

The summary statistics for each speaker were displayed on a series of plots, with one plot per framework where the LPC order was fixed across formants, and one plot per formant per framework where the LPC order was variable across formants. In each plot the speakers were ordered according to increasing median LPC order and range. An example plot for the condition where the LPC order is fixed within tokens and across formants with the absolute error criterion is shown in Figure 6.28.

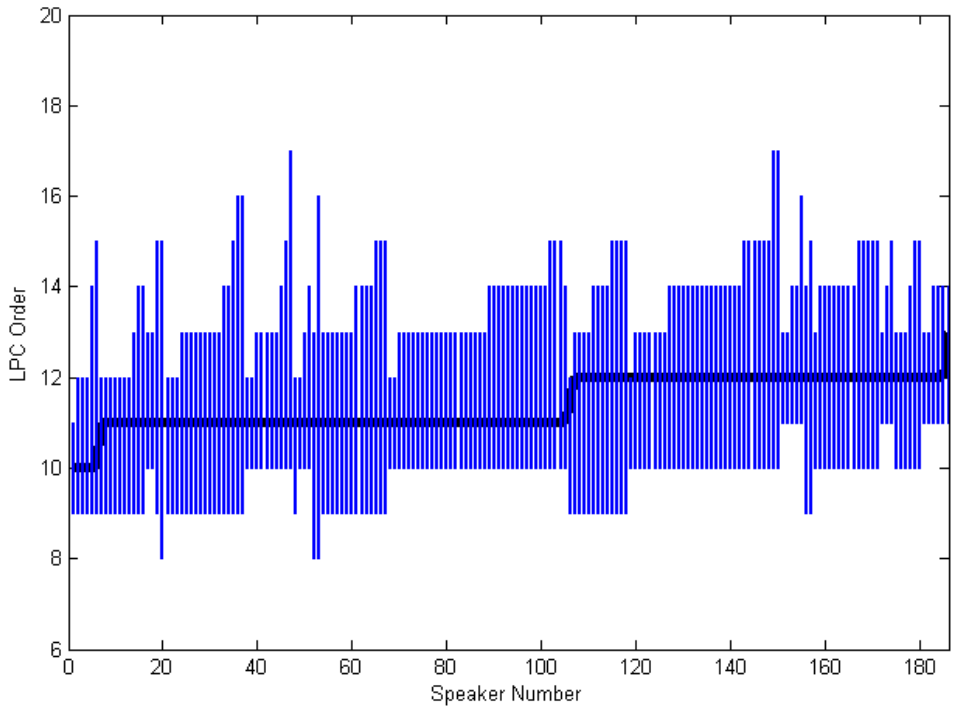

Figure 6.28 Plot of median LPC order (thick horizontal line) and range (thin vertical line) for all speakers ordered by increasing median value and range. The results originate from the framework where LPC order is fixed within tokens and across formants, with the absolute error criterion.

In Figure 6.28, the range of median LPC orders is from 10 to 13. Overall, the lowest LPC order encountered across the speakers was 8 whilst the highest was 17. The smallest range displayed by one speaker is 2, whilst at the other extreme one speaker has a range of 7. In addition to the plots, the same summary data was calculated for each analysis framework. For the frameworks where the LPC order could vary across the

formants, the typical median LPC orders decreased from F1 to F3, as did the range of the medians. This is the same as the overall results for the entire set of results presented in Section 6.3.2. Also, the typical range of LPC orders encountered decreases from F1 to F3. Comparison of the benchmark framework with the frameworks with LPC order fixed just within tokens reveals a reduced range of LPC orders.

The frameworks where the LPC order was restricted across the formants show a much reduced range of median LPC orders. Where the LPC orders are variable within a token the range of LPC orders is greater than where the LPC order is fixed within the tokens.

Overall, as the LPC order becomes more restricted across the frameworks, the range of LPC orders used by the individual speakers is reduced. However, even within the most restrictive framework, it is apparent that different speakers are using different ranges of LPC orders.

6.3.4.14 LPC Variation Across Formants in Frameworks

For three of the frameworks the LPC order is permitted to vary independently for each formant. To find out whether or not there was a linear relationship across the three formants in terms of the mean LPC order used by each speaker, scatter plots of the median LPC orders for F1 vs F2, F1 vs F3 and F2 vs F3 were generated for each framework. Within each of the plots a general positive tendency was apparent for each formant pairing but the data points were relatively dispersed. The relationships were not as strong as those described in the following section for the cross framework comparisons. However, they were somewhat stronger than the relationships seen across the formants in terms of the magnitude of the errors, discussed in Section 6.3.4.7.

6.3.4.15 LPC Use Across Frameworks

Comparison of the median LPC orders for speakers across the analysis frameworks provides an indication of how stable the speakers' use of LPC orders is across the frameworks. Again, this was done by examining scatter plots of the median LPC orders across the frameworks. For the frameworks where the LPC order is fixed across the formants, all the cross framework comparisons showed a very strong degree of linear dependence. When these frameworks were compared with the frameworks where LPC order could vary across the formants, the strongest relationship was with the median LPC orders obtained for the F3 measurements. The relationship for F2 was slightly less

strong, and even less so for F1. These results suggest that when the LPC order is fixed across formants the behaviour of the LPC orders for individual speakers is most similar to that for F3 when the LPC order is not restricted.

Comparisons of the LPC orders for formants across frameworks showed that for individual formants, the usage of LPC orders by individual speakers is very similar across frameworks.

6.3.4.16 LPC Use Compared With Sex

Having previously considered whether the mean errors for speakers are associated with various attributes of the speakers, such as sex, mean fundamental frequency and location within the vowel space, these parameters are now examined against the median LPC order usage for speakers. As stated above, if strong relationships could be shown then they could be used to help determine suitable LPC orders for speakers.

In order to consider the sex of the speaker against the median LPC order usage the same approach of determining the rank order of the speakers was used, but on this occasion it was derived from the median LPC order rather than the mean speaker error. Histograms were again generated showing the number of speakers of each sex within the bins. For all frameworks a clear pattern emerged where the tendency was for the female speakers to use lower LPC orders whilst the male speakers used higher LPC orders. The patterning seen was very similar to that shown in Figure 6.26 for the distribution of the sexes across mean percentage speaker error. As observed in the plot there was a large degree of overlap between the sexes across the LPC orders but the gradual transition across the speakers was evident for all frameworks. The same pattern was also observed for the frameworks where LPC order was not restricted across the formants and median LPC orders had been calculated for the three individual formants.

6.3.4.17 LPC Use Compared With Fundamental Frequency

Scatter plots were generated to compare the median LPC order for each speaker against their mean fundamental frequency for all of the frameworks. The data points in the plots were again colour coded to allow the male and female speakers to be easily identified. The separation of the two sexes was clearly evident in the fundamental frequency direction as already noted above when considering the mean errors. Across the frameworks a weak tendency was apparent in the data with speakers with a lower mean

fundamental frequency tending towards having higher LPC orders and speakers with higher fundamental frequencies tending towards lower LPC orders. However, this relationship was only moderate.

The patterning within the data is clearly linked to the fact that male speakers tend to have a lower mean fundamental frequency than female speakers. Since the results from the previous section showed that male speakers tended toward higher LPC orders it is no surprise that a similar pattern is seen again when fundamental frequency is compared with mean LPC order.

6.3.4.18 LPC Use Compared With Vowel Space Usage

The final section of data analysis considers the relationship between the location of speakers within the vowel space and their usage of LPC orders. Again, the location of speakers within the vowel space was represented by the mean of the reference values from the VTR database. A series of scatter plots were generated to compare each of the three formants with the median LPC orders used by each of the speakers for each of the analysis frameworks. For the frameworks where the LPC order was constrained across the three formants, the same median LPC order was plotted against the mean reference values for each formant. For the frameworks where the LPC could vary across the formants the median LPC order for each formant was only compared with the corresponding mean reference values. In all of the plots the sex of the speakers was identified by different coloured data points.

All of the scatter plots showed a moderate negative linear tendency, with some variation across the formants and frameworks, i.e. the higher median LPC orders were generally associated with lower mean reference formants, whilst the lower median LPC orders were located with the higher mean reference formants. The patterning was somewhat more apparent than for the comparison of mean fundamental frequency described in the previous section. Again, the two sexes formed two relatively distinct groupings in the plots with the male speakers generally having lower mean reference formant values than the females.

6.4 Summary

The methodology employed in this chapter involved comparing formant measurements made using Praat's LPC tool across a range of LPC orders for a subset of the TIMIT

speech corpus with a set of reference formant values (Deng et al 2006). The reference values allow the accuracy of the measurements to be determined. The experiments undertaken were focused on the second and third research questions:

RQ 2. How does altering the LPC analysis parameters affect formant measurement accuracy?

RQ 3. To what extent does the accuracy of LPC formant measurements vary across speakers?

Using the reference formant values from the VTR database allowed the performance of a large number of speakers to be assessed, which provided answers to RQ3. By considering this performance over a range of LPC orders and applying different analysis frameworks that replicate the decision analysts might make when measuring formants, the findings also addressed RQ2.

The key outcomes can be summarised as follows:

- Overall, the measurements that were obtained when the LPC order was constant across all speakers and analysis frames revealed the same behaviour as seen for the synthetic speech but the errors were larger.
- Allowing the LPC order to vary across the formants and analysis frames via the application of different analysis frameworks showed a clear reduction in the magnitude of the errors. The greater the restriction on the variation of the LPC order, the smaller the increase in performance. Keeping the LPC order the same across the three formants within a token resulted in worse performance than if it was restricted across the frames of a token.
- The performance of individual speakers was shown to vary within the group, and their relative performance was reasonably consistent across the frameworks.
- No strong relationships were found between the performance of speakers and the parameters of speaker sex, fundamental frequency and their location within the vowel space.
- Examination of the variation in LPC orders over the vowel space from the different analysis frameworks highlighted the tendency for different LPC orders to be used within different regions. Also, the patterning across the three formants was different.

- The LPC orders used by individual speakers for the analysis frameworks were shown to have different central tendencies and ranges across the speakers.
- Comparison of the LPC orders used by speakers with the parameters of speaker sex, fundamental frequency and location within the vowel space showed relationships that were stronger than those found when compared with speaker performance. However, the correlations were only moderate and showed negative linear tendencies, i.e. higher LPC orders were more aligned with speakers with lower F0.

These findings support the guidance provided in Section 4.5, that LPC order should be adjusted where necessary for each vowel token and formant, and again highlight the magnitude of the errors that can occur if inappropriate orders are used. Further guidance that is motivated by these findings is that LPC order should be tailored to individual speakers to obtain the most accurate measurements. Unfortunately, the speaker attributes of sex, mean fundamental frequency and vowel space position do not provide strong indicators for suitable LPC orders or likely performance.

Chapter 7 Performance of Formant Trackers

7.1 Introduction

The work presented in Chapter 6 examined the performance of Praat's normal formant measuring tool when analysing real speech. As previously discussed, the measurements made by this tool are not subject to any formant tracking processes. To assess whether formant trackers produce more accurate measurements, and investigate the potential differences in performance between tools, the experiments reported in the current chapter consider formant measurements from the same speech material, made with three formant trackers. In doing so the findings address all three research questions:

RQ 1. What influence does the LPC formant measuring tool have on the accuracy of formant measurements?

RQ 2. How does altering the LPC analysis parameters affect formant measurement accuracy?

RQ 3. To what extent does the accuracy of LPC formant measurements vary across speakers?

The reference values from the VTR database are again used to assess the performance of the formant trackers and the results are examined in terms of their variation across the vowel space and across speakers. As well as making the measurements over a range of LPC orders, the influence on performance of other parameters relating to the tracking functions of the tools is considered. The results are also compared with those from other studies that have used the VTR database, which raises several methodological issues.

The findings from this chapter build on those from the previous chapters concerning the speakers and the analysis parameters, and strengthen the conclusions drawn. They also highlight the differences in performance that can be expected between different measuring tools, and their influence on the behaviour of the measurements. The guidance derived from these results highlight further pitfalls when using formant trackers, but also suggest ways to further improve measurement accuracy.

7.2 Formant Trackers

Three LPC-based formant trackers were selected in order to examine the differences in performance across formant measuring tools. The first two trackers are those in the

Praat and Wavesurfer[6] software. They were chosen because they are freely available to download from the internet and the software is widely used within the forensic speech science and phonetic communities. WaveSurfer was also chosen because it has been used to provide benchmark results in several other studies that have used the VTR database, for example, Mehta et al. (2012), Smit et al. (2012) and García Laínez et al. (2012). The third tracker has been developed by Dr Frantz Clermont (Clermont 1991, 1992) and is known as the Iterative Cepstral Analysis by Synthesis (iCAbS) tracker. This tracker applies a novel approach to the formant tracking problem and whilst not currently freely available, variants of it have been used by the author, and others, in several research projects, for example, Clermont (1991), Clermont et al (2008) and French et al. (2012).

The three trackers selected all follow a two stage measurement process. In summary, the first stage is an LPC analysis that produces a set of candidate formant values, whilst the second stage processes those candidates via a series of rules to arrive at the estimated formant values. Praat's Burg tool, used in the previous chapters, does have this structure, but the rules that are applied to the candidate formant values are very limited. The rules are that candidates below 50 Hz and within the upper 50 Hz of the analysis bandwidth are rejected. For those that remain, the lowest candidate is designated as F1, the next lowest as F2 and so on. However, the term tracker is not applied to this method since the rules are very basic and no frame to frame information is used to arrive at the formant estimates.

The following sections discuss the fundamental operating principles of the three chosen trackers, as well as the specific implementations employed in this study, together with the analysis settings used in the various test conditions.

7.2.1 Praat Tracker

7.2.1.1 Principles

The tracker function within Praat extracts a specified number of formant tracks from a set of candidate values derived from an LPC analysis. This is done by considering all the possible combinations of the candidate values for each analysis frame and all of the possible tracks through them from one frame to the next. For each set of possible tracks,

[6] The formant tracking algorithm that is used in WaveSurfer is the same as that used in the Entropic ESPS X Waves software.

a series of values, known as costs, are calculated. These costs are based on how far each candidate formant deviates from a set of reference values, how wide the bandwidth of each formant is relative to its centre frequency, and how large the jump in frequency is between consecutive frames. The set of candidates that are chosen as the formant tracks are those that overall produce the smallest costs. This process favours candidate values that are closest to the specified reference values, have the smallest bandwidths and have the smallest jumps in frequency across frames. The calculation of the costs and the method of combining them are given in the Praat manual (Boersma, 2002).

The function requires several parameters to be specified. The first of these is the number of tracks to extract. For the function to operate there must be at least this number of candidates in each frame. Reference formant values, from F1 to the number of formant tracks to be extracted, must also be provided. The default values suggested in Praat's manual are for a neutral vowel derived from the odd harmonics of a lossless tube which is open at one end and has the length of a typical female vocal tract. The final three settings are the frequency cost, the bandwidth cost and the transition cost, which weight each of the calculated cost values described above. No parameters relating to the LPC analysis are specified for the tracker function since it can only be applied to candidate formant values, not directly to a sound file.

7.2.1.2 Implementation & Settings

The formant tracker function within Praat, called by the command 'Track…', is distinct from the 'Sound: To Formant (burg)…' function previously used in this thesis. The 'Track…' function only operates on formant objects within Praat, where formant objects are a data structure containing formant values obtained from applying a function, such as 'Sound: To Formant (burg)…', to a sound object. The candidate formants must already have been created before the tracker process can be run.

Since the tracker function requires that formant measurements are made first, a modified version of the script file used in the previous chapter for Praat's standard formant measuring tool (see Section 6.2.7) was used to both obtain the candidate formant measurements and perform the tracking. The script was altered to include the tracking command after the initial formant measurements were made. The settings used to obtain the initial candidate formant values were the same as those used in the previous chapter (see Sections 6.2.4 to 6.2.6). Again, the upper analysis frequency was

set at 5,000 Hz for the male speakers and 5,500 Hz for the female speakers. The only formant measurement parameter that was varied was the LPC order.

To investigate the effect of altering some of the formant tracking parameters, three different sets of measurements were made, each with different combinations of analysis settings. For all three, the tracker settings of 'Frequency cost', 'Bandwidth cost' and 'Transition cost' retained their default values of 1. Whilst it would be interesting to explore the effects of altering these parameters, such work is outside the scope of this study.

For the first two sets of measurements the reference formant values were kept at the values that represent a typical neutral vowel, i.e. 500 Hz, 1,500 Hz, 2,500 Hz and 3,500 Hz for F1 to F4 for the male speakers and 550 Hz, 1,650 Hz, 2,750 Hz and 2,850 Hz for F1 to F4 for the female speakers. The first series of tests, named the 'Default' condition, were run with the 'Number of tracks' at the default value of 3, with the LPC order varying from 6 to 20.

The second series, referred to as the '4 formant' condition, was made with the 'Number of tracks' set to 4 formants and the LPC order was varied from 8 to 20. This parameter was selected as a variable because whilst its function is obvious, and may well be changed by an analyst, its impact on the accuracy of the measurements does not appear to be documented.

For the final set of measurements, referred to as the 'Optimum' condition, the 'Number of tracks' was set at 4 formants, with the LPC order varying from 8 to 20, and the reference formant values were altered to an 'optimum' set for every vowel token. The values used were the mean reference values for the specific token obtained from the VTR database, i.e. a value very close to the true value of the formant. This required the values from the VTR database to be read by the Praat script and passed to the tracker function for each vowel token. The reference values were chosen as a parameter to alter because again, the influence of varying them is not documented. The parameter is of interest because many speech corpora have time aligned segmental transcripts and such information could be used to select a set of relevant average reference values for each token based on the category of the transcribed vowel. The reference values from the VTR database were used in this instance, rather than average values for each vowel category, so that the approach could be tested using what are effectively the best

possible reference values. A similar approach of specifying reference values on a token by token basis could also be adopted by analysts making formant measurements interactively.

A summary of the analysis settings used for the three conditions is shown in Table 7.1.

No	Condition Name	LPC Range	Number of Formants	Reference Formant Values
1	Default	6 to 20	3	Default
2	4 formants	8 to 20	4	Default
3	Optimum	8 to 20	4	Optimum from VTR database

Table 7.1 Analysis parameters used for the three conditions used to measure formants in the VTR database with the Praat tracker.

One significant difference between the analysis approach used for the Praat tracker and that used in the previous chapter concerns how the measurements were made in the sound file. For the previous chapter, formant measurements were made across the entirety of each file. This was done to simplify the analysis procedure within Praat since the measurements from individual frames were not influenced by those surrounding them. The determination of which frames corresponded to vowels was undertaken at a later stage when the formant measurements were analysed within Matlab. However, this approach is not necessarily desirable when using the formant tracker since the transition cost element means that the selection of candidate formant values is influenced by those surrounding them. As the effect of including non-vocalic segments within the measurement process was not known, and the Praat manual suggests that the function should only be applied to vowels (Boersma, 2002), it was decided to restrict the measuring of the formants to the vocalic sections only. Also, the Optimum condition could not be tested if measurements were made across the entirety of the file. However, this issue was examined for WaveSurfer (see Section 7.2.2.2).

The vocalic sections of the sound files that were subject to analysis were determined from the phonetic transcripts that accompany the TIMIT sound files. The timings for the analysis frames were selected so that they aligned with those made in the previous chapter. This also ensured that the alignment of the measurements with the VTR reference values remained constant within sound files. However, the overall alignment of the measurements from the three trackers with the VTR reference values was different from the previous chapter. This is discussed in Section 7.3.1.

7.2.2 WaveSurfer

7.2.2.1 Principles

The formant tracker within WaveSurfer follows the same basic approach as Praat's tracker in order to arrive at the formant estimates, i.e. it selects the formant candidates that produce the minimum cost values associated with the formant frequency, formant bandwidth and frame to fame differences (Talkin, 1987). However, at a practical level, the function combines both the LPC analysis that produces the candidate values, and the tracking process. Therefore, a number of parameters for the LPC analysis must be specified. The parameters and the values used for them in this study are described in Section 7.2.2.2. Unlike Praat, only one parameter relating to the tracking element of the function can be specified in WaveSurfer. That parameter is the nominal or reference value for the first formant. Rather than being able to specify the reference values for each formant, WaveSurfer calculates the reference values for the higher formants based on the value given for F1. There is no option to modify the behaviour or weighting of the other elements involved in the tracking process. These are fixed within the software.

7.2.2.2 Implementation & Settings

The WaveSurfer software package itself is not scriptable. However, WaveSurfer is built on a set of functions which are known as the Snack Sound Toolkit[7] (Sjölander, 1997). These functions can be utilised from within programming languages such as Tcl/Tk or Python. For this study the Tcl/Tk language was chosen as this had been used for the author's Masters research (Harrison, 2004) discussed in Chapter 3. The script used in that study was modified for the current tests. The modified script follows the same basic procedure used for the Praat tracker script, i.e. it sequentially opens all of the sound files and for each one performs the formant analysis, and logs the formant measurements and the settings used to obtain them. The operation of the Snack script was checked to confirm that the measurements obtained in this way were identical to those made using the WaveSurfer software (version 1.8.8p3) with the same analysis settings. Identical results were obtained from the Snack script and WaveSurfer when a number of files were compared[8]. To avoid confusion, the results in the following sections will be

[7] The specific version of the Snack toolkit used in this thesis was 2.2.10 which was part of the Tcl software. The Tcl software was Active State Tcl version 8.4.19.6295590 win32 ix86 threaded, released on 8th February 2012.

[8] The TIMIT audio files contain a 1024 byte header at the start of the file that needs to be ignored when opening them in WaveSurfer and with the Snack Toolkit.

attributed to WaveSurfer but Snack will continue to be discussed in this section in relation to the script and the implementation.

A number of different analysis conditions, i.e. combinations of settings, were employed when measuring the formants in order to investigate the change in performance caused by altering them. For most of these conditions, the Snack script was configured to measure the formants across the entire sound file, rather than within vowel tokens. The frames that corresponded to vowels were determined later in the processing within Matlab when the measurement errors were calculated. The reason this approach was adopted was so that the results would be comparable with those from other studies discussed in Section 7.4, where WaveSurfer was tested using the VTR database. However, as discussed above for the Praat tracker, such an approach could have a detrimental effect on performance since the formants must also be tracked through non-vocalic segments. In order to examine the effect of this approach on the measurements, one of the measurement conditions only measured the formants within the vowel tokens (see the Vowels condition below). For this condition, the same method described for the Praat tracker was used to determine the timings of the analysis frames.

The six different analysis conditions for which formant measurements were made are summarised in Table 7.2. The sets of parameters were chosen to reflect WaveSurfer's default settings ('Default'), the equivalent configuration used for the Praat tracker ('Vowels'), the setting used by García Laínez et al. (2012) ('Hamming') and two intermediate states ('25 ms' and '25 ms Hamming'). The final condition, '3 formants', was chosen in order to investigate the effect of the 'Number of Formants' parameter. Snack imposes a condition on the LPC order so that it must be at least four more than twice the number of formants, i.e. for 4 formants the minimum LPC order is 12. So for all the conditions, apart from the '3 formants', the range of LPC orders tested was 12 to 20.

No	Condition Name	LPC Order Range	Number of Formants	Window Length (ms)	Window Type	Speech Analysed
1	Default	12 to 20	4	49	Cos^4	All
2	25 ms	12 to 20	4	25	Cos^4	All
3	Hamming	12 to 20	4	49	Hamming	All
4	25 ms Hamming	12 to 20	4	25	Hamming	All
5	Vowels	12 to 20	4	25	Hamming	Vowels
6	3 formants	10 to 20	3	49	Hamming	All

Table 7.2 Analysis parameters used for the six conditions used to measure formants in the VTR database with the Snack tracker.

The remainder of the analysis parameters were given their default values across all of the conditions. The time step, or frame advance setting, was 0.01 seconds, so that it corresponded with the time difference of the measurements in the VTR database. Within WaveSurfer this parameter is referred to as 'Frame interval', but for the Snack command it is confusingly called 'Frame length'. The LPC analysis type was set to autocorrelation, which is specified by a 0 both within the script and in WaveSurfer. The pre-emphasis factor was set at 0.7. The sampling rate, which determines the upper frequency limit of the signal when it is resampled before the LPC analysis, was set at 10,000 Hz. The nominal value for the first formant frequency was kept at 500 Hz[9]. Unlike the Praat tracker, these two values were not adjusted according to the sex of the speaker. This was done to provide comparability with the WaveSurfer results reported in the other studies discussed below.

7.2.3 iCAbS

7.2.3.1 Principles

The Iterative Cepstral Analysis by Synthesis (iCAbS) tracker is a development of the earlier Cepstral Analysis by Synthesis (CAbS) tracker (Clermont, 1991; 1992). Rather than using heuristic approaches to determine the candidate values most likely to be formants, such as those used by Praat and WaveSurfer, the CAbS tracker uses an objective measure derived from the speech signal. The tracker selects the candidate formants which produce the best alignment with the linear prediction cepstrum measured from the speech signal. For a given frame, synthetic cepstra are generated for all of the possible combinations of the candidate formants. Each synthesised cepstrum is compared with the measured cepstrum and the cepstral distance is calculated. The

[9] In WaveSurfer the default value for the nominal F1 frequency is -10 Hz. Tests were conducted in both Snack and WaveSurfer and they confirmed that identical measurements are produced when the setting is either -10 Hz or 500 Hz.

cepstral distance measure used is more sensitive to differences around spectral peaks (Yegnanarayana and Reddy, 1979) and allows the frequency range of the comparisons to be specified (Clermont and Mokhtari, 1994). It is also possible to apply a dynamic programming approach to minimise the cepstral distances across a number of frames.

Previous work has shown that the settings chosen for the LPC order and the upper cepstral comparison frequency can markedly affect the reliability of the tracker (Clermont et al. 2007). Also, the optimum values for these parameters can vary across speakers and conditions. As a consequence of this the iterative version of the CAbS tracker was created, which automatically cycles through a specified range of LPC orders and upper cepstral comparison frequencies (Clermont et al. 2007). For each combination of the two parameters the CAbS tracker is applied and a set of tracked formants are produced. To determine which combination of parameters has produced the best formant tracks a continuity quality value is calculated, which is the average frame to frame difference for F1 to F3. The set of tracked formants with the minimum continuity quality value are presented as the final output values for the tracker.

7.2.3.2 Implementation & Settings

The implementation of the iCAbS tracker used in this chapter was coded by the author using a combination of the programming language Perl and the scripting capabilities within Praat. The tasks of opening the sound files, determining where the vowel tokens occurred and the logging of the formant values was undertaken by a Praat script that was essentially the same as the one used for the Praat tracker. In terms of the actual formant measuring, the 'To LPC (autocorrelation)...' function within Praat was used to obtain the initial LPC values which are then passed to a Perl script. The Perl script processes the LPC values to obtain the LP cepstrum from the signal as well as the candidate formant values by root solving. It is only necessary to generate the LPC coefficients for the highest LPC order being considered because the coefficients for the lower LPC orders are derived from the initial set (Clermont, 2011, personal communication). The remainder of the processing, including the determination of the combinations of candidate formants, the generation of the synthetic cepstra, the cepstral distance calculations and calculations of the continuity quality, is conducted by the Perl script. The final tracked formant values are passed from the Perl script back to the Praat script for logging.

The analysis parameters used by the LPC function were a window length of 25 ms, a time step (frame advance) of 10 ms and a pre-emphasis setting of 50 Hz, i.e. the same as those used previously for Praat's Burg tool and the tracker. The iCAbS tracker uses the autocorrelation method within Praat for the LPC analysis, rather than the Burg method. Again, the maximum formant frequency setting, which determines the sample rate of the speech file before being subject to the LPC analysis, was set to 5,000 Hz and 5,500 Hz for male and female speakers respectively.

A total of six sets of analysis parameters were tested with the iCAbS tracker. The parameters that were varied were the LPC order range, the number of formants to be extracted and the upper cepstral comparison frequency range. The settings used for each of the six conditions are set out in Table 7.3. The settings were chosen to represent a number of combinations that could be adopted, including limiting the range of LPC orders to a single order, i.e. removing the iterative aspect for the LPC order. The step size for the upper cepstral comparison frequency was 250 Hz for all conditions.

No.	Condition Name	LPC Order Range	Number of Formants	Upper Cepstral Comparison Frequency Range (Hz)
1	Default	8 to 16	4	3,000 to 5,000
2	3 formants	8 to 16	3	3,000 to 5,000
3	LPC 8 to 14	8 to 14	4	3,000 to 5,000
4	LPC 12	12	4	3,000 to 5,000
5	LPC 16	16	4	3,000 to 5,000
6	LPC 12, upper comp freq 4 kHz	12	4	3,000 to 4,000

Table 7.3 Analysis parameters used for the six conditions used to measure formants in the VTR database with the iCAbS tracker.

Like the Praat tracker, the iCAbS tracker was only used to measure formants within the vowel tokens, not across the entire sound files. The same process for determining their location was used. For each vowel token analysed the script also logged the LPC order and upper cepstral comparison frequency that the tracker had selected as providing the best formant values, as well as the continuity quality value for those formants. These results are examined in Section 7.3.2.3 as part of the analysis of the performance. The dynamic programming option of the CAbS tracker was not included in the current implementation of the iCAbS tracker. This was done to reduce the computational requirements and because the frame to frame continuity is already considered in the

continuity quality calculation used to determine the optimum LPC order and upper cepstral comparison frequency.

7.3 Analysis of Measurement Errors

The following sections examine the measurement errors from the three formant trackers with the analysis conditions described above. The first section discusses the alignment of the measurements both across the trackers and with the reference values from the VTR database, which is different from that in the previous chapter. This is followed by a brief re-examination of some of the measurements from the previous chapter with the new alignment. The next section examines the overall results from the three trackers for all of the tests undertaken. This is followed by a re-examination of a subset of the results from the Praat tracker and WaveSurfer to determine the minimum possible errors that can be achieved if the LPC order is permitted to vary across vowel tokens. The following section considers how the overall and minimum errors behave over the vowel space. The next section examines the results across the speakers. The final section considers how the results compare with those from other studies that have used the VTR database.

7.3.1 Alignment of Measurements

The remainder of this chapter concerns the analysis and comparison of the measurement errors from the three different formant trackers. To ensure that the results are comparable it was necessary to consider the alignment of the measurements across the trackers. This is discussed in Section 7.3.1.1. Related to this is the issue of the alignment of the measurements with the VTR reference values. This topic was revisited following the initial calculation of the measurement errors and the comparison of them with the results from other studies. This is discussed in Section 7.3.1.2.

7.3.1.1 Alignment of Measurements Across Trackers

In the case of the Praat and the iCAbS trackers, it was simple to confirm the alignment of the measurements, as both sets had been produced using very similar scripts and, more importantly, each measurement was assigned a timing within Praat that corresponded to the centre of the analysis frame. Comparison with the WaveSurfer results was less straightforward as neither WaveSurfer nor Snack assigns timings to the exported or logged measurements. But, information is provided in the Snack documentation which states, ‘the first row corresponds to a start time of half the

window length' (Sjölander, 2004). This can be interpreted in two different ways; firstly, it could mean that the first analysis frame is left aligned with the start of the recording, so the timing of the first measurement is half the window length, or, secondly, it could mean that the start of the first analysis frame is located at a time which is half the window length, so the centre of the first analysis frame occurs at a time equivalent to a whole analysis frame. With the WaveSurfer default window length of 0.049 seconds the timing difference between the two interpretations is 0.0245 seconds, which, with a frame advance of 0.01 seconds, is almost two and a half frames different. Examination of a number of sound files and formant measurements at different frame lengths within WaveSurfer confirmed that the first interpretation was the correct one. This meant that for all three trackers the timings of the frames were the equivalent of having the first frame aligned with the start of the sound file.

For the analysis conditions where the frame length was 25 ms, the alignment of the frames across the three trackers was identical. However, for the WaveSurfer conditions with a 49 ms frame length, the centres of the frames were offset by 2 ms from those with a 25 ms frame length. No attempt was made to compensate for this offset since WaveSurfer was tested with both frame lengths allowing the 25ms frame length results to be directly compared with the other trackers.

7.3.1.2 Alignment of Measurements with VTR Reference Values

Initially, the formant measurement errors were calculated using the same alignment with the VTR reference values described in the previous chapter. This was based on the assumption that the start of the first VTR analysis frame was aligned with the start of the sound file. However, comparison of the results obtained for WaveSurfer with those presented by Deng et al. (2006), the creators of the database, showed them to be somewhat different, especially for F2. Contact was again made with the authors of the database in order to discover if their methodology for obtaining the measurements from WaveSurfer was significantly different from that described above. It was confirmed that their analysis settings remained constant across all utterances, that the formants were tracked across entire files, that segments were determined according to the TIMIT segmentation information, and that the measurements were made with WaveSurfer rather than the Snack toolkit (Deng, 2013, Cui, 2013).

To further investigate the difference in performance, the measurement errors for WaveSurfer's Default condition at LPC order 12 were re-calculated across a number of different alignments with the VTR reference values. The mean absolute error values for the three formants at the different alignments are shown in Figure 7.1. The offset values show the shift in the alignment of the VTR reference values relative to the WaveSurfer measurements. The 0 ms offset corresponds to the alignment used in the previous chapter. The negative offset values correspond to the VTR values shifted backwards in time relative to the WaveSurfer measurements, whilst the positive values are a forwards shift in relative time.

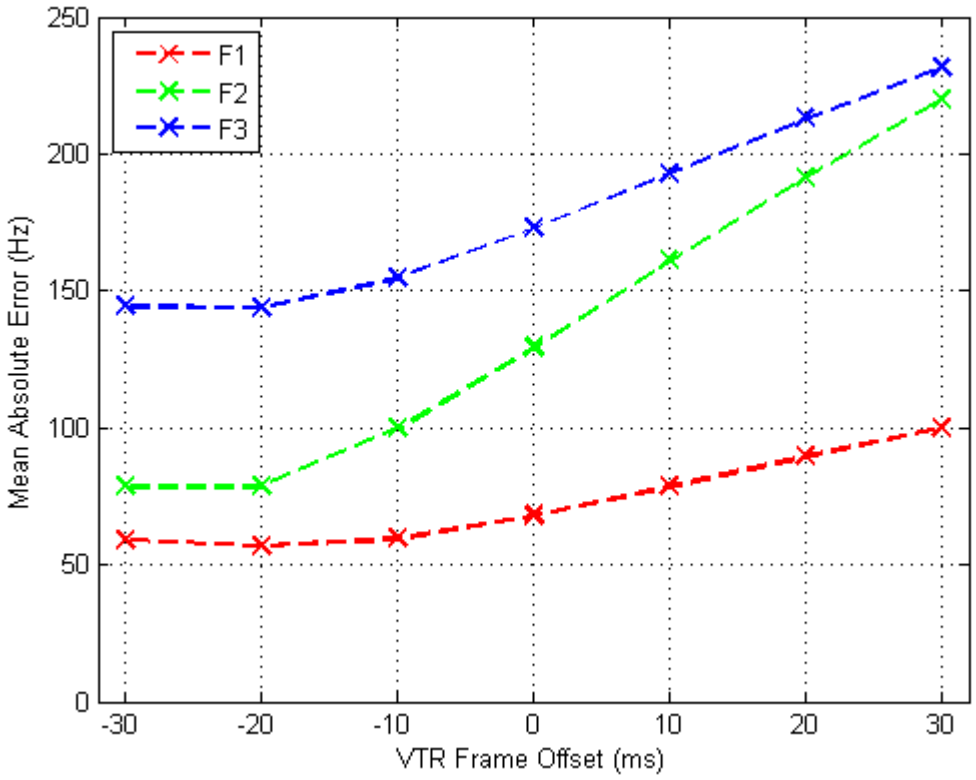

Figure 7.1 Mean absolute error values from WaveSurfer's Default condition at LPC order 12 across different time alignments with the VTR reference values for F1 (red), F2 (green) and F3 (blue).

The plot clearly shows that the magnitude of the errors is dependent on the alignment of the measurements with the VTR reference values. The best performance is achieved with an offset of -20 ms. At this offset, the performance is better than that reported by Deng et al. (2006).

Given these findings, it was decided to re-examine the alignment of the VTR reference values with the speech signal. This was done using WaveSurfer, which allows log files

from other sources to be loaded into the program and overlaid on spectrograms. It is also possible to easily adjust the offset or time alignment of the data points. Several TIMIT samples and their accompanying VTR reference values were loaded into WaveSurfer and the alignment of the VTR values was adjusted across a range of offsets. It became clear that the alignment applied in the previous chapter did not provide the best overall visual match. This was most apparent for diphthongs with large changes in F2. A range of negative offsets did result in better alignments but no objective visual criterion could be used to determine which one was the most appropriate. The offset value within WaveSurfer that was selected was -0.0165 seconds. If it is assumed that the frame length used to create the VTR values is 25 ms then this shift corresponds to a change in alignment of -29 ms relative to the alignment used in the previous chapter. This alignment was also selected since it resulted in the centre of the analysis frames being equidistance between the frames for the 49 ms and 25 ms frame length conditions for the trackers.

7.3.1.3 Comparability with Results from Chapter 6

The change in alignment from the previous chapter has an effect on the comparability of the earlier results with those presented below for the trackers. It is clear from Figure 7.1 that varying the alignment alters the magnitude of the errors, with the greatest change occurring for F2. However, it is assumed that the overall tendencies and patterns seen within the results will not be significantly affected. This is because the bias within the results that is introduced by the misalignment is consistent through the entire set of results. This assumption is also supported by a comparison of the results presented below with the equivalent results calculated with the same alignment as the previous chapter. Arrangements of the error surfaces and the relative performance of the trackers both across trackers and across conditions for the same tracker were not markedly different.

In order to allow a more direct comparison of the results from the trackers with those from Praat's Burg tool, some of the Burg errors were recalculated with the new alignment. The mean absolute error and standard deviation values for LPC 10 across all vowel frames are shown in Table 7.4.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	76.25	16.93 %	86.91	5.65 %	123.34	5.01 %	95.50	9.20 %
Standard Deviation (Hz)	108.63	28.44 %	152.09	10.59 %	208.83	8.74 %	163.94	19.30 %

Table 7.4 Mean absolute error and standard deviation from the Praat Burg tool at LPC order 10 for all vowel frames with modified VRT alignment.

Comparison of these results with the equivalent values in Table 6.4 show that the performance measured as mean absolute error for F1 has actually decreased from 75.61 Hz to 76.25 Hz, but has increased for F2 and F3 from 125.67 Hz to 86.91 Hz and from 144.12 Hz to 123.34 Hz, respectively. Again, with the new alignment, the best performance for F1 was achieved at an LPC order of 15, resulting in a mean absolute error of 59.75 Hz. This compares with a mean absolute error of 63.68 Hz in Table 6.1.

The error values were also recalculated at the new alignment for the minimum errors achieved using the 'Tkn Fix, F Fix, Abs' analysis framework. These are shown in Table 7.5.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	71.55	15.67 %	70.10	4.53 %	87.90	3.56 %	76.52	7.92 %
Standard Deviation (Hz)	82.60	20.42 %	106.89	6.83 %	134.66	5.45 %	113.65	14.22 %

Table 7.5 Mean absolute error and standard deviation from the Praat Burg tool for minimum errors ('Tkn Fix, F Fix, Abs' framework) for all vowel frames with modified VTR alignment.

These results show a similar change in performance in comparison with the previous results in Table 6.13. Again, the performance for F1 is slightly worse (71.55 Hz vs 70.84 Hz), with the greatest change being for F2 from 108.76 Hz to 70.10 Hz. The LPC orders that produced these results are very similar to those reported previously, with a median and mode LPC order of 11, an IQR of 1 and a range from 7 to 16.

7.3.2 Overall Tracker Results

The following sections present the overall results from each of the test conditions for the three trackers. For the Praat tracker and WaveSurfer results the errors are calculated at each LPC order. The calculation of the measurement errors followed the same process used in the previous chapter (see Section 6.2.7). The formant measurements were loaded in to Matlab, together with the VTR reference values, and the measurement errors were calculated. Even though many of the analysis conditions for the trackers

have the number of measured formants set to 4, the errors were only calculated and analysed for F1 to F3.

The results in the following sections address the questions raised by RQ1 and RQ2 concerning the variation in performance across software and across analysis parameters.

7.3.2.1 Praat Tracker

Like the analyses in the previous chapters, boxplots were generated to show the behaviour and distribution of the errors from all of the analysis frames at each LPC order. The configuration of the boxplots was the same as described previously (see Sections 4.4.2 and 6.3.1.1). The boxplots for the results for the Praat tracker for the 4 formant condition are shown in Figure 7.2 to Figure 7.4. Also, mean absolute error values from all frames were calculated for each formant across the LPC orders.

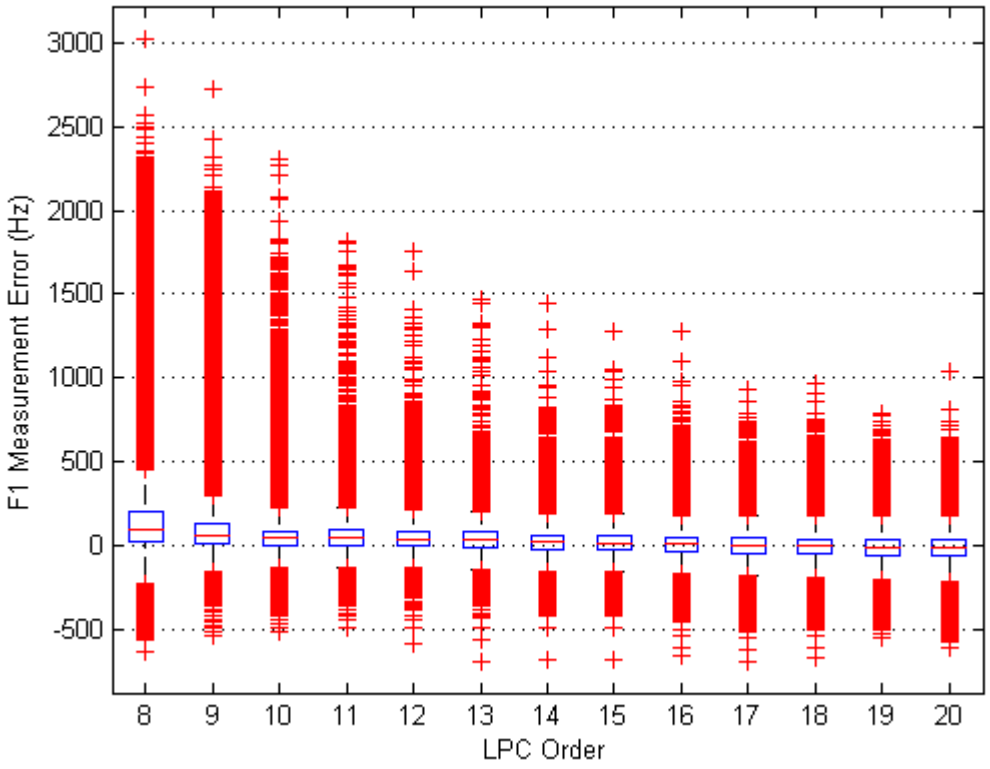

Figure 7.2 Boxplot of F1 measurement errors for all frames from Praat tracker, 4 formant condition.

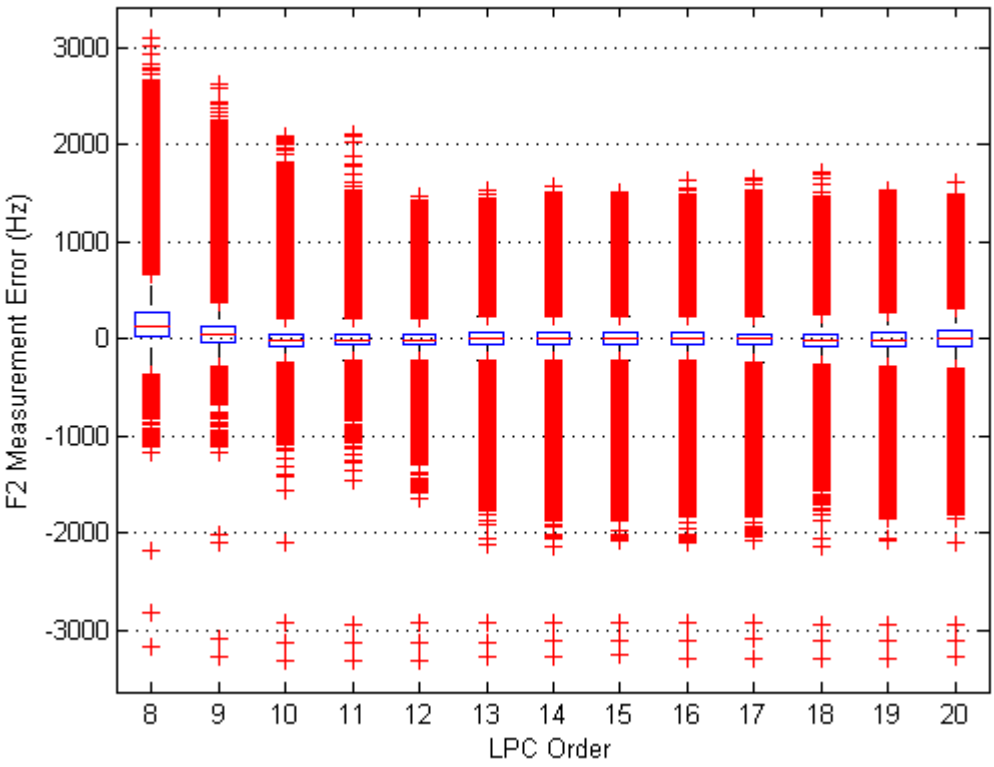

Figure 7.3 Boxplot of F2 measurement errors for all frames from Praat tracker, 4 formant condition.

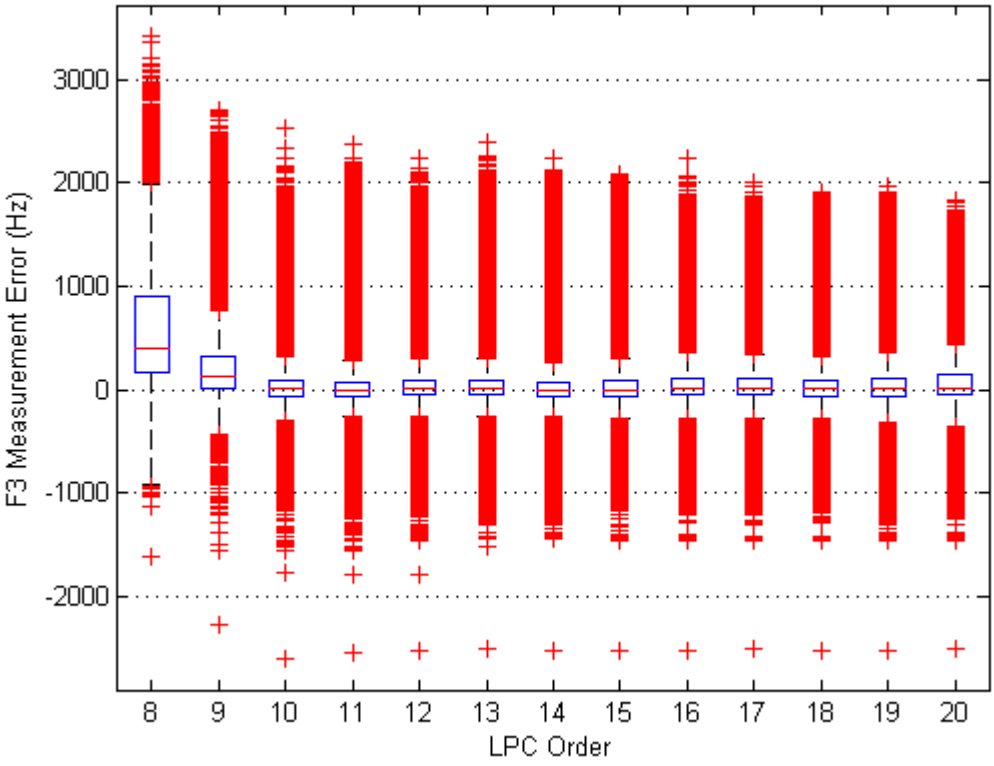

Figure 7.4 Boxplot of F3 measurement errors for all frames from Praat tracker, 4 formant condition.

The overall structure of the results in Figure 7.2 for the F1 errors is very similar to that in Figure 6.1 in the previous chapter for the F1 errors from the non-tracked Burg measurements. For both sets of results the minimum mean absolute error occurs at LPC order 15 and is 60.73 Hz for the tracker and 59.75 Hz for the recalculated Burg results at the new alignment. In contrast, comparison of the structure of the results for F2 and F3 in the figures above with those from the previous chapter reveals marked differences. For the tracker results the behaviour across the LPC orders for F2 and F3 is very much like that of F1. Unlike the non-tracked results, the errors for F2 and F3 do not change from a tendency of being overestimates to underestimates and continue to increase in magnitude as the LPC order increases. Rather, they remain positive across the LPC orders and are relatively stable.

A consequence of this is that the mean absolute errors for F2 and F3 are also relatively stable across the higher LPC orders and show only a slight increase across the LPC orders. The reason for this difference in behaviour is that the second and third formants are not restricted to taking the values of the second and third candidate formants. As the LPC order increases the number of candidate formants within the analysis bandwidth increases and as extra poles appear, the frequency of the second and third candidates will reduce, which causes the effect seen in the results from Praat's Burg tool.

The numeric results for Praat's 4 formant tracker condition are given in Table 7.6 for LPC order 11, which is the order at which the mean absolute error across the three formants is the minimum. Comparison of these values with those from the non-tracked condition at LPC order 10 (Table 7.4), which also produced the minimum combined mean absolute error, shows that whilst F1 and F2 performed better for the tracker, F3 was worse. The difference in F3 means that the average performance across all three formants was worse for the tracker. The same is also true for the standard deviation results.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	74.06	16.29 %	83.75	5.75 %	144.95	6.55 %	100.92	9.53 %
Standard Deviation (Hz)	90.07	22.82 %	145.94	12.02 %	287.29	15.01 %	194.58	18.07 %

Table 7.6 Mean absolute error and standard deviation for Praat Tracker 4 formant condition at LPC order 11.

Across the LPC orders the overall behaviour of the summary statistics for the Default and Optimum conditions are the same as those for the 4 formants condition. For the

three conditions the LPC order at which the best performance occurs, as determined by the minimum combined absolute mean error, is different. For the Default condition this is at LPC order 14, for the 4 formant condition it is at LPC order 11 and for the Optimum condition it is at LPC order 15. However, given the relative stability of the errors over the LPC orders the differences between the errors across the LPC orders within each of the conditions is relatively small. The statistical summaries for the Default and Optimum conditions at these LPC orders are given in Table 7.7 and Table 7.8.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	61.75	13.37 %	102.42	7.29 %	178.75	8.39 %	114.31	9.68 %
Standard Deviation (Hz)	81.65	19.73 %	192.27	15.55 %	356.89	19.02 %	240.39	18.31 %

Table 7.7 Mean absolute error and standard deviation for Praat Tracker Default (3 formants) condition at LPC order 14.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	60.93	13.01 %	74.89	4.96 %	90.66	3.69 %	75.49	7.22 %
Standard Deviation (Hz)	82.31	19.20 %	115.01	7.86 %	139.36	5.87 %	114.99	12.84 %

Table 7.8 Mean absolute error and standard deviation for Praat Tracker Optimum condition at LPC order 15.

Comparison of the results from the three analysis conditions reveals differences in performance across all three formants. Overall, the Default settings produce the worst performance, then the 4 formant condition, with the Optimum condition producing the best. Altering the reference values that the tracker uses on a token by token basis for the Optimum condition has markedly improved the performance compared with the other two conditions. The number of formants to be measured also has an influence on the results, with the default 3 formant condition producing better performance than the 4 formant condition for F1 but a relatively worse performance for F2 and F3. Both these conditions also produced results that were worse than those using Praat's standard formant measuring tool with the same LPC order of 10 across all the material. This result highlights the finding that using a tracker does not necessarily result in better performance.

7.3.2.2 WaveSurfer

Examination of the statistical summary results for the WaveSurfer test conditions across the LPC orders showed the same relative stability in the errors that were seen above for

the Praat tracker for the higher LPC orders. Again, a slight increase in the mean absolute errors occurs as the LPC order increases. The stability was even more apparent since the lower LPC orders, which tend to produce the largest errors, could not be tested using WaveSurfer due to the minimum permitted LPC order being 12 when extracting 4 formants. The only unstable set of errors across the LPC orders was for F3 in the 3 formants condition. At LPC order 10 the mean absolute error was 560.00 Hz. This reduced across the LPC orders to a still relatively large minimum value of 278.61 Hz at LPC order 20. However, the errors for F1 and F2 showed the same stability found across the other conditions.

The summary statistical results at the LPC order that produced the best performance for each of the conditions are shown in Table 7.9 to Table 7.14. Again, the criterion for determining the best performance is the minimum mean absolute error across the three formants. Given the poor performance of F3 for the 3 formants condition, the results are shown for LPC order 12, which produced the best performance for F1 and F2.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	61.01	12.48 %	80.75	5.16 %	136.69	5.68 %	92.82	7.77 %
Standard Deviation (Hz)	86.57	18.71 %	156.97	9.32 %	267.03	11.61 %	186.85	13.88 %

Table 7.9 Mean absolute error and standard deviation for WaveSurfer Default condition at LPC order 13.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	68.05	13.57 %	89.69	5.66 %	142.11	5.86 %	99.95	8.36 %
Standard Deviation (Hz)	98.41	20.62 %	181.03	10.41 %	268.98	11.65 %	196.83	14.96 %

Table 7.10 Mean absolute error and standard deviation for WaveSurfer 25ms condition at LPC order 14.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	59.33	12.24 %	78.63	5.03 %	132.30	5.50 %	90.09	7.59 %
Standard Deviation (Hz)	81.40	17.79 %	151.23	8.93 %	262.83	11.41 %	182.31	13.35 %

Table 7.11 Mean absolute error and standard deviation for WaveSurfer Hamming condition at LPC order 13.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	61.70	12.64 %	83.07	5.32 %	135.82	5.64 %	93.53	7.86 %
Standard Deviation (Hz)	86.92	18.83 %	158.22	9.45 %	262.84	11.40 %	185.13	13.92 %

Table 7.12 Mean absolute error and standard deviation for WaveSurfer 25 ms Hamming condition at LPC order 13.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	60.33	12.48 %	79.77	5.18 %	143.81	6.08 %	94.64	7.91 %
Standard Deviation (Hz)	84.97	18.93 %	139.12	8.75 %	282.62	12.78 %	190.11	14.23 %

Table 7.13 Mean absolute error and standard deviation for WaveSurfer Vowels condition at LPC order 12.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	58.18	12.12 %	85.21	5.70 %	403.02	17.43 %	182.14	11.75 %
Standard Deviation (Hz)	80.13	17.96 %	160.74	11.76 %	574.11	26.40 %	382.68	20.57 %

Table 7.14 Mean absolute error and standard deviation for WaveSurfer 3 formants condition at LPC order 12.

For all of the conditions, apart from the 3 formants condition, the performance is very similar. Even for the 3 formants condition the results for F1 and F2 are comparable with the other conditions. Apart from the 3 formants condition, none of the others stand out as being dramatically better or worse than the others. However, the best overall performance is achieved by the Hamming condition. These results show that the analysis parameters that were modified only have a limited impact on the performance of WaveSurfer.

Comparison of the WaveSurfer statistical summary results with those from the Praat tracker show them to generally be better than Praat's Default and 4 formant conditions, but worse than the Optimum condition results. Comparison with Praat's non-tracked Burg results at LPC order 10 shows that WaveSurfer consistently outperforms Praat in terms of F1 but the situation is reversed for F3, whilst for F2 they are similar.

7.3.2.3 iCAbS

For each of the iCAbS tracker analysis conditions only a single set of measurement errors were generated, unlike the Praat tracker and WaveSurfer, where a set of errors were generated across a number of LPC orders. This is because the optimum LPC order is automatically selected by the iCAbS tracker as part of the measurement process. The summary statistics for each of the analysis conditions are presented in Table 7.15 to Table 7.20.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	66.79	14.43 %	77.74	5.02 %	117.84	4.85 %	87.46	8.10 %
Standard Deviation (Hz)	88.40	21.21 %	134.06	8.34 %	229.18	9.98 %	162.30	15.02 %

Table 7.15 Mean absolute error and standard deviation for iCAbS Default condition.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	78.09	17.27 %	112.20	7.61 %	360.91	15.44 %	183.73	13.44 %
Standard Deviation (Hz)	137.07	36.14 %	232.58	17.99 %	573.77	26.00 %	385.83	28.10 %

Table 7.16 Mean absolute error and standard deviation for iCAbS 3 formants condition.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	69.09	15.06 %	76.41	4.96 %	118.87	4.93 %	88.12	8.32 %
Standard Deviation (Hz)	89.37	21.99 %	123.49	8.09 %	229.13	10.18 %	159.85	15.55 %

Table 7.17 Mean absolute error and standard deviation for iCAbS LPC 8 to 14 condition.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	68.99	15.20 %	77.28	5.00 %	116.82	4.80 %	87.70	8.33 %
Standard Deviation (Hz)	87.85	22.26 %	126.39	8.29 %	216.62	9.14 %	154.74	15.63 %

Table 7.18 Mean absolute error and standard deviation for iCAbS LPC 12 condition.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	61.50	12.78 %	89.17	5.66 %	123.06	4.96 %	91.24	7.80 %
Standard Deviation (Hz)	89.16	20.12 %	174.67	10.07 %	235.65	9.47 %	177.35	14.31 %

Table 7.19 Mean absolute error and standard deviation for iCAbS LPC 16 condition.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	60.11	12.63 %	207.42	11.65 %	254.48	9.43 %	174.00	11.24 %
Standard Deviation (Hz)	81.87	18.41 %	384.28	19.08 %	417.70	14.80 %	343.18	18.51 %

Table 7.20 Mean absolute error and standard deviation for iCAbS LPC 12, upper comparison frequency 4 kHz condition.

The results from the Default, LPC 8 to 14 and LPC 12 conditions are very similar, both overall and for each formant. Each of these conditions has the number of formants parameter at 4 and the range of the upper cepstral comparison frequency is from 3,000 Hz to 5,000 Hz. The only parameter that is altered across these conditions is the range of the LPC orders that are used. These results show that for the iCAbS tracker, altering the range of the LPC orders and even restricting it to a single value, at least within a

sensible range, does not adversely affect the performance. The results for the LPC 16 condition are not dramatically different from these three conditions and the F1 performance is actually better, although for F2 and F3 it is worse.

Reducing the number of formants setting to 3 produced a marked change in the performance, especially for F2 and F3. This was also seen in the results for the Praat tracker and WaveSurfer. For both WaveSurfer and the iCAbS tracker, reducing the number of formants extracted resulted in the overall mean absolute error doubling, with the greatest increase in the errors occurring for F3. For the iCAbS tracker the results for the LPC 12, 4 kHz upper comparison frequency condition also show a large deviation from the other sets of results. This condition resulted in a significant increase in the errors for both F2 and F3.

The best four conditions for the iCAbS tracker produced results that are comparable in terms of their magnitude with those from WaveSurfer. They are better than the Default and 4 formant conditions for the Praat tracker, but Praat's Optimum condition produced the best results of all the trackers.

In addition to logging iCAbS' measured formant values, a record was kept of the LPC orders, upper cepstral comparison frequencies and mean frame to frame distances that led to the selection of the tracked formants for each vowel token. Table 7.21 shows the median, mode and IQR of the LPC orders for the three conditions in which the LPC order was varied.

Condition	**LPC Order Median**	**LPC Order Mode**	**LPC Order IQR**
Default	13	16	4
3 formants	13	16	5
LPC 8 to 14	12	14	3

Table 7.21 Summary statistics for the LPC orders used by iCAbS to produce measurements in the conditions with variable LPC order.

The Default and 3 formant conditions both had an LPC order range set from 8 to 16. The median and mode are the same for both conditions, whilst the IQRs only differ by one. The distributions of the orders showed them to be relatively uniform with a peak at the highest order. The same distribution was seen for the LPC 8 to 14 condition, but with a reduced range of LPC orders, the median, mode and IQR values are lower.

The upper cepstral comparison frequencies and mean frame to frame distances for all six of the iCAbS tracker measurement conditions were also examined. However, the interpretation of these values was found to be less straightforward as they relate to non-standard measures which are specific to the implementation of the tracker. No clear patterns were seen in the data.

7.3.2.4 Discussion

Overall, the iCAbS tracker performed slightly better than WaveSurfer, which in turn was slightly better than the Praat tracker in the 4 formant condition. The best performer was the Praat tracker in the Optimum condition. This shows that providing the tracker with specific information about the vowel being measured leads to more accurate formant measurements. This improvement in performance could be harnessed when automatically measuring formants in speech samples that have accompanying segmental information.

For all three trackers, the condition in which three formants were measured, rather than four, produced the worst performance. The sensitivity of the performance to this setting is perhaps not obvious and deserves to be highlighted. An analyst who only requires values for the first three formants may simply assume that the 'number of formants' parameter simply controls the number of formant values that are returned by the software. Whilst the setting does serve this function, it also dramatically alters the performance of the three trackers, especially for the third formant. Having the default value of this setting as 3 for the Praat tracker may well lead to poor performance which could be easily avoided if the setting is changed. Since it is the default setting, analysts may well choose not to change it and assume, albeit incorrectly, that it is an appropriate value.

The better performing conditions for WaveSurfer produced results that were comparable with those from Praat's non-tracking Burg tool when the LPC order was 10 across all tokens. The Praat tracker results for the Default and 4 formant conditions were worse. These findings are perhaps surprising as they show that these two formant trackers do not consistently outperform Praat's Burg tool. However, the results from the trackers show that they are much less sensitive to the choice of the LPC order than Praat's standard tool.

7.3.3 Minimum Errors

The general approach used to obtain the measurements summarised above can be likened to an automated unsupervised measurement process, i.e. the analysis parameters are selected and no decision is made by the analyst in relation to the accuracy of the formant tracks. Such an approach is likely to be adopted when making a large number of measurements on segmented speech material. However, formant trackers can also be used by analysts in an interactive way, where parameters are adjusted on a token by token basis until satisfactory formant tracks are achieved.

In the previous chapter a number of analysis frameworks were imposed on the measurements in order to simulate a number of different approaches that a human analyst may take when adjusting the analysis parameters. The same technique is adopted in the following sections for some of the tracker results. However, only one such framework is adopted, which requires the LPC order to remain constant within a vowel token and across the formants ('Tkn Fix, F Fix, Abs' framework in the previous chapter). The framework determines the LPC order at which the minimum errors occur for each vowel token. In this instance the minimum error criterion is the sum of the absolute errors across the formants. Even though this approach was shown in the previous chapter to be most influenced by the errors produced by F3, since these tend to be the largest, this approach was selected since it is more representative of the decision an analysts would make when visually inspecting formant tracks overlaid on a spectrogram with a linear frequency scale. The errors that result from the application of the framework are then summarised using the same statistical measures used previously.

The framework was applied to the results from the three conditions for the Praat tracker and to the results for the Default, 25 ms Hamming and Vowels conditions for WaveSurfer. Since the iCAbS tracker already considers measurements across a number of LPC orders it is not possible to retrospectively apply this framework to those results.

7.3.3.1 Praat Tracker

Table 7.22 to Table 7.24 show the minimum errors obtained across the range of LPC orders for the three conditions tested with the Praat tracker.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	54.27	11.65 %	73.73	5.10 %	116.18	5.40 %	81.39	7.38 %
Standard Deviation (Hz)	74.21	17.51 %	123.40	10.01 %	234.38	12.65 %	159.87	13.90 %

Table 7.22 Mean absolute error and standard deviation for Praat Tracker Default condition with minimum error framework for each token.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	55.28	11.94 %	66.31	4.42 %	89.71	3.91 %	70.43	6.76 %
Standard Deviation (Hz)	76.18	18.22 %	102.67	7.40 %	169.87	8.59 %	123.08	12.72 %

Table 7.23 Mean absolute error and standard deviation for Praat Tracker 4 formant condition with minimum error framework for each token.

	F1		F2		F3		F123	
Mean Absolute Error (Hz)	50.33	10.76 %	61.11	3.98 %	68.61	2.80 %	60.02	5.85 %
Standard Deviation (Hz)	68.00	15.96 %	89.08	5.78 %	100.85	4.22 %	87.40	10.40 %

Table 7.24 Mean absolute error and standard deviation for Praat Tracker Optimum condition with minimum error framework for each token.

Comparison of these figure with the results from the constant LPC order tests presented in Table 7.6 to Table 7.8 for the same conditions show an improvement in performance. The improvement is seen across all three conditions and for all three formants. The largest improvements are seen for the F3 errors, which is to be expected since the minimum error criterion tends to minimise the F3 errors as they are the largest. The biggest improvement in overall performance is for the Default condition. However, in overall terms, the Default condition results are still worse than the results for Praat's non-tracked Burg results with the same minimum error framework applied (see Table 7.5 for comparable results produced with the same alignment with the VTR reference values).

Table 7.25 summarises the LPC orders that were selected in order to obtain the minimum error results presented above.

Condition	LPC Order Median	LPC Order Mode	LPC Order IQR
Default	15	14	6
4 formants	14	11	6
Optimum	16	20	5

Table 7.25 Summary statistics of the LPC orders used to obtain the minimum error values for the Praat tracker across the three conditions tested.

For each condition the full range of LPC orders tested was utilised. The wide spread of LPC orders is reflected by the large IQR values. The median LPC orders are higher than

those selected by the iCAbS tracker, but this most likely due to the restricted range made available to the iCAbS tracker. The results are similar to the LPC orders that produced the minimum errors for the Optimum and Default conditions in the fixed LPC order tests discussed above. The distribution of the orders for all conditions were relatively uniform, apart from the lowest orders which were used infrequently.

7.3.3.2 WaveSurfer

Table 7.26 to Table 7.28 show the minimum errors obtained for three of the better performing conditions for WaveSurfer. These three conditions were selected since they represent the Default analysis parameters, those with the same window length as the other studies discussed (25 ms Hamming condition), and the Vowel condition which is the equivalent of the approaches adopted for the Praat tracker and iCAbS.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	52.94	10.62 %	65.53	4.22 %	85.26	3.49 %	67.91	6.11 %
Standard Deviation (Hz)	74.75	15.55 %	111.75	6.62 %	151.09	6.38 %	116.85	10.43 %

Table 7.26 Mean absolute error and standard deviation for the WaveSurfer Default condition with minimum error framework for each token.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	53.67	10.79 %	68.08	4.39 %	86.22	3.53 %	69.32	6.24 %
Standard Deviation (Hz)	75.66	15.76 %	115.09	6.83 %	150.59	6.37 %	117.90	10.58 %

Table 7.27 Mean absolute error and standard deviation for the WaveSurfer 25ms Hamming condition with minimum error framework for each token.

	F1		**F2**		**F3**		**F123**	
Mean Absolute Error (Hz)	53.21	10.74 %	67.20	4.34 %	85.14	3.50 %	68.52	6.19 %
Standard Deviation (Hz)	74.92	15.85 %	111.88	6.74 %	150.07	6.44 %	116.46	10.63 %

Table 7.28 Mean absolute error and standard deviation for the WaveSurfer Vowels condition with minimum error framework for each token.

The application of the minimum error framework to the WaveSurfer results again causes a reduction in the errors with the greatest improvement being for F3. The resulting mean absolute error values for each formant are more similar than for the fixed LPC order results. The summary statistic of the LPC orders that obtained the minimum orders for each condition was identical. The median LPC was 15, the mode was 12 and the IQR was 5. For each condition all of the LPC orders from 12 to 20 were used.

7.3.3.3 Discussion

The results presented above show that allowing the LPC order to vary between vowel tokens reduces the overall magnitude of the errors relative to the situation where the LPC order is the same across all tokens. This reflects the findings in the previous chapter (Section 6.3.2). These results suggest that a human analyst who is adjusting the analysis parameters on a token by token basis can produce more accurate measurements than using the same LPC order across all tokens. The overall performance is better if an analyst modifies the LPC order when using a formant tracker. However, even the results obtained from applying the minimum error frameworks to the measurements from Praat's non-tracking Burg tool are better than those produced by the trackers with a fixed LPC order.

The iCAbS tracker can generate candidate formants across a range of LPC orders and applies an objective criterion, which is based on the signal, to select the best formant values. This can be seen as an approach which is similar to that used by the minimum error framework. However, the results from the application of the minimum error framework outperform the iCAbS tracker. However, the minimum error framework does utilise the VTR reference values in order to determine the minimum errors. The same performance may not be achieved by a human analyst who would be visually comparing the formant tracks with a spectrogram.

7.3.4 Variation of Errors Across the Vowel Space

The following sections examine the behaviour of the measurement errors from the three trackers over the vowel space. The same approach used in the previous chapter was applied to the error values from the trackers, i.e. error surface plots were created over the F1~F2 and F2~F3 vowel spaces. The results from the Praat tracker and WaveSurfer are considered for both the constant LPC order analyses and for the minimum error framework. The error surfaces were generated for all of the Praat tracker conditions, but only for the better performing conditions for WaveSurfer and iCAbS.

The behaviour of the LPC order across the vowel space is also examined for the minimum error frameworks for the Praat and WaveSurfer, and for the iCAbS results.

7.3.4.1 Praat Tracker

The F1 mean error surfaces for the three Praat tracker conditions for LPC order 10 and above are very similar in structure to that shown in Figure 6.7 in the previous chapter. The structure also remains relatively stable over this range of LPC orders. The overestimates, with positive errors, occur at the lower F1 values, whilst the underestimated measurements, with negative errors, are at the higher F1 values. The same structure is seen across the minimum error framework conditions but the magnitude of the errors at the limits of the F1 range is smaller and the surfaces are at a shallower angle, indicating that there is less variation over the surface in the F1 direction.

The structure of the F2 mean error surfaces for the Default condition and the 4 formant condition are somewhat different to that shown in Figure 6.8 in the previous chapter. The F2 error surface for the Default condition at LPC order 14, which produced the overall minimum error, is shown in Figure 7.5. The central region of the surface is relatively flat, whereas at the lowest and highest F2 values the magnitude of the errors is very large. There is also no apparent dependency in the surface on the F1 value.

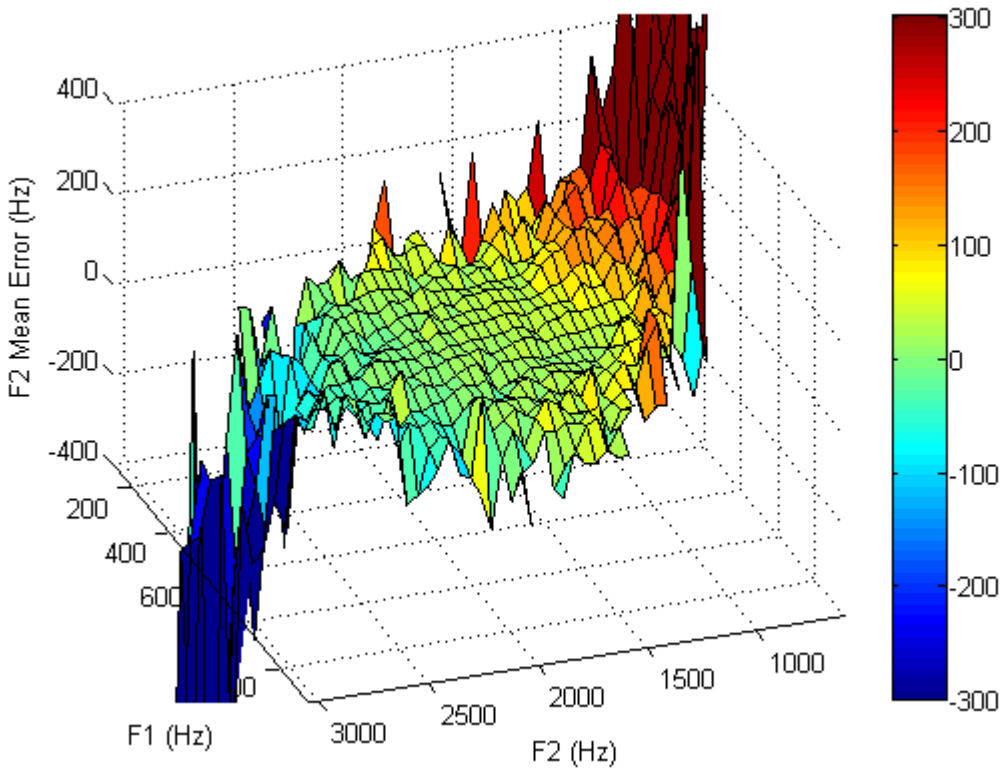

Figure 7.5 F2 mean error surface over the F1~F2 vowel space for the Praat tracker Default condition at LPC order 14.

The F2 error surface for the 4 formant condition at LPC order 11, which produced the minimum overall error, is very similar to that shown in Figure 7.5. For both the Default and 4 formant conditions as the LPC order increases the structure changes and the surface becomes very steep, with the highest positive errors at the lower F2 values and the lowest negative errors at the higher F2 values.

For the Optimum condition, above LPC order 10 the error surface is relatively flat and shows very little variation in either the F1 or F2 direction. Comparison of this error surface with those for the Default and 4 formant conditions show that altering the tracker's reference values from the default settings prevents the very large errors occurring at the extremes of the F2 range. In the non-Optimum conditions, at the extremes of the F2 range, the tracker is tending to gravitate towards candidate values which are closer to the fixed reference values rather than the correct ones. Figure 7.6 shows a typical example of such an error, in this instance for F3.

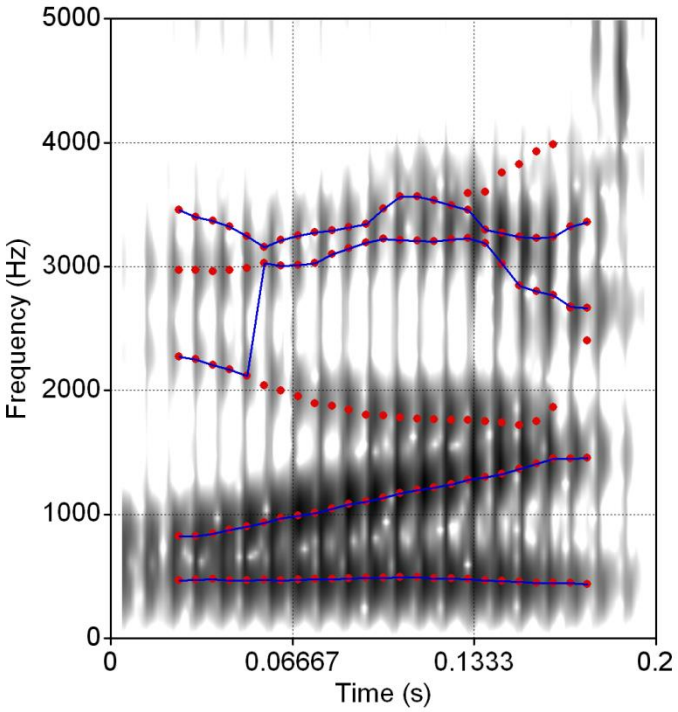

Figure 7.6 Spectrogram of vowel sequence /ɔɚ/ in the word 'towards' from file 'SI1154.WAV' spoken by 'mcdr0', with overlaid candidate formant values as red dots produced by an LPC analysis at order 12. The formant tracks produced by Praat's tracker with the 4 formant condition settings are shown as blue lines. A tracking error has occurred from the sixth analysis frame onwards for F3.

At the sixth analysis frame the track for the third formant shifts from the candidate values that follow the true F3 values as shown by the spectrogram, up to those that are

associated with the fourth formant. Across the incorrectly tracked section the candidate values associated with F4 are closer to the reference value of 2,500 Hz than the competing candidate values aligned with the true F3. This has caused the track to jump to the higher candidates. If the candidate values are accepted using the numbering approach of Praat's normal measuring tool then then measurements would have been acceptable measurements that would not have included such a large error in F3.

For the minimum error framework results, the F2 error surface for the Optimum condition is very similar to the constant LPC order situation. For the Default condition, the minimum error surface does not display the region of large negative errors at the higher F2 values and the region of large positive errors at the lower F2 values is smaller. The surface for the 4 formants condition is similar, but the magnitude of the errors in the lower F2 region is even smaller.

The F3 error surfaces for the Default and 4 formant conditions across the F1~F2 vowel space show a distinct region in the lower F1~F2 area of the vowel space where the errors are large and positive. This is in contrast to the rest of the surface which is relatively stable. Figure 7.7 shows the F3 errors across the F2~F3 vowel space, which reveals a similar pattern.

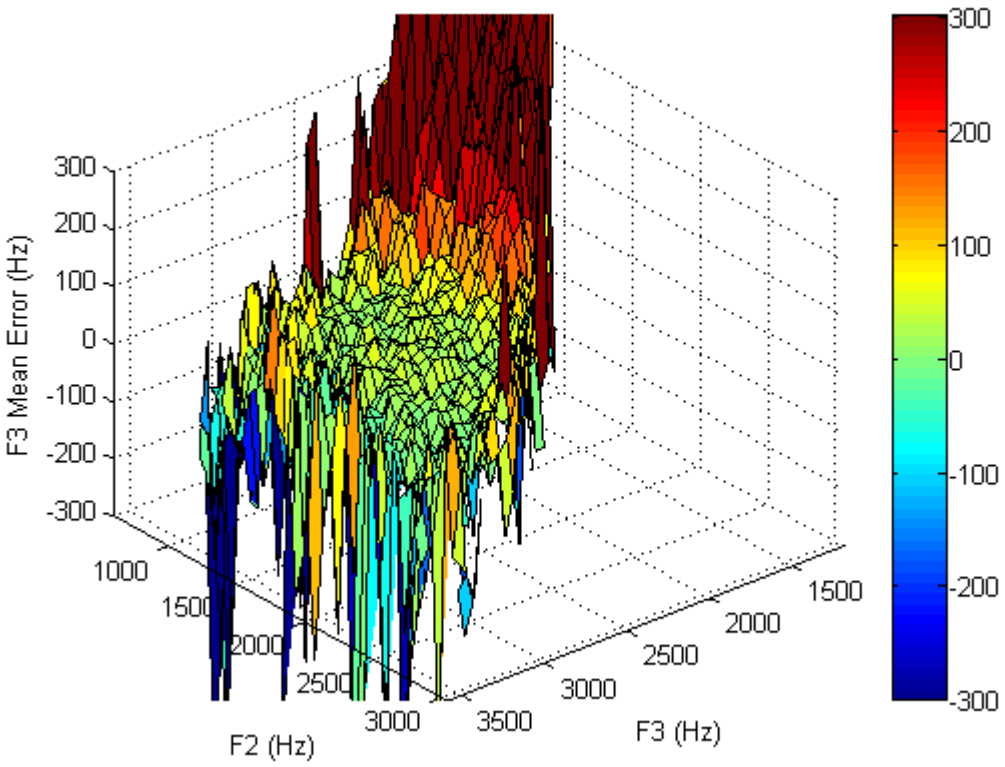

Figure 7.7 F3 mean error surface over the F2~F3 vowel space for Praat tracker 4 formant condition at LPC order 11.

For both conditions, as the LPC order increases, the region of large positive errors spreads towards the centre of the surface and a similar region containing large negative errors appears at the higher F3 values. The minimum error framework results for both conditions have the same structure as seen in Figure 7.7, but the magnitude of the largest errors is reduced.

The F3 error surfaces for the Optimum condition are again relatively flat and stable above an LPC order of 10. No regions of large errors are present. This means that the same effect described above for F2 also occurs for F3, i.e. at the extremes of the F3 values the tracker tends to favour candidate values towards the centre of the range, resulting in large errors at the extremes. Providing the tracker with information about the true location of the formants removes this effect.

7.3.4.2 WaveSurfer

The F1 error surfaces for the WaveSurfer results across all of the three conditions examined are again very similar to the F1 error surfaces described above and in the previous chapter. Minimal variation is seen across the surfaces as the LPC order

changes. Again, the same structures are seen for the minimum error framework results, with somewhat smaller errors.

At LPC order 12 the F2 error surfaces across the conditions are relatively flat and show little variation. As the LPC order increases to 13 and above, a region of large negative errors appears in the higher F2 region. However, unlike the Praat tracker F2 error surfaces, no region of large positive errors occurs in the lower F2 region. The F2 error surfaces for the minimum error framework results are also relatively flat.

The F3 error surfaces over the F2~F3 vowel space show tendency for positive errors, i.e. overestimates, in the lower F3 region of the space, and negative errors in the higher F3 region. This can be seen in Figure 7.8 for the Default condition at LPC order 13.

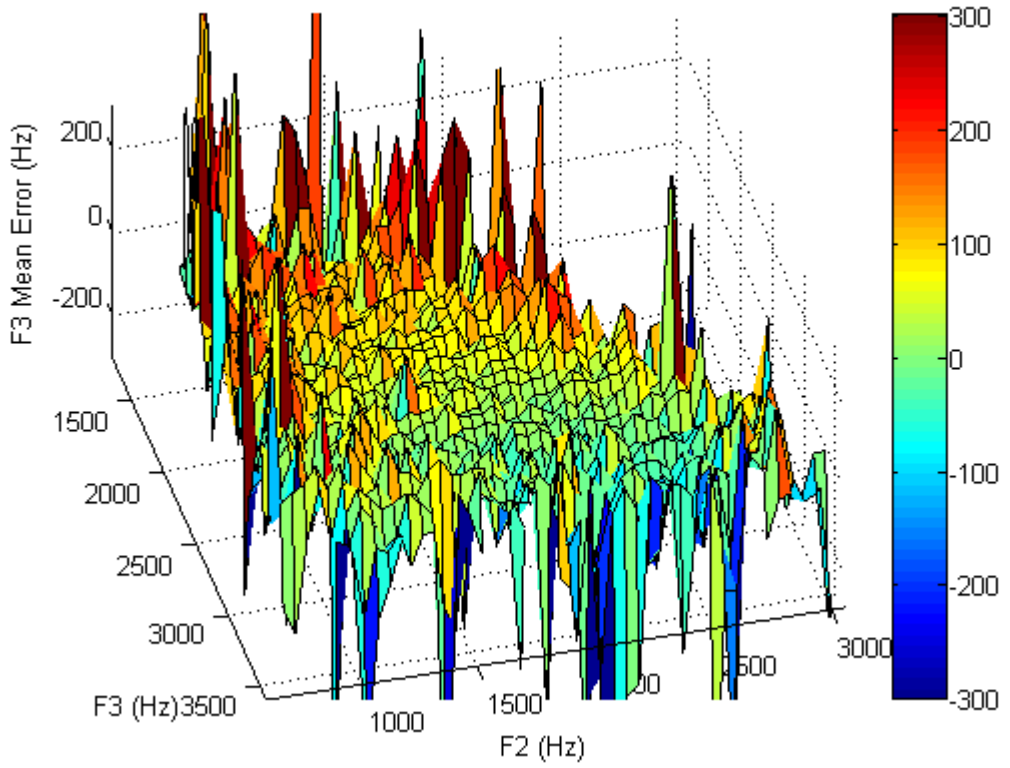

Figure 7.8 F3 mean error surface over the F2~F3 vowel space for WaveSurfer Default condition at LPC order 13.

From LPC order 14 and higher in the region of the highest F3 values large negative errors occur. Unlike the F3 errors for the Default and 4 formant conditions for the Praat tracker, the behaviour of the errors at the lower F3 values does not significantly change. The F3 error surfaces for the minimum error framework results are much more uniform and do not show the variation across the F3 values.

7.3.4.3 iCAbS

For the iCAbS tracker the F1 errors across the vowel space are again very similar to those seen already. Across the four conditions examined there is very little to distinguish them. However, the F2 error surfaces are somewhat different to those described previously. Figure 7.9 shows the F2 error surface for the iCAbS Default condition. For each of the conditions positive errors tend to occur at the lowest F1 values for central F2 values. Another obvious feature for all conditions is a small dip towards the right hand edge of the surface at the lower F2 values. One feature that is only present for the F2 mean error surface for the LPC 16 condition is a region of large negative errors across the left hand edge of the surface with the highest F2 values.

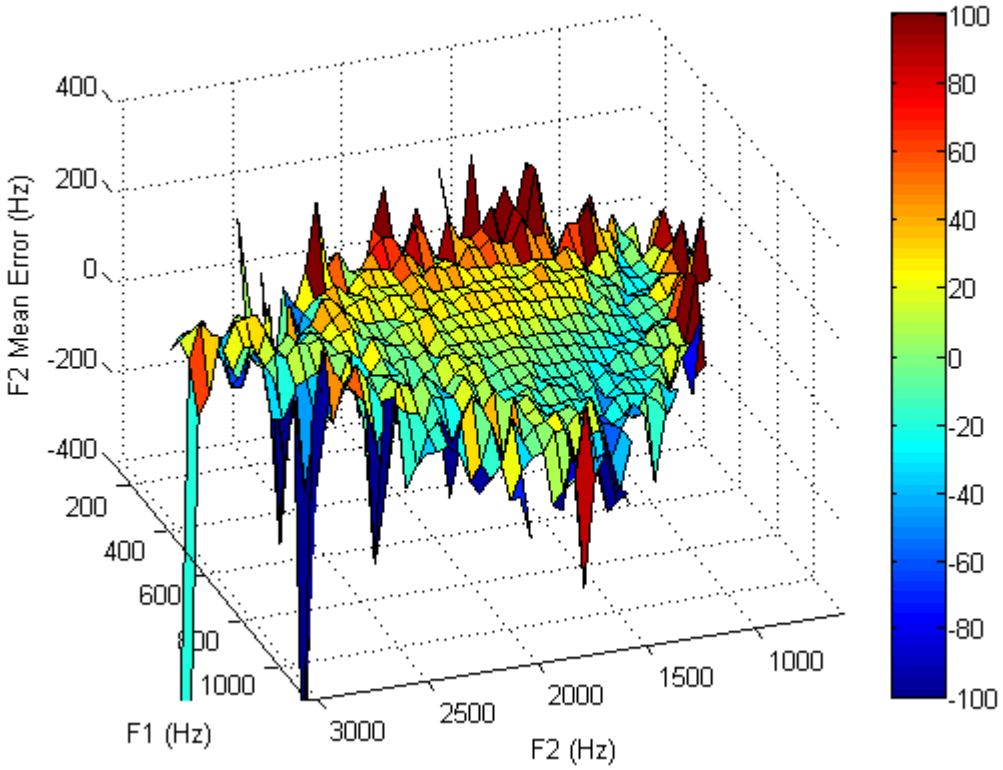

Figure 7.9 F2 mean error surface over the F1~F2 vowel space for iCAbS Default condition.

The F3 errors for all four conditions are relatively stable across the majority of the F1~F2 surface with some larger errors occurring at the edges of the surfaces. A similar pattern is seen for the F3 errors when considered over the F2~F3 vowel space, which also contains isolated pockets of positive and negative errors.

7.3.4.4 LPC Order Variation

Plots were generated to show how the median LPC order varied across the vowel space for the minimum error tracker conditions. Since the minimum error criterion required that the LPC order was the same across the three formants within a vowel token, only one set of values per tracker condition existed. The same plots were also generated for the iCAbS tracker results where the LPC order was varied.

Figure 7.10 shows the distribution over the F1~F2 vowel space of the median LPC order for the Praat tracker 4 formant condition. The distribution for the Default condition was very similar. The plot shows a tendency for the higher LPC orders to occur towards the lower F1 values and it also exhibits some dependence on F2 within that region. Figure 7.11 show the median LPC order distribution for the Optimum condition, which in comparison shows less F2 dependence at the lower F1 values. The LPC orders are generally higher, which is to be expected given the results in Table 7.25 that show a higher overall median LPC order for the Optimum condition when locating the minimum errors.

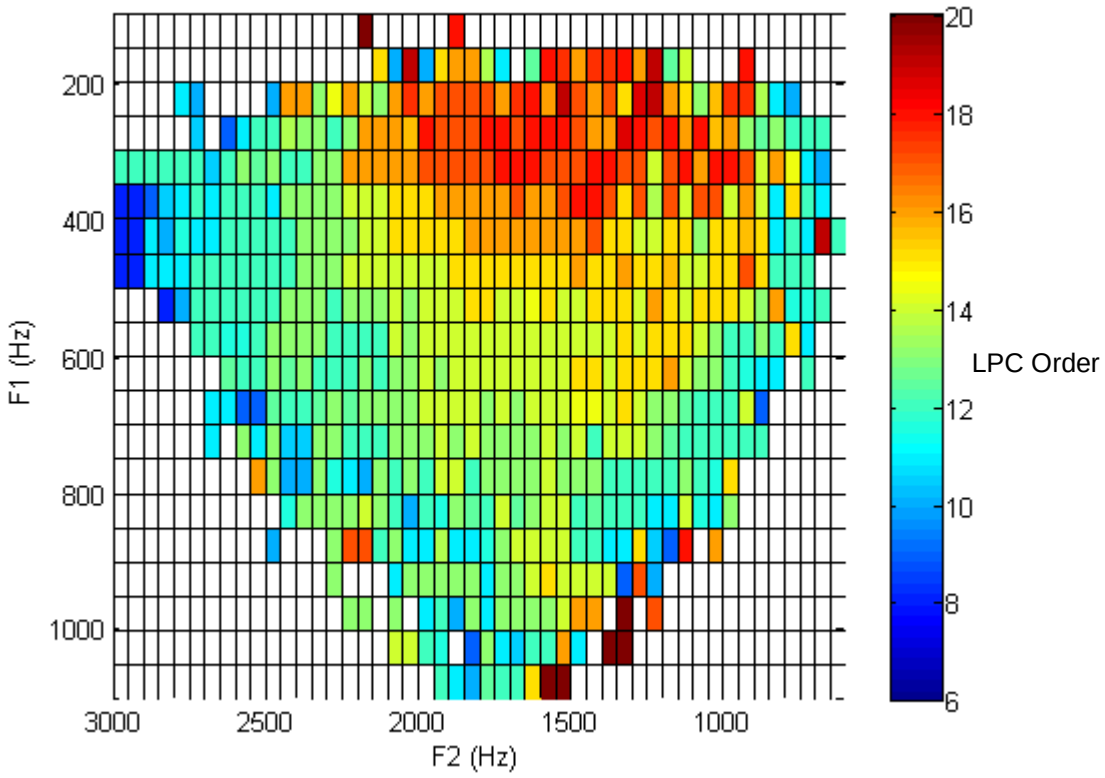

Figure 7.10 Median LPC order usage across the F1~F2 vowel space for Praat tracker 4 formant condition minimum errors.

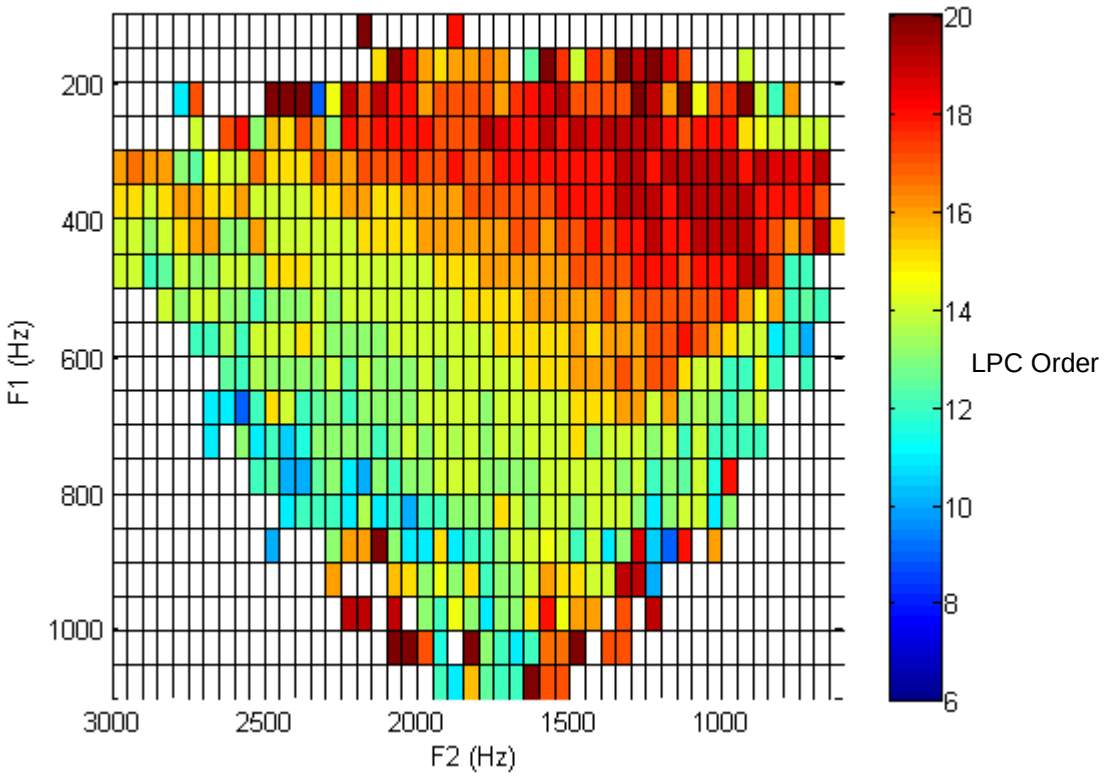

Figure 7.11 Median LPC order usage across the F1~F2 vowel space for Praat tracker Optimum condition minimum errors.

The plots generated for WaveSurfer over the F1~F2 vowel space showed distributions similar to that shown above for the Optimum Praat tracker condition, with similar LPC order magnitudes. The distributions for the iCAbS tracker were much more uniform, showing no apparent dependency on the location in the vowel space. Also, the median LPC orders were lower, but since the upper LPC order was 14 or 16, compared with 20 for Praat and WaveSurfer, this is not surprising.

7.3.4.5 Summary of Errors Across the Vowel Space

The behaviour of the F1 errors over the vowel space for the three trackers is very similar and reflects the patterns seen in the previous chapter for the non-tracked results. The variation in the behaviour of the F2 and F3 errors for the Praat tracker and WaveSurfer clearly demonstrates the impact of the tracking decision being partly determined by a set of reference values. The behaviour of the F2 and F3 errors for iCAbS are different to those from the Praat tracker and WaveSurfer. The behaviour seen cannot be clearly linked to the tracking methodology employed.

The application of the minimum error framework to the results showed a reduction in the degree of variation seen in the error surfaces. Examination of the LPC orders that

gave rise to the minimum errors revealed some dependence on the location within the vowel space. However, this dependence was not seen in the LPC orders that were selected by the iCAbS tracker.

7.3.5 Variation of Performance Across Speakers

The analyses described in the following sections concern the performance of the speakers across the three trackers. A subset of the results is examined using the same methods applied in the previous chapter to the data from Praat's Burg tool. The subset of results consists of the 3 conditions for the Praat tracker, the Default, 25ms Hamming and Vowels conditions for WaveSurfer, and the Default, LPC 8 to 14, LPC 12 and LPC 16 conditions for iCAbS. The results for Praat and WaveSurfer are considered only at the LPC orders that resulted in the smallest combined errors as shown in the tables above. The subset also includes the minimum error conditions for these Praat and WaveSurfer conditions. All of the analyses are based on the mean absolute error for each speaker. Only the numeric errors, rather than the percentage equivalents, are considered in these analyses.

7.3.5.1 Analysis of Mean Speaker Errors Across Trackers

The results were first examined by determining the mean and the standard deviation of the speakers' mean absolute errors across the three formants. These values are shown in Figure 7.12 for the subset of the normal tracker conditions and in Figure 7.13 for the minimum error conditions. The mean values are represented by circles, whilst the standard deviations are shown by the vertical error bars. The colours red, green and blue are used for F1, F2 and F3 respectively.

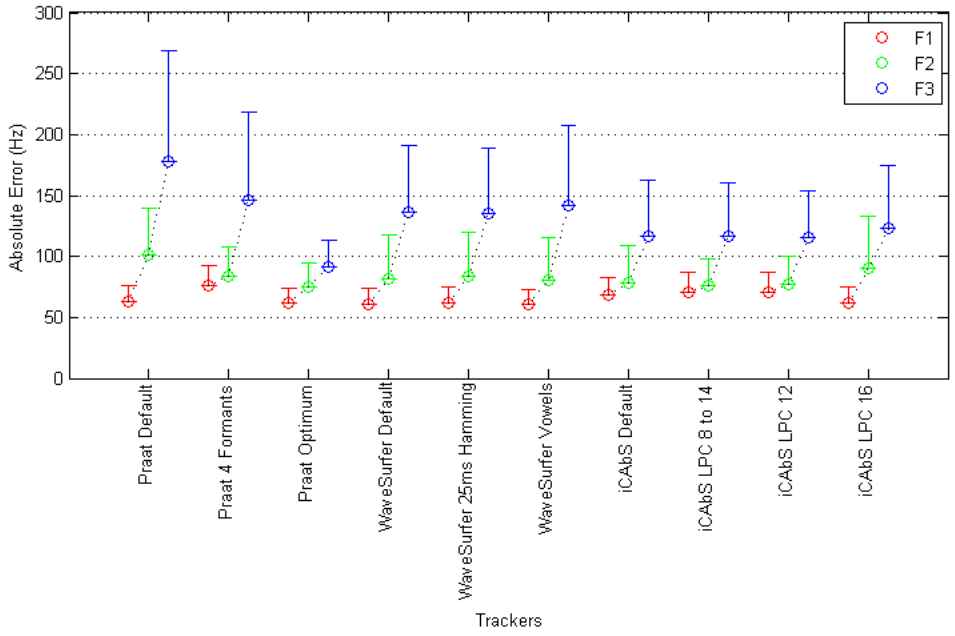

Figure 7.12 Mean and standard deviation of the mean absolute errors from 186 speakers for each tracker condition for F1 (red), F2 (green) and F3 (blue).

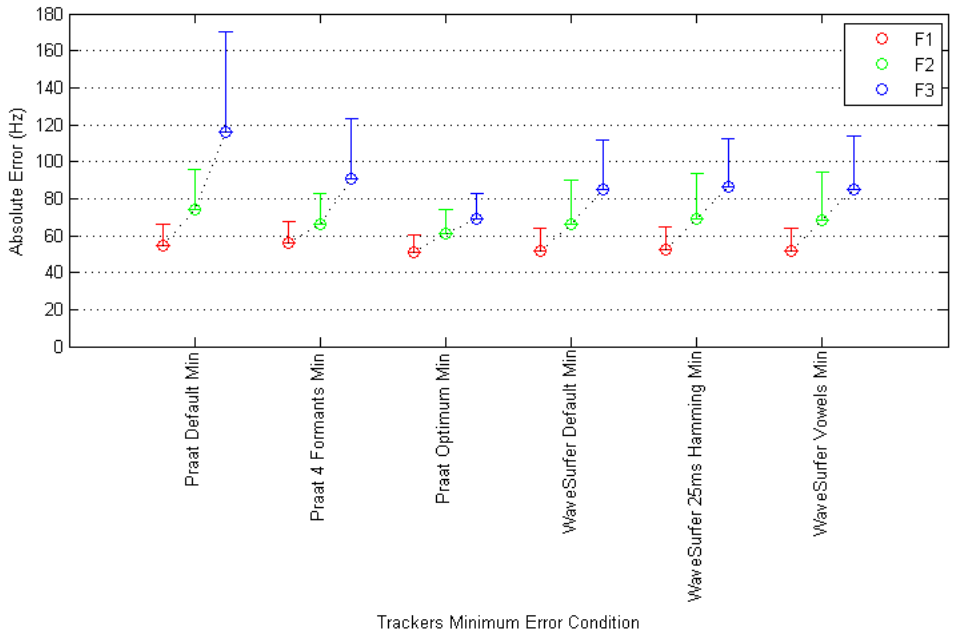

Figure 7.13 Mean and standard deviation of the mean absolute errors for 186 speakers for each tracker condition with the application of the minimum error frameworks for F1 (red), F2 (green) and F3 (blue).

As expected, the mean values in Figure 7.12 and Figure 7.13 are very similar to those in the earlier tables for the entire set of errors across all frames for each tracker condition. The standard deviations of the speakers' absolute mean errors are much lower than the

standard deviations for the entire set of errors. In Figure 7.12 the improvement in performance across the three Praat tracker conditions from the Default to the Optimum is clear, albeit that the F1 performance for the 4 Formants condition is the worst of the three. The WaveSurfer conditions have very similar results, which is also the case for the first three iCAbS conditions, with the fourth showing performance that is better for F1, but worse for F2 and F3.

The results in Figure 7.13 reflect the earlier findings that the results from the minimum error conditions are better than those from the constant LPC order conditions. For Praat, the minimum error conditions show the same improvement across the three conditions seen in Figure 7.12. The results for WaveSurfer are again very similar across the condition.

7.3.5.2 Analysis of Mean Errors Within & Across Trackers

In order to examine the mean absolute errors for individual speakers, a series of plots, similar to that at Figure 6.19 in the previous chapter, were generated. These showed the mean absolute errors for each formant for each speaker as a stacked bar chart, ordered according to increasing combined error across the three formants. Whilst the plots showed the magnitude and range of errors encountered for speakers across the different trackers and conditions, they did not reveal any systematic differences when compared with the equivalent results from Praat's Burg tool. Like the results from the previous chapter, the plots from the tracker data did not show any obvious relationships between the magnitudes of the errors across the formants for individual speakers. However, this point is addressed in more detail in the following section.

7.3.5.3 Relationship Of Errors Across Formants

To further examine the relationships across the formants for speakers' mean absolute errors, a series of scatter plots were generated to compare F1 with F2, F1 with F3 and F2 with F3. The plots were similar to those shown in Figure 6.20 to Figure 6.22 in the previous chapter and showed similar amounts of dispersion in the data. Again, there were some plots that showed a weak tendency towards positive correlations. These were generally for the F2 vs F3 comparisons, but several of the F1 vs F2 plots also showed similar patterning. To summarise these relationships Pearson's r correlation coefficients were calculated for each of the formant pairs across the formant conditions. These showed that the relationships between the formants are, in general, greatest for F2 and

F3, followed by F1 and F2. The size of the correlation coefficients is comparable with those from the previous chapter. This suggests that overall the use of a formant tracker does not increase the level of dependency in the errors across formants. Again, the majority of the correlations were significant, but the relatively weak relationships are apparent from the magnitude of the correlation coefficients and the amount of dispersion seen in the scatter plots.

7.3.5.4 Relationship of Errors Across Trackers

The final section examining the error results from speakers considers their behaviour across the different tracker conditions. The results from the previous chapter, presented in Section 6.3.4.8, showed that the speakers behaved in a very similar manner across the analysis frameworks, i.e. those that had large errors for one framework also tended to do so for the others. To determine if this also applied to the tracker results a series of scatter plots were generated to compare the mean absolute measurement errors from all speakers across all possible combinations of tracker conditions. This was done for each formant individually as well as for the combined error summed across all three formants.

The scatter plots all showed a positive correlation, but a much wider range of dispersions existed across the tracker conditions in comparison with those seen across the minimum error frameworks in the previous chapter. In general, the strongest correlations existed between the conditions for the same tracker, whilst the weakest correlations were between the conditions for different trackers. Again, to summarise the results Pearson's r correlation coefficients were calculated for all combinations. All of the correlations were significant at the 0.01 level (two tailed).

Both the scatter plots and the correlation coefficients show that the performance for individual speakers varies both within and across the different formant trackers. This means that while some speakers will perform well for a particular tracker they may well perform badly for another, or even for a different analysis condition with the same tracker. The comparisons of the Praat Default and 4 formant conditions with those from the other trackers shows relatively small coefficients around 0.3, whilst for most of the WaveSurfer and iCAbS comparisons the coefficients are around 0.7. The correlation coefficients for the errors from the individual formants showed somewhat different patterns especially for F1.

7.3.5.5 Variation Across Speaker Parameters

In Section 6.3.4 in the previous chapter speakers' mean errors were considered against the parameters of speaker sex, fundamental frequency and vowel space location, in order to determine if these factors were related to performance. The findings were that only weak correlations existed between performance and the parameters examined. In view of those findings and the variation in performance for speakers seen across the different tracker conditions, it was considered unlikely that the tracker results would yield results that were substantially different those in the previous chapter. Therefore, these comparisons were not undertaken.

7.3.5.6 LPC Order Variation Across Speakers

The speaker results were also examined in terms of the LPC orders that were selected when the minimum error framework was applied to the WaveSurfer and Praat tracker results, and those that were selected by iCAbS. Following the same approach described in the previous chapter (Section 6.3.4.13), the median, mode, minimum and maximum LPC orders were determined for each speaker for each tracker condition. The summary values were displayed in plots like the one shown at Figure 7.14, which shows the results for the 4 formants condition for the Praat tracker. Examining the plots for the different tracker conditions revealed that for most speakers the minimum errors were obtained across a wide range of LPC orders.

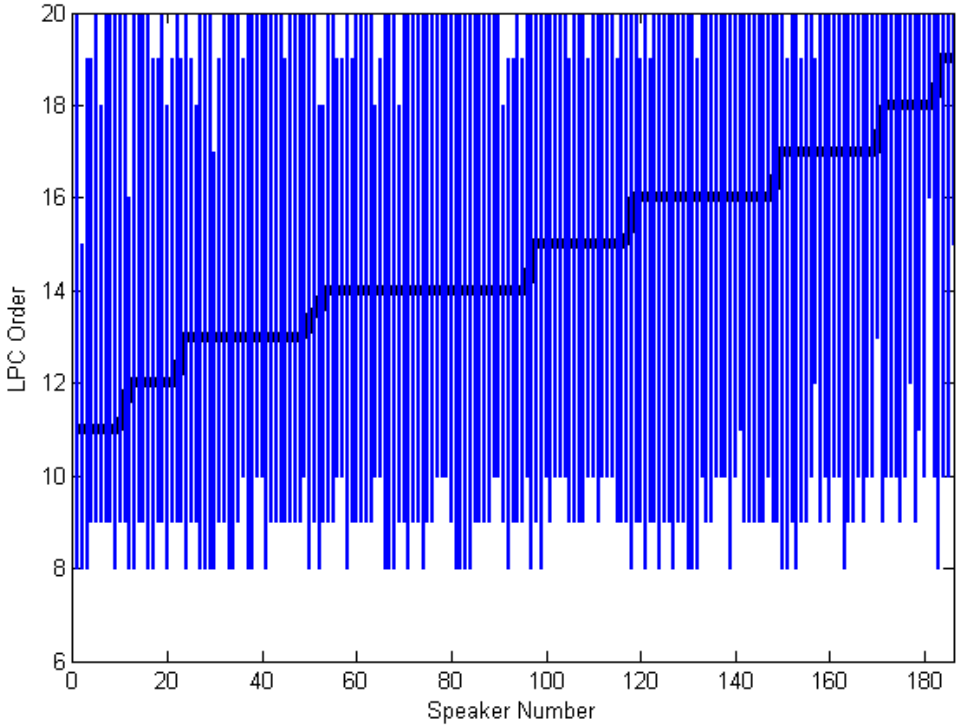

Figure 7.14 Plot of median LPC order (thick horizontal line) and range (thin vertical line) for all speakers ordered by increasing median value and range. The results are from the Praat tracker with the 4 formant analysis parameters and the minimum error framework.

Comparison of the results across the three trackers show that the median LPC orders across the Praat tracker and WaveSurfer were similar. The differences that were seen in the range results are a consequence of the ranges of the LPC order tested for the two trackers being different. The mean LPC order results for iCAbS are different from the other two, but it is not apparent if this is a consequence of a more limited range of LPC order being tested or whether it is a result of a different tracking approach.

7.4 Results from Other Studies Using the VTR Database

A number of other studies have used the VTR database to test the performance of formant trackers, often ones that employ novel approaches to the formant tracking problem. In principle, it is useful to compare the results from the Praat tracker, WaveSurfer and iCAbS with those from these other studies as it provides further information about their performance relative to other trackers. Such comparison would shed further light on RQ1. However, the comparison of the results is problematic. Some of the studies, such as Gläser et al. (2007), Gläser et al. (2010) and Rudoy et al. (2007), do not present their results in a form that are comparable with those presented above. These studies only provide summary statistics for the measurement errors across all

speech segments rather than breaking them down by segment type, i.e. there are no results provided for just the vowel segments. Also, they use non-comparable measures of the errors. Rudoy et al. (2007) have reported root mean square error reduction relative to their results from WaveSurfer, whilst Gläser et al. (2007) have employed an error measure based on formant specific thresholds. In Gläser et al. (2010) the results from their tracking method are presented as mean relative percentage improvements compared to the results from Praat, WaveSurfer and Mustafa and Bruce's (2006) tracking method.

One feature common to most of the studies is that in addition to testing a new formant tracking technique, which is the main focus of each of the articles, they also report the results from tests with WaveSurfer. This is done so that the performance of the new techniques can be compared with a benchmark tool that has 'wide use among voice and speech researchers' (Mehta et al., 2012, p. 1737). In most of the studies the reported performance of WaveSurfer has been determined by the authors. However, in the case of Özbek and Demirekler (2008) they simply quote the performance figures for WaveSurfer and the MSR algorithm provided by Deng et al. (2006). The inclusion of results from WaveSurfer in these studies also highlights a significant issue. In each of the reported studies, and in the tests undertaken for this thesis, the performance for WaveSurfer is different. This means that there must be differences between the implementation of the testing processes across the studies. Some of these differences may be in the analysis parameters that have been selected, but it is highly likely that some of them relate to other aspects of the testing procedure, which will most probably have also been applied to the testing of the new tracking methods. The consequence of this is that the results from the different studies cannot be considered as directly comparable since they have not been tested in identical ways.

7.4.1 Results From Deng et al. (2006) & Smit et al. (2012)

Table 7.29 shows the results presented in Deng et al. (2006, p. 370, Table 1) for WaveSurfer and their MSR tracker, as well as the results from Smit et al. (2012, p. 899, Table 3) for WaveSurfer, Praat and their novel approach based on spectral peak picking and contour integration. The values are mean absolute errors across frames for the vowel segments only.

Study	Tracker	F1 (Hz)	F2 (Hz)	F3 (Hz)
Deng et al 2006	WaveSurfer	70	94	154
	MSR	64	105	125
Smit et al 2012	WaveSurfer	53	84	172
	Praat	90	116	167
	wGDF-CI	57	86	131
	GDF-CI	57	87	133

Table 7.29 Mean absolute error values expressed in Hertz for measurements of vowel frames referenced to the VTR database reported in Deng et al. (2006, p. 370, Table 1) and Smit et al. (2012, p. 899, Table 3) obtained from different formant trackers.

Comparison of the WaveSurfer results across the two studies shows that the F1 and F2 performance was better for Smit et al. (2012), whilst the F3 performance was better for Deng et al. (2006). In percentage terms the differences are 32 %, 12 % and 12 % for F1 to F3 respectively. These error values are also different to those reported above for WaveSurfer. The results from the current study of WaveSurfer presented above have mean absolute error values ranging from 58 to 68 Hz for F1, from 79 to 90 Hz for F2 and from 132 to 144 Hz for F3, ignoring the 3 formants condition (see Table 7.9 to Table 7.14). These results are better than those presented by Deng et al. (2006) for all three formants. Compared with Smit et al (2012) they are better for F3, similar for F2 and worse for F1.

Contact was again made with the authors of the VTR database in order to discover if their methodology was significantly different from that used in the present study. It was confirmed that their analysis settings remained constant across all utterances, that the formants were tracked across entire files, that segments were determined according to the TIMIT segmentation information, and that the measurements were made with WaveSurfer rather than the Snack toolkit (Deng, 2013, Cui, 2013). However, there are two differences between the set of VTR reference values used in Deng et al. (2006) and those that are in the publically available database. Firstly, the Deng et al. (2006) results relate to measurements taken from 538 sentences, not the 518 that are provided in the VTR database, and secondly, they were calculated from the first pass correction of the VTR values, not the second pass ones in the released version of the database. A warning is provided in the user manual that accompanies the database that the use of the second pass values may lead to different results. However, given the differences found between the results from the current study and those reported elsewhere, it is not apparent how

much of the disparity with the Deng et al. (2006) results can be attributed to these issues.

Closer examination of the description of the methodology in Smit et al. (2012) reveals that certain relevant information is not provided and there are a number significant differences in the approach used. No information is provided about the analysis settings used for either Praat or WaveSurfer, and the raw results from Praat and the other trackers being tested were post-processed with a three point running median filter. The filter was applied in order to make the results more comparable with those from WaveSurfer, since they had been subject to a tracking procedure. Furthermore, they list the total number of vowel frames analysed as 61,238 (Smit et al. 2012, p. 899, Table 2) compared with the 67,242 frames analysed in this thesis; a difference of 6,004. Since there are a total of 6,601 vowel tokens, the difference in frames is equivalent to just less than 1 frame per token. Given the number of known and potential differences between the approaches it is difficult to assess their impact on the results.

7.4.2 Results From Mehta et al. (2012)

One study which provides comparably more information in relation to the analysis parameters used is Mehta et al. (2012). Again, both the Praat tracker and WaveSurfer are used as benchmark conditions for comparison with the performance of a KARMA (Kalman autoregressive moving average) formant tracker. Some of the analysis parameters for Praat and WaveSurfer were matched to those of the KARMA algorithm, presumably in an attempt to provide comparability between the results. The sampling frequency was 7 kHz, giving an upper analysis frequency of 3.5 kHz, the analysis frames were 20 ms in duration, with a Hamming window and 50 % overlap, i.e. a frame advance of 10 ms, and the LPC order was 12. The default tracking, or 'smoothing' settings as they are referred to in the text, were used for WaveSurfer and Praat. Again, these settings do not completely match those used in the current study. Also, the performance is expressed as root mean square error (RMSE), rather than mean absolute error. To assess the similarity of the results for Praat and WaveSurfer from Mehta et al. (2012) with those from the tests described above, some of the error summary values were recalculated as root mean square errors. This was done for the default condition of WaveSurfer at LPC order 12, for the default condition of the Praat tracker at LPC order 12 and for the 4 formants condition for the Praat tracker, again at LPC order 12. These conditions were chosen as they were the most similar to those tested by Mehta et al.

(2012). These RMSE values, together with the results for the vowel segments reported in Mehta et al. (2012, p.1738, Table IV), are shown in Table 7.30.

Study	**Tracker**	**F1 RMSE (Hz)**	**F2 RMSE (Hz)**	**F3 RMSE (Hz)**
Mehta et al. 2012	KARMA	82	258	336
	WaveSurfer	112	254	262
	Praat	134	269	341
Current	WaveSurfer Default LPC order 12	85.56	145.35	292.16
	Praat Default LPC order 12	95.72	173.28	380.48
	Praat 4 formants LPC order 12	95.47	160.47	301.91

Table 7.30 Root mean square error values expressed in Hertz for vowel frames from the VTR database reported in Mehta et al. (2012, p.1738, Table IV) and WaveSurfer default condition, Praat tracker default condition and 4 formants condition at LPC order 12 from the current study.

These results show a clear difference in performance for both Praat and WaveSurfer across the two sets of results. For WaveSurfer, the F1 and F2 RMSE values are smaller in the current study but the Mehta et al. (2012) result for F3 is better. The situation is the same between the default condition of the Praat results and Mehta et al (2012), but for the 4 formant condition the current study's results are better for all three formants.

The differences in the results highlights a potential issue that adjusting the analysis parameters of the benchmarking tools, i.e. WaveSurfer and Praat, to those of the tracker being tested, may make the benchmark performance worse. During the development of the tracker under test it is likely that the analysis parameter values will be optimised to ensure that it performs at its best. However, it would appear that in the studies considered above that no such optimisation is applied to the benchmarking tools. Adjusting the analysis parameters so that they reflect those of the system being compared may actually decrease the performance of the tool being used as the benchmark. The difference in the Praat tracker results for F3 between the default and the 4 formant conditions clearly demonstrate the effect that altering a single parameter can have.

7.4.3 Results From García Laínez et al. (2012) & González et al. (2012)

Two further studies which employ the VTR database to test a novel tracking technique, and use WaveSurfer as a benchmark, are García Laínez et al. (2012) and González et al. (2012). Both test the same tracking method, based on a beam-search algorithm, with the first providing a detailed description of the technique and test results for different configurations of the tool, whilst the second considers its performance for degraded speech material. These studies are of note for two reasons. Firstly, the overall performance, for both the new algorithm and WaveSurfer, is markedly better than any of the other studies discussed above. Secondly, even though the limited description of the testing procedures used for the two studies appear to be identical, the results reported for WaveSurfer are different in each one. However, identical results are reported in the studies for one of the configurations of the new tracker being tested.

The apparent difference in performance of WaveSurfer between the two studies could simply be a typographical error or it may be a consequence of the tests having been re-run differently for the second study. If this were the case then this further demonstrates the sensitivity of the results to changes in the testing process, albeit in this case unknown changes. The results are not directly comparable with any of the other studies since another different methodological approach has been applied. Rather than use all of the 518 sentences in the database, only 420 were analysed. Also, the errors are calculated for all voiced frames, rather than all speech frames or vowel frames only. In García Laínez et al. (2012) some of the analysis parameters are provided; the sample rate was 10 kHz, the pre-emphasis factor was 0.7, the LPC analysis was autocorrelation, the frame duration was 49 ms with a Hamming window, the frame advance was 10 ms and the LPC order was 12. The results from the two studies for WaveSurfer and one of the test algorithm conditions (designated 'Quad+Mp') are shown in Table 7.31.

Tracker	F1 (Hz)	F2 (Hz)	F3 (Hz)
WaveSurfer García Laínez et al. (2012)	18.46	30.84	46.33
WaveSurfer González et al. (2012)	29.95	57.66	76.53
Beam-search 'Quad+Mp'	18.39	27.96	35.26

Table 7.31 Mean average error values for WaveSurfer and the beam-search tracking algorithm (condition 'Quad+Mp') for vowels in the VTR databased reported in García Laínez et al. (2012, p. 754, Table 1) and González et al. (2012, p. 44, Table 1).

Comparison of the results from WaveSurfer with the other studies shows a marked difference in performance. Compared with the results from this study in Table 7.9 for

the default condition at LPC order 13, the three formants show an improvement in performance of over 60%. It is not apparent why the results in these two studies should be so different to all of the others.

7.5 Summary

This chapter has examined the effects of using different formant trackers on formant measurement accuracy across a range of settings, for a large number of speakers. In doing so it has provided answers to all three research questions. The important outcomes are summarised as follows:

- The alignment of the measurements with the reference values in the VTR database was found to have a marked impact on the magnitude of the measurement errors. Because of this, the alignment was changed from that used in the previous chapter and some results were re-calculated to allow comparison with measurements in the current chapter.
- Whilst the behaviour of the F1 errors across the LPC orders from the trackers was similar to that seen for Praat's Burg tool, the behaviour of the errors for F2 and F3 were different. They behaved more like the F1 errors as the LPC order increased and they remained relatively constant. This was a consequence of the trackers being able to select formant candidates other than the second and third ones for F2 and F3 respectively.
- Comparison of Praat's Burg tool results with the tracker conditions revealed them to be similar. However, the tracker results were much less sensitive to changes in LPC order. It is not necessarily the case that trackers will produce more accurate measurements.
- Altering the reference formant values for Praat's tracker on a token by token basis produced the best performance.
- Setting the 'number of formants to extract' parameter to 3 for all of the trackers produced unexpectedly poor performance, particularly for F3.
- Applying the minimum error framework to the results (the equivalent of an interactive measurement process) showed improvement in the performance across all trackers and combinations of settings considered.
- Examination of the variation in measurement errors over the vowel space showed clear effects of the tracking algorithms of Praat and WaveSurfer for F2

and F3. The extremes of the vowel space showed where formant candidates were incorrectly selected due to the bias of the tracker reference values.

- A range of performance was seen across the speakers for individual tracker conditions and there were no strong relationships across individual speaker's errors for the three formants.
- Comparison of speakers' performance across the tracker conditions showed a range of relationships, with some speakers having similar performance across a range of trackers and conditions, whilst others showed varying performance across different combinations of parameters for the same tracker.
- Examining results from studies that have used the VTR database to assess the performance of other formant trackers revealed a range of reported performances for WaveSurfer and Praat. Given the lack of relevant information it was not possible to determine the specific reasons for the differences. Nevertheless, the they illustrate the problems of comparing reported performance for formant measurement tools.

The guidance that follows from these findings echoes that from previous chapters in respect of the need to select appropriate LPC orders, and that the best performance can be obtained by tailoring this setting, and measuring formants in an interactive way. Advice concerning whether or not to use a formant tracker must be motivated by the situation in which it is to be used, since in certain circumstances, namely when LPC order is not altered, Praat's simple Burg tool can outperform Praat's tracker. If information is available about the vowel being analysed then this can be used to obtain better performance. If a tracker is to be used then care must be taken to use appropriate settings for parameters such as the number of formants to be extracted. A warning is offered concerning the reliability of reported performance of formant measuring tools, since differences in the testing methodology, even with the same material, can markedly alter performance.

Chapter 8 Discussion & Guidance

This chapter brings together the findings from the previous five chapters and discusses them in the context of the three research questions. This is followed by the presentation of practical guidance on the measurement of formants for forensic speech scientists, and phoneticians more generally, that arises from these results.

8.1 Software

The first research question asked:

RQ 1. What influence does the LPC formant measuring tool have on the accuracy of formant measurements?

The results from all of the analysis chapters show that the behaviour and accuracy of formant measurements differs across software tools. This is clearly demonstrated by the reanalysis of the measurements from the pilot study presented in Section 3.4, and the comparison of the results from the formant trackers in Chapter 7, with the non-tracker results in Chapter 6. The findings show that the overall behaviour of the measurements is influenced by the measurement method employed by the software. Differences in the measurement method are apparent when the performance of F2 and F3 are examined across LPC order, for example in Figure 3.6 to Figure 3.9. For the simple formant measuring tools in Praat and Multi-Speech, which do not apply any formant tracking, the performance for F2 and F3 deteriorates at higher LPC orders as the number of poles in the LPC model increases. This results in F2 and F3 measurements being below their true values, i.e. they are underestimates. In Chapter 3, the worst performance was seen at LPC orders 16 and 18 for the F3 measurements from Praat, where almost all measurements across all vowel categories fell outside of the 300 Hz acceptable limit. The results in Chapter 6, also from Praat's measurements of real speech, showed F2 and F3 to have median errors of around -500 Hz and -800 Hz at LPC order 15, which increased in magnitude at higher orders. In contrast, the results from Praat's tracker function, and WaveSurfer, show that their performance remains relatively constant for F2 and F3 as LPC order increases, since the trackers are able to select more appropriate pole frequencies for the measured values. The alternative approach to the tracking problem adopted by the iCAbS tracker also displayed differences in the behaviour and accuracy of its measurements, again demonstrating that the underlying approach of the

software influences the measurements. The influences from the different software were seen at a specific level, by considering the distribution of errors over the vowel space, and more generally through the comparison of summary statistics.

Since the formant trackers apply prior knowledge about the nature of formants to the measurement process, it may seem a reasonable assumption that they would produce more accurate results than Praat's normal measuring tool. However, the results in Chapter 7 show that this is not necessarily the case, and making an overall assessment of performance is not straightforward. This is partly because the performance of the two approaches is not as different as might be expected, and it is also influenced by the analysis settings used, which are discussed in the following section. Another factor that makes straightforward comparisons difficult is the way in which performance is assessed. It is clear from many of the analyses that the measurements for each of the three formants behave differently, so it is important to consider them separately. However, this makes the comparison of software and analysis settings problematic when the relative performance is different across the formants. Combining the errors from all three formants to give a single measure of performance is one way in which this problem was overcome.

For Praat's normal formant measuring tool, at an LPC order of 10, the combined mean absolute error across the three formants from the TIMIT speakers was 95.5 Hz. This outperformed the best results from Praat's tracker at its default settings, which only achieved a combined mean absolute error of 100.9 Hz at LPC order 11. However, WaveSurfer's default settings achieved 92.9 Hz, whilst iCAbS gave the best performance of all with 87.5 Hz at its default settings. These results demonstrate that formant trackers do not always give the most accurate results and that the performance across trackers can vary.

The results discussed so far have only considered the accuracy of measurements when the analysis parameters were kept constant across the vowel tokens examined. Whilst this may reflect the situation where measurements are made in an entirely automated way, a significant proportion of formant measurements are made interactively by analysts. To simulate an interactive approach, a number of measurement strategies and frameworks were constructed. Since the true formant values were known, it was possible to choose measurements closest to the true values in a similar way to how a

real analyst would, with the LPC formant values overlaid on a spectrogram. In Sections 4.4.7 and 5.3.5 the numbering assignment of formants to pole frequencies employed by Praat's formant measuring tool was ignored, and a marked improvement in the accuracy of the measurements from the synthetic vowels was seen. For one speaker, the mean absolute error was reduced from 94.3 Hz to 37.6 Hz for F2 and from 268.8 Hz to 60.2 Hz for F3. Whilst this approach was not tested on the real TIMIT speech, a similar improvement is to be expected for real speech. The other strategies applied to the TIMIT speech involved allowing the LPC order to vary across tokens and formants. The application of these strategies improved the accuracy of the measurements in all circumstances. For Praat's normal tool, for a constant LPC order across all tokens and formants, the combined mean absolute error across the three formants was 115 Hz. By allowing the LPC order to vary across tokens this was reduced to 98 Hz. Allowing LPC order to vary across individual analysis frames reduced the error even further to 83 Hz, whilst the best performance of 80 Hz occurred when LPC order could vary across tokens. In general, the strategies which were the least constrained in terms of how the LPC order could vary showed the greatest improvements.

Other studies in which the performance of different software tools have been tested report variation across them (Jemaa et al 2009, Woehrling and Mareüil 2007 and Chen et al 2009). However, comparison of the results in those studies with the current work is problematic since the speech material, methodology employed and performance measures are not consistent. One of the benefits of using the VTR database is that, in principle, it allows the performance of tools that have been tested with it to be compared, since a common dataset is used. However, when attempting to compare the tools tested in the current experiments with those reported elsewhere, a number of issues arose. Some concerned the measure used to represent performance, such as RMS error, or how the results were combined for different speech segments. However, more fundamental issues were highlighted by the different performances reported for the same software, i.e. WaveSurfer, which was employed as a benchmark, in many of the studies (e.g. Deng et al 2006, Smit et al 2012, García Laínez et al. 2012 and González et al. 2012). The fact that similar results were not reported for WaveSurfer suggests differences in the testing methodology used, which raises questions about the comparability of the test results from the tools being examined. These findings highlight the difficulties, and dangers, of using reported data to make assessments and

comparisons of performance, especially when insufficient information is provided about the methodology used.

Overall, these results confirm previous observations and research studies in showing that performance does vary across different software. They also show that the accuracy of formant measurements is not only dependent on the software but also the way in which it is used to make the measurements. These findings make it clear that the performance observed for one tool cannot be assumed to apply to another, especially one with a different measurement method. This highlights the need for the testing of specific tools to understand their individual performance.

8.2 The Analysis Settings

Research question two asks:

RQ 2. How does altering the LPC analysis parameters affect formant measurement accuracy?

It is already clear from the previous section that the analysis settings influence the accuracy of formant measurements. They are inextricably linked to the software, and its measurement approach, since they control its operation and resulting behaviour.

The findings from the pilot study reported in Chapter 3 showed that LPC order had a far greater influence on the formant measurements than the frame duration or pre-emphasis. Based on this outcome, the testing of the synthetic and real speech reported in Chapter 4, Chapter 5 and Chapter 6 only considered the effects of LPC order. These chapters demonstrated the very strong influence of LPC order on measurement accuracy. It was clear that for Praat's normal formant measuring tool there is a relatively narrow range of LPC orders that produce accurate results. This range was found to be between 9 and 11 depending on the material being analysed. When the order is outside that range, the performance was markedly reduced, particularly for F2 and F3. Since LPC order controls the complexity of the LPC model, and governs the number of poles that can exist in it, the degree of influence seen on the formant measurements was unsurprising, especially given Praat's underlying measurement method.

The application of measurement strategies to replicate the approaches adopted by human analysts, where different LPC orders could be chosen, showed that the performance could be improved. The implication of this finding is that a single LPC order will not produce the most accurate formant measurements across a range of vowels or speakers. The findings demonstrated that freedom to employ different LPC orders and ignore Praat's formant numbering system were sensible approaches to improving performance. However, these both require some additional decision making in order to determine whether the chosen LPC order and pole frequencies have produced accurate measurements.

One solution to this problem is formant trackers, which often use theoretically-driven information about the nature of formants to process the results of an LPC analysis in an attempt to improve measurement accuracy. However, this approach brings with it additional analysis settings, either user-controlled or hard coded in the software, which may further influence the accuracy of the measurements. The testing of the formant trackers in Chapter 7 therefore considered some of the additional parameters related to the tracking function. Sets of parameters were selected for the testing which represented typical values an analyst might choose, and which had some comparability across the systems.

The analysis of the tracker measurements demonstrated that the rules they applied could make good decisions about formant measurements, but not for all trackers or combinations of parameters. As discussed above, the performance was not always better than Praat's normal tool, which was demonstrated by the poorer performance of Praat's tracker with its default settings. Praat's normal tool gave a combined mean absolute error of 95.5 Hz at LPC order 10, whilst the tracker with its default settings gave a combined error of 114.3 Hz at order 14. Again, application of measurement strategies to allow LPC order to vary across tokens for Praat and WaveSurfer showed an improvement in performance, with the Praat tracker achieving a combined error of 81.4 Hz. This finding shows that using trackers with a single LPC order does not achieve the best possible performance.

The analysis parameters that produced the best results were when the reference values for Praat's tracker were adjusted on a token-by-token basis to reflect the formant values of the vowel being measured. This gave a combined error of 75.5 Hz at LPC order 15.

In the error surface plots for both the Praat Tracker and WaveSurfer with their default settings, the effect of using reference values for a neutral vowel for all tokens was apparent, as shown in Figure 7.5. At the extremes of the vowel space, when the target formant value was the furthest away from the reference value, the trackers often selected the incorrect candidate formant, resulting in a large error. Comparison with the error surface from the Praat tracker with the optimum reference values showed these errors were not present. This again demonstrates that applying more information, in this case making the trackers reference values similar to the expected formant values, improves the measurement accuracy. These results were improved further still by again allowing LPC order to vary across tokens, leading to a combined mean absolute error of 60.0 Hz. Similar findings have been reported by Evanini et al (2009) who also used expected formant values for different vowel categories as part of a novel format measurement process. Statistical models were trained for each vowel category with a set of centre frequency and bandwidth values. The models were subsequently used in the tool to obtain formant measurements by selecting the pole frequency combinations from an LPC analysis that were closest to the model for the specific vowel category. The tests showed a 10% improvement for F1 and 20% improvement for F2 in the global mean absolute difference compared with hand measurements.

Not all combinations of parameters showed good performance for the trackers. The one parameter which had a marked negative influence on performance for all three trackers was ‘number of tracks’ or ‘number of formants’, when the setting was 3, rather than 4. This parameter may seem to be quite innocuous, and its function could be misinterpreted as simply defining the number of formant values that the software displays or logs. However, the results show that this is not the case. For all three trackers the measurements obtained when this parameter was set to 3 produced the worst sets of results. The mean absolute errors combined across the formants were 114.3 Hz for Praat, 182.1 Hz for WaveSurfer and 183.7 for iCAbS. For all trackers, the decrease in performance for individual formants was greatest for F3. Given these findings, some of the differences in performance across the software that were found in the pilot study can be attributed to the influence of this parameter on the results from WaveSurfer. The selection of the value for this parameter in the pilot study was done without appreciating its influence on the measurement process.

In summary, these results illustrate the influence the analysis parameters have on the resulting measurements. They show that to avoid poor performance the parameters chosen must be appropriate for the software and the material examined. It is therefore important that the influence of the parameters for a particular tool are understood by those who use it. Accepting the default settings without considering their suitability presents the danger of making inaccurate measurements. Using the same settings for all measurements also poses the same risk. The advice offered by Rose (2002, p. 267) to keep the settings the same is not supported by the findings discussed.

8.3 The Speaker

The final research question asked:

RQ 3. To what extent does the accuracy of LPC formant measurements vary across speakers?

All the experiments conducted provided some insight into the extent of variation in formant measurement accuracy that can exist across speakers. Whilst these differences in performance are relatively easy to observe and quantify, their causes are harder to determine. Since speakers display differences across many speech parameters, such as size and shape of the vowel space, fundamental frequency (F0), and voice quality, determining the source of the variation is problematic as the influence of each can be difficult to isolate.

Synthetic speech was used in the first instance to investigate the extent of variation in formant measurements across speakers. Two advantages that synthetic speech has over real speech are that the formant values can be specified in the synthesis process, so true measurement errors can be calculated, and the speech can be precisely controlled, so that the influence of speech parameters can be observed in the results. The analysis in Chapter 4 showed a strong dependence between F0 and the structure of the error surfaces, which was governed by the frequencies of the harmonics of the F0. This was also observed for the speakers in Chapter 5. A clear relationship was also seen between F0 and the magnitude of the errors, with larger errors occurring at higher F0s. These findings were in agreement with Atal and Schroeder (1974), who reported maximum errors of 11, 30 and 67 Hz for fundamental frequencies of 100, 200 and 400 Hz from synthetic speech with a single formant. Figure 4.12 shows that for the LPC orders that

produced relatively accurate measurements, the errors increased across F0, and the magnitude of the errors was different at each order. Modifying the structure of the third formant showed changes in the F3 error surfaces, which were again linked to F0, but little impact was seen in the errors for F1 and F2. The magnitude of the F3 errors were shown to be influenced by the specified F3 values as well as their location within the F1~F2 vowel space. These findings show that the errors produced by the synthetic speakers are dependent on speech parameters that vary across speakers, which implies that performance is speaker dependent.

The different glottal source speakers showed greater variation in the magnitude of the errors, and differences were again seen in performance across LPC orders. The glottal source parameter that was altered did result in changes in the behaviour of the measurements, but a clear pattern was not evident.

The findings from the synthetic speakers show that some systematic relationships do exist between speech parameters and the formant measurement errors. Whilst some of these remained relatively stable, LPC order was shown to be another factor which influenced the behaviour. Given the differences between real speech and the simple speech production model used to generate the synthetic speech, it was not apparent if these findings would be replicated for real speakers.

The formant measurements from the real speech in the TIMIT corpus showed variation in performance for Praat's standard tool across the speakers. When the overall results were divided according to sex, the mean absolute error for all three formants combined was 105.6 Hz for the men and 120.1 Hz for the women. However, when the speakers were ranked according to performance there was no obvious relationship with the speaker sex. The mean F0 values for the speakers were distributed as expected, and when the performance was compared with F0 no clear relationships were established. The same was true when performance was compared with vowel space usage, which was also distributed as expected across the sexes.

These findings can be interpreted in several ways. Since the patterns found for the synthetic speech were not seen in the real speech, it could be argued that the synthetic speech did not sufficiently model the complexity of real speech, and that performance is governed by the interactions of several factors. Alternatively, comparing the average performance of each speaker with their mean F0 may have hidden patterns, which could

have been seen by considering the F0 value for each frame or token analysed. The same could be true for vowel space usage since this was also represented by average values. Another view is that the reference values from the VTR database were not accurate enough to provide a clear picture of the behaviour of the measurements. Perhaps the reality is that a combination of all three factors has led to these findings. In any case, based on these results, none of these parameters provide a basis on which to estimate the performance of a speaker.

When the measurements from Praat's standard tool were subject to the different analysis frameworks, the relative performance of speakers tended to stay the same across the frameworks i.e. those speakers who performed well for one framework did so for the others. However, when the performance of speakers was considered across the different formant trackers and analysis conditions, the relationships were more complex. These results show that the accuracy of measurements for individual speakers is to some extent dependent on the analysis tool used.

For the analysis frameworks where the LPC order was permitted to vary across tokens, the smallest measurement errors were obtained when a range of LPC orders were used for each speaker. The specific orders, and the range of orders, varied across speakers. This finding is in agreement with Vallabha and Tuller (2004) who also showed that speakers have a range of optimum LPC orders rather than a single one.

In summary, these findings demonstrate that the speaker is an integral part of the formant measurement process and that the performance of tools is not only governed by the analysis settings, but by the speaker being analysed. This makes the assessment of formant measurement tools problematic as their performance is to some extent determined by the material they are analysing, and the behaviour of the material can be different across tools. These issues all serve to highlight the complexities and potential difficulties in measuring formants.

8.4 Guidance

The ultimate aim of this thesis is to offer information and practical guidance which will assist analysts when making formant measurements. Whilst the discussion above and the results presented in the previous chapters provide helpful information and insights,

the current section sets out practical guidance and advice based on the findings of this research.

The most general advice that can be offered is that measuring formants by LPC analysis should not be treated as a simple automatic process which can be left to a computer to carry out blindly. It should be clear from the research summarised in Chapter 2, and the tests described in this thesis, that formant measurements are influenced by many factors, including the software tool used, the analysis settings and the speaker. If analysts acknowledge that these factors will have a bearing on the formant measurements they make, and an understanding of their effects, then they will be in a better position to make more reliable measurements.

To use a formant measuring tool properly, analysts must understand the process it follows to obtain the measurements. This must include, as a starting point, an understanding of the principles of LPC analysis, as described in Section 1.2.3, including its limitations. Without this knowledge it is difficult to understand the process used within a specific tool. The knowledge of a particular implementation must also include, as an absolute minimum, whether the tool performs formant tracking or not. In the author's experience it is common for the term 'formant tracker' to be incorrectly used to refer to Praat's normal formant measuring tool. Whilst the term is certainly more elegant than 'normal formant measuring tool', its use may well lead to the user believing that the tool conducts formant tracking, when it does not. The problem can be compounded by referring to the formant measurements overlaid on a spectrogram, as shown in Figure 1.8, as 'formant tracks', again reinforcing the misapprehension that they are the result of a formant tracking process. This may lead to the incorrect assumption that the results are more likely to be accurate as they come from a tracker and that the analyst will not intervene and adjust parameters, such as LPC order, in an attempt to obtain better measurements. By understanding issues such as these, and how the tool works, analysts should be able to better interpret the measurements, and adjust and select appropriate settings to achieve more accurate results.

In addition to understanding the underlying measurement process of their software, analysts should also be aware of the influence of analysis parameters on the resulting formant measurements and any peculiarities that exist for their chosen tools. This is especially important for LPC order, as this has been shown to have a significant effect

on accuracy. Analysts should understand how LPC order relates to the underlying LPC analysis and how it influences the modelled LPC spectrum, as shown in Figure 1.9, and the number of poles, from which the formant measurements are derived. They should also be aware of the different ways in which the parameter may be specified in software. For instance, in Praat, LPC order is specified via the 'Number of formants' setting, leading to an LPC order that is twice this value. Specifying the parameter in this way gives rise to the unconventional and potentially confusing situation where the number of formants can be specified as half integers to select odd numbered LPC orders, i.e. if the 'Number of formants' setting is 4.5 then the LPC order is 9. Additionally, the name used for this parameter may also reinforce the misconception that the tool is a formant tracker. Another example of a parameter that is specified in a non-conventional way is pre-emphasis in Praat. This specifies the frequency above which the pre-emphasis is applied rather than the more commonly used pre-emphasis filter coefficient.

Care should also be taken with the selection of analysis settings for formant trackers, as these can also have a marked influence on measurement accuracy. The parameter 'Number of tracks' in Praat or 'number of formants' in WaveSurfer and iCAbS was shown to affect the tracking process and does not simply determine how many formant values are logged or presented by the software. This is particularly important if measurements are to be made automatically and not compared with spectrograms.

An understanding of the behaviour of the software used is perhaps best gained through the use of it with commonly encountered speech materials, such as poor quality or telephone recordings in the forensic context, combined with an understanding of its specific underlying measurement process and analysis parameters. Without some knowledge of the software's underlying process it may be difficult to interpret the effects seen on measurements when analysis parameters are adjusted. So some responsibility must lie with the authors of the software to provide sufficient information. A user manual cannot be expected to include detailed information concerning the behaviour of the software in a wide range of scenarios, but information about the measurement process, and parameters of the tool, will be invaluable to analysts when attempting to interpret the behaviour of the software. In this respect, the Praat manual (Boersma 2010) is particularly helpful and contains a sufficient level of detail for analysts to understand the measurement process adopted by the software. In contrast, the WaveSurfer manual page (Sjölander and Beskow 2006b) contain no information

about the analysis method or the analysis parameters. Information is available within the documentation for the Snack Sound Toolkit (Sjölander 2004), which WaveSurfer is built from, but this documentation is unlikely to be found, or even known about, by a typical analyst.

The way in which the software tools are used will have an influence on the accuracy of the measurements that can be made with them. The most accurate results will be obtained when the tools are used in an interactive way, where measurements are overlaid on spectrograms and the analysis parameters, such as LPC order, are adjusted, where necessary, on a token by token or even formant by formant basis. The results in Section 6.3.2 showed that the smallest errors occurred when the variation of LPC order across tokens and formants was least constrained. Overlaying the measurements on a spectrogram allows the analyst to make a visual comparison between the spectral representation of the signal and the formant values obtained over a range of analysis parameters. A decision as to whether to accept or reject the measurements obtained with a specific combination of analysis settings will be based on the degree of visual alignment between the measurements and the representation of the formants seen in the spectrogram. Such an approach cannot be guaranteed to obtain the most accurate measurements possible, since certain combinations of analysis parameters will produce similar measurements and attempting to determine which is the most accurate is problematic. This is due to the difficulty in determining the centre of formants within spectrograms and the fact that their appearance is also governed by the analysis settings used to generate the spectrogram. Additional consultation of FFT or LPC spectra may assist where the interpretation of the spectrogram is problematic. Whilst not a perfect approach, allowing analysis parameters to vary means that obviously erroneous measurements are rejected which would otherwise be accepted if the analysis parameters remained constant across all tokens and formants. The advice to adopt this method is counter to that offered by Rose (2002, p. 267) who recommends all settings be kept constant, but echoes the approach suggested by Ladefoged (1996, p.212) to try a range of settings.

One significant drawback of this approach is that it is a time consuming process. Automated approaches in which measurements are made without any direct intervention by an analyst can still yield relatively accurate results, but there is a danger that the analysis parameters used may not be suitable for certain speakers, vowels or formants.

In general, it was found that the more specific the tailoring of the analysis parameters, the greater the level of accuracy that can be achieved. The issue then arises of how to tailor the settings. If a speech corpus contains both male and female speakers then it is likely that the sound files will be coded for sex. This can easily be factored into the analysis settings by applying a different maximum analysis frequency based on the sex of the speaker, as was done in this research. If the speech material has been segmented and the vowels labelled then this information can be used in Praat to modify the tracker settings on a vowel by vowel basis, which was shown to produce the most accurate results.

The suitability of other approaches for determining appropriate settings will be governed to some extent by the amount of material being analysed and its variability in terms of factors such as the diversity of recording channels and the number of speakers. However, a method should be adopted to check the suitability of the chosen parameters. In the absence of any standardised approaches, this is perhaps best achieved by examining the measurements. This could involve overlaying measurements on spectrograms for a representative sample of material to gain an impression of their accuracy, and allow the identification of any problematic tokens. An alternative approach could involve examining the distributions of measurements to detect the occurrence of obviously erroneous values. Since automated measurements can be made repeatedly, it could be beneficial to obtain the measurements with different sets of analysis parameters and examine the variation in the resulting measurements. This could involve checking the distribution of measurements, in order to obtain an overall impression of the data, and determine how sensitive the measurements are to the adjustment of parameters. Until more systematic methods for determining suitable parameters are developed, such as the one discussed by Vallabha and Tuller (2002, 2004), then the application of knowledge and checking of measurements provides the best solution.

For measurements obtained by either an interactive or automated method, the default settings of the software may produce accurate measurements. However, they should be treated as a useful starting point and not universally applied without due consideration to the material being analysed. An obvious example is Praat's 'Maximum formant frequency' setting, which has a default value 5,500 Hz that is more appropriate for female rather than male speakers. Since the vast majority of recordings encountered in

forensic casework involve men rather than women, this default setting is not appropriate most of the time. Another example from Praat is the default setting of 3 for the 'Number of tracks' parameter for the tracker. This was shown to give particularly poor performance for F3 measurements. Given the factors that have been shown to influence formant measurements, both in this thesis and in those studies described in Chapter 2, recommending suitable settings for different scenarios would be unwise. As already stated above analysts should select appropriate settings based on their knowledge of the software and the material being analysed.

The validation of methods and tools used in forensic analysis was discussed in Section 1.3.5. The research presented in this thesis will be helpful to those tasked with designing and implementing validation exercises, and writing standard operating procedures. Based on the outcomes of this research it is apparent that a key element of validating formant measurement methods must be the competency testing of analysts who make the measurements. An analysis tool may be shown to produce accurate results for a range of speakers and recording conditions, but if an analyst is not able to use the tool effectively by selecting suitable analysis parameters and making accurate measurements, then the method will not have achieved its aim. Another important consideration in the forensic context is the consistency of measurements both for individual analysts and across analysts. Achieving consistency in measurements will reduce the dependency of the results on the individual analyst and allow measurements and findings to be repeated by others. This can be achieved by following the guidance given so far and adopting analysis approaches such as those used by Duckworth et al. (2011). These included using standard settings for some parameters, but allowing LPC order to vary, and making measurements at a single time point in a relatively stable part of the token around its maximum intensity (2011, p. 40).

Because the analysis parameters selected can have a large influence on the accuracy of measurements, the analysis settings used and the location in a sound file where the measurements were made should be logged. This not only provides an accurate record of the work carried out, which is a general requirement of forensic analysis, but makes the reviewing of measurements easier. In the forensic context, this may be done by a colleague who is peer reviewing an analysts work, or by another expert. Similar advice is given by Duckworth et al. (2011).

The logging and reporting of analysis settings is recommended for speech research more generally, where formant analysis is conducted. Accurately reporting the method followed, including the analysis parameters used, allows others to critically assess the methodology and potential accuracy of the measurements. This permits a better understanding of the results and their implications. Simply stating the software that was used to make the measurements is wholly insufficient. Accurate reporting also allows work to be replicated. Had further information been provided on the method followed in the studies reported by García Laínez et al. (2012) and González et al. (2012), which are discussed in Section 7.4.3, it may have been apparent why their claimed performance for WaveSurfer was much better than the other studies discussed in Section 7.4 and those presented in this thesis.

Since the research presented has compared the performance of different software, a potential outcome might have been to recommend one tool over the others. However, the results did not indicate a clear and universally valid choice. In common with the guidance offered above, it is the analyst's knowledge of the particular tool, the analysis parameters chosen and the way in which it is used that will have greater influence on its performance than any fundamental differences between the tools. Other factors may influence the choice of software, including the ease with which parameters can be adjusted, how the measurements are displayed and logged, and how easily the measurement process can be automated. The ability to automate certain tasks via scripts, as permitted in e.g. Praat, is particularly useful and has facilitated the research reported in this thesis. Some of the recommendations given above involving repeated measurements with different settings would be particular difficult to undertake, especially on large datasets, without some form of automation. The automatic logging of measurements, together with the settings used to obtain them and the timings from which they originate, can significantly reduce the time burden when measuring formants interactively (French and Harrison, 2004).

8.4.1 Impact on Forensic Analysis

The main motivating factor for the research conducted in this thesis is the lack of information and guidance for forensic practitioners concerning the measurement of formants, as discussed in Section 2.3.1. This thesis not only presents results that demonstrate the magnitude and behaviour of errors that can be encountered across different software, settings and speakers, but it also provides specific guidance. If

forensic analysts apply the guidance presented in this chapter, and critically assess their measurements in light of the experimental findings, then they should make more accurate formant measurements and be less likely to misinterpret them. This should lead to more reliable conclusions concerning speaker identity and disputed content. Ultimately, the impact of this work is that the risk of a miscarriage of justice is reduced.

The results of the experiments have shown the magnitude of errors that can occur, and provided insights into the influence of the measurement tool, the analysis settings and the speaker on the accuracy of formant measurements. This work should raise awareness of these factors within the forensic community, and by gaining an understanding of them, analysts should give more critical consideration to them when interpreting measurements and drawing conclusions. At present, the telephone effect on formants is well known within the field, often being cited in research and taken into account when interpreting measurements. However, little explicit acknowledgement is given to the influence of software, settings or speakers either in forensic research or in casework. By drawing attention to the importance of these factors this situation will hopefully change.

At the most fundamental level, the findings reinforce the point that all formant measurements must be assumed to be inaccurate to some degree. Forensic analysts must always consider this when making and interpreting measurements. If this is coupled with an understanding of the basis of LPC analysis and knowledge of the operating principles of the software used, then these factors alone should allow forensic analysts to begin critically assessing the likely accuracy of measurements rather than blindly accepting them as accurate.

At a more specific level, the findings show the errors obtained at LPC orders which produce the most accurate measurements tend to be distributed symmetrically around zero (see for example Section 6.3.1.1). The implication of this is that distributions of relatively accurate formant measurements can appear to be wider than they truly are. In sparsely populated distributions the opposite is also possible, so distributions may appear narrower. This has implications for speaker comparison analysis where distributions of measurements are compared across recordings and is especially relevant where there is only limited or non-existent overlap, since this may be a consequence of errors rather than a true reflection of the similarities of the distributions. This applies

equally to tests which rely on manual comparison as well as automatic ones such as the MVKD or GMM-UBM approaches. Analysts must therefore take this factor into account when conducting such comparisons and interpreting the outcomes.

The experimental results also revealed systematic interactions between the errors and a number of speech and analysis parameters. For example, the results from the synthetic speech showed a tendency for errors to be larger for higher fundamental frequencies. This relationship is particularly important for analysts to take into account when interpreting measurements from the speech of women and children as they generally have higher fundamental frequencies than men. Interactions were also seen across the F1~F2 vowel space with the largest errors tending to occur at the edges of the space. However, the specific patterning was different across the formants and the measurement approaches adopted. Therefore, as a general principle, analysts should be more cautious of measurements originating from the edges of a speaker's vowel space. This guidance is particularly relevant to measurements made using WaveSurfer, or Praat's tracker with default settings, as the errors at the edge of the space were shown to be associated with the tracker employing reference formant values for a central vowel.

The factors that were investigated revealed general tendencies in the errors rather than rigid relationships which applied universally and could be accounted for in a consistent manner. Whilst this finding is less problematic where a large number of tokens are analysed and the emphasis is on the overall distribution of the measurements, it presents issues where an individual token or small number of tokens are concerned. For a single token, it cannot be known how the influence of each factor has combined to affect the overall accuracy of the measurements. Analysts should be particularly cautious when undertaking disputed utterance cases, where the focus is often on the measurements from a single token. Great care should be taken to ensure that the most appropriate settings are used and that the measurements are checked against spectrograms and other spectral representations to ensure they are as reliable as possible. Extreme caution should be applied when interpreting the measurements if such checks cannot be satisfactorily done, which may be the situation with poor quality or noisy recordings.

8.4.2 Guidance Summary

The guidance and advice offered above can be distilled into the following three key points, which if applied, should lead to analysts making more accurate formant measurements:

- Understand the principles of LPC analysis and how the analysis parameters can affect the resulting measurements
- Understand how the LPC analysis process is implemented in the software being used and how the analysis parameters configure the underlying measurement process
- Based on this knowledge, tailor the analysis approach and the analysis parameters to the speech being analysed, at the formant, token or speaker level, where practical.

Chapter 9 Conclusions

9.1 Thesis Summary

In Chapter 1, formants were introduced and defined with reference to the source-filter model of speech production. Methods of measuring formants were discussed, with particular attention paid to the LPC approach, as this was the method used throughout the thesis. The chapter concluded with a discussion on the use of formant measurements within the field of forensic speech science, this being the area from which the motivation for this research originated.

Chapter 2 summarised previous research relating to the accuracy of formant measurements. This focused on the measurement method, variability introduced by the analyst and technical aspects of the speech signal. The chapter discussed the limited advice concerning formant measurements, and highlighted the lack of guidance relating to commonly used software. The overall aim of the thesis was presented, which was to provide such guidance. The research questions were stated, which focused on investigating the influence of three factors on formant measurement accuracy: the software, the analysis settings and the speaker.

In Chapter 3, the findings and a further analysis of results from a pilot study were presented, which examined the variation in formant measurements encountered across three commonly used software tools. Formants were measured in recordings of two speakers reading a word list, made both directly via a microphone, and over a telephone line. The analysis parameters of LPC order, frame length and pre-emphasis were varied. Differences were found in the variability of the measurements across the software, the speakers, the vowel categories and recording conditions. Of the three analysis parameters, LPC order was found to have the greatest influence on the measurements. Whether the measurement tool used a formant tracking process or not also had a marked effect on the measurements. The outcomes of this study helped to shape the main body of this research.

The formant measurement errors from a single synthetic speaker were examined in Chapter 4. Synthetic vowel tokens based on realistic formant values were generated across a range of fundamental frequencies. The formants of the resulting vowels were analysed using Praat's standard formant measuring tool for a range of LPC orders. The

measurement errors were examined for the three formants across the vowel space. Systematic differences were observed in the error surfaces for the three formants, which were influenced by LPC order. The magnitude of the errors was found to increase as fundamental frequency increased. Imposing different measurement strategies on the measurements, which allowed LPC order to vary across tokens, and which ignored Praat's formant numbering approach, showed a marked improvement in the accuracy of the measurements.

In Chapter 5, two sets of synthetic speakers were generated in order to examine the influence that different speaker characteristics might have on measurement accuracy. For the first set of speakers the structure of their third formant across the F1~F2 vowel space was altered. Measurements from these speakers showed that changing this structure had the greatest effect on F3 errors, which were largest for the synthetic speakers with higher average F3 values. The second set of speakers that were created had different glottal source waveforms. These speakers exhibited larger errors than the first set at certain LPC orders, which were a consequence of localised regions in the vowel space that had poor performance. Again, the behaviour of the measurements was found to differ across LPC orders.

Chapter 6 returned to the analysis of real speech and examined 518 sentences from the TIMIT corpus for which a set of hand corrected formant values were available. These were used as reference values from which formant measurement errors were calculated. Again, the measurement were made using Praat's normal measuring tool across a range of LPC orders. A number of measurement strategies were imposed on the results to reflect the ways a real analyst might make measurements. Allowing the LPC order to vary across the three formants and across tokens was found to produce the greatest reduction in the overall error. The performance of individual speakers was compared with the speaker characteristics of sex, average fundamental frequency and location within the vowel space. However, no strong relationships were found. Performance was shown to vary across speakers, as was a preference for different ranges of LPC orders.

In Chapter 7, the same speech material was analysed using three formant trackers with a number of different analysis parameter combinations. The measurement errors were analysed in similar ways to those in the previous chapter. Unsurprisingly, the performance of the trackers was found to be less sensitive to variation in LPC order.

However, certain combinations of parameters resulted in poor performance. Allowing the LPC order to vary across tokens again produced a reduction in the magnitude of the errors. Overall, the most accurate measurements were obtained when the reference values for Praat's tracker were altered on a token-by-token basis. The results from the individual speakers were compared across the formant trackers and speaker characteristics, but no clear patterns emerged.

The discussion in Chapter 8 brought together the key findings from the previous chapters and considered them in light of the three factors investigated: the software, the analysis settings and the speaker. This was followed by statements of guidance based on the findings. The guidance suggested that understanding the principles of LPC analysis, how it was implemented in specific software and the influence of analysis parameters were important when making formant measurements. By using this knowledge to tailor the analysis approach and analysis parameters, analysts could be expected to make more accurate measurements.

9.2 Summary of Research Contribution

The research presented in this thesis has fulfilled its overall aim by providing guidance and information that will assist analysts in making more accurate formant measurements. The experiments conducted explored the influences of the software tool, the analysis parameters and the speaker on formant measurement accuracy. Not only do the results of these experiments form the basis of the guidance presented, they contribute to knowledge concerning the accuracy of formant measurements and the factors that affect them.

The research is the first comprehensive investigation of the performance of software currently used by analysts to measure formants, which is based on a large set of data from a wide range of speakers. It is also the first study to provide guidance on the measurement of formants based on such an empirical investigation.

The contribution of this research is of particular importance as the formant measurements were made in software commonly used by analysts. The results and findings are therefore directly applicable to those specific tools. Since these tools are not restricted to a specific narrow discipline, the outcomes of this research can influence formant analyses conducted across a range of fields. By aiming to increase the accuracy

of the formant measurements this work has the potential to improve the performance of other analysis techniques which are based on formant data.

9.3 Further Research

The research presented in this thesis raises questions and opportunities for further investigation that also have the potential to improve formant measurement accuracy. From a practical perspective, it is sensible to ask to what extent the guidance provided leads to greater accuracy in formant measurements. This could be tested by assessing the performance of analysts before and after receiving the guidance. Comparison of their performance would demonstrate the effectiveness of the guidance and could reveal opportunities for it to be improved or refined. As this research has suggested, and prior research has shown, the analyst is a key component of the measurement process, so a better understanding of the strategies they employ, and their ability to apply the guidance, would provide useful insights.

In terms of the investigation of the effects of analysis parameters, a potentially important one that was not considered in this study is the maximum analysis frequency. This parameter has a marked influence on the LPC analysis, which is inextricably linked with the LPC order. However, it is not apparent to what extent adjusting this parameter, perhaps in combination with the LPC order, could improve the accuracy of formant measurements. The scripts for performing the measurements and analysing the measurement errors already exist, so examining its effects would be straightforward. Since the parameter controls the frequency bandwidth over which the LPC analysis occurs, it may have a significant influence on recordings with a restricted frequency bandwidth, such as those from telephone lines, which are often encountered in forensic casework.

Whilst the findings of the present study are applicable to the forensic context, one forensically relevant parameter that was not considered in detail is the quality of the recording. The findings from the pilot study showed differences in the accuracy of the measurements between the microphone and telephone conditions. Since telephone recordings are frequently encountered in casework, the investigation of their influence on formant measurements would be a welcome extension of the present research. The findings would also be of relevance more widely where phonetic data are collected via the telephone. Again, this could be achieved easily using the materials already analysed,

by re-recording the speech via various telephone connection types. To ensure comparability of the results, care would need to be taken to ensure that the VTR reference values were correctly aligned with the new recordings, since this has been shown to affect the results. The analysis could be conducted in parallel with investigations into the influence of the maximum analysis frequency parameter.

Another aspect of this work which deserves further investigation is the relationships between the speaker, the analysis settings and their performance. A greater understanding of the interdependency between these factors would allow a better informed selection of analysis parameters and understanding of the variation in performance that can be expected across speakers. A starting point may be to examine in greater detail the results from the 24 speakers for whom eight sentences were analysed. The creation of a simple test to pre-determine suitable analysis parameters would potentially lead to improvements in performance.

Finally, the further development of formant trackers may negate the requirement to pre-determine standard analysis parameters. The iCAbS tracker showed strong potential in this respect. The principle of comparing the signal with LPC models obtained from different sets of parameters mirrors the analysis strategies that were shown to give better performance. Again, based on the results presented, developing a tracker in which the LPC order could vary across formants would likely yield improvements in performance.

9.4 Conclusion

The thesis has demonstrated that the software used, the analysis settings employed and the speakers being analysed all influence the accuracy of formant measurements. By using knowledge of LPC analysis, its specific implementation in software and understanding the influence of analysis parameters, analysts can make more accurate measurements. It is hoped that the guidance provided will be followed and that more accurate formant measurements will be made.

Bibliography

Aitken, C. G. G. & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics,* 53(1), 109-122.

Alexander, A., Forth, O., Jessen, M. & Jessen, M. (2013). Speaker recognition with phonetic and automatic features using VOCALISE software. Presented at IAFPA 22nd Annual Conference, Tampa, USA.

Atal, B. S. & Hanauer, S. L. (1971). Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. *Journal of the Acoustical Society of America*, 50(2), 637-655.

Atal, B. S. & Schroeder, M. R. (1974). Recent advances in predictive coding - applications to speech synthesis. *Proceedings of the 1974 Stockholm Speech Communications Seminar*, C.G.M. Fant, ed. John Wiley and Sons, NY, NY, 27-31.

Atal, B. S. (2006). The History of Linear Prediction. *IEEE Signal Processing Magazine,* 23, 154-161.

Baken, R. J. & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice 2nd ed*. San Diego, CA: Singular.

Becker, T., Jessen, M. & Grigoras, C. (2008). Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models. *Proceedings of Interspeech 2008 Incorporating SST 2008, International Speech Communication Association*, 1505-1508.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Boersma, P. (2002). *Formant: Track...* [Online] University of Amsterdam. Available at: http://www.fon.hum.uva.nl/praat/manual/Formant__Track___.html [Accessed 29th June 2014]

Boersma, P. (2010). *Sound: To Formant (burg)...* [Online] University of Amsterdam. Available at: http://www.fon.hum.uva.nl/praat/manual/Sound__To_Formant__burg____.html [Accessed 28th June 2014]

Broad, D. J. & Wakita, H. (1977). Piecewise–planar representation of vowel formant frequencies. *Journal of the Acoustical Society of America*, 62(6), 1467-1473.

Byrne, C. & Foulkes, P. (2004). The 'Mobile Phone Effect' on vowel formants. *International Journal of Speech Language and the Law*, 11(1), 83-102.

Castro, A. D., Ramos, D. & Gonzalez-Rodriguez, J. (2009). Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking. *Proceedings of Interspeech 2009*, Brighton, UK, 2343-2346.

Chandra, S. & Lin, W. (1974). Experimental comparison between stationary and nonstationary formulations of linear prediction applied to voiced speech analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 22(6), 403-415.

Chen, N. F., Shen, W., Campbell, J. & Schwartz, R. (2009). Large-scale analysis of formant frequency estimation variability in conversational telephone speech. *Proceedings of Interspeech 2009*, Brighton, UK, 2203-2206.

Childers, D. G. (1978). *Modern spectrum analysis*. New York: IEEE Press.

Clark, J. & Yallop, C. (1995). *An introduction to phonetics and phonology. 2nd edn*. Oxford: Blackwell.

Clément, P., Hans, S., Hartle, D. M., Maeda, S., Vaissière, J. & Brasnu, D. (2007). Vocal tract area function for vowels using three-dimensional magnetic resonance imaging. A preliminary study. *Journal of Voice*, 21(5), 522-530.

Clermont, F. (1991). *Formant-Contour Models of Diphthongs: A Study in Acoustic Phonetics and Computer Modelling of Speech.* PhD Thesis, The Australian National University, Institute of Advanced Studies.

Clermont, F. (1992). Formant-contour parameterisation of vocalic sounds by temporally-constrained spectral matching. *Proceedings of 4th Australian International Conference on Speech Science & Technology*, Brisbane, 48-53.

Clermont, F. & Mokhtari, P. (1994). Frequency-band specification in cepstral distance computation. *Proceedings of 5th Australian International Conference on Speech Science & Technology*, Perth, vol. 1, 354-359.

Clermont, F., French, J. P., Harrison, P. T. & Simpson, S. (2008). Population data for English spoken in England: A modest first step. Presented at IAFPA 17th Annual Conference, Lausanne, Switzerland.

Cui, X. (2013). Email to P Harrison re. VTR Database, 21st February 2013.

Delaunay, B. (1934). Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, 7(793-800), 1-2.

Deng, L., Lee, L. J., Attias, H. & Acero, A. (2004). A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances. *Acoustics, Speech, and Signal Processing, 2004. ICASSP 2004 Proceedings. IEEE International Conference on*, 1, 557-560.

Deng, L., Cui, X., Pruvenok, R., Chen, Y., Momen, S. & Alwan, A. (2006). A database of vocal tract resonance trajectories for research in speech processing. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 1, 369-372.

Deng, L. (2011). Email to P Harrison re. VTR Database, 21st March 2011.

Deng, L. (2013). Email to P Harrison re. VTR Database, 21st February 2013.

Duckworth, M., McDougall, K., de Jong, G. & Shockey, L. (2011). Improving the consistency of formant measurement. *International Journal of Speech, Language and the Law*, 18(1), 35-51.

Enzinger, E. (2010). Measuring the effects of adaptive multirate (AMR) codecs on formant tracker performance. *Journal of the Acoustical Society of America*, 128(4), 2394.

Epps, J., Smith, J. R. & Wolfe, J. (1997). A novel instrument to measure acoustic resonances of the vocal tract during phonation. *Measurement, Science and Technology*, 8, 1112-1121.

Evett, I. W. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, 38(3), 198-202.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fant, G. (1973). *Speech Sounds and Features*. MIT, Cambridge, MA.

Fant, G., Liljencrants, J. & Lin, Q. G. (1985). A four-parameter model of glottal flow. *STL-QPSR*, 26(4), 1-13.

Fitch, J. L. & Holbrook, A (1970). Modal vocal fundamental frequency of young adults. *Archives of Otolaryngology*, 92, 379-382.

Flanagan, J. L. (1972). *Speech analysis synthesis and perception*. New York: Springer-Verlag.

Forensic Science Regulator (2011). *Codes of Practice and Conduct for forensic science providers and practitioners in the Criminal Justice System*. Forensic Science Regulator.

Foulkes, P. & French, J. P. (2012). Forensic speaker comparison: a linguistic-acoustic perspective. In P. Tiersma & L. Solan (eds.) Oxford Handbook of Language and Law. Oxford: Oxford University Press. 557-572.

French, P. (1990). Analytic Procedures for the Determination of Disputed Utterances. In H. Kniffka (ed.), *Texte zu Theorie und Praxis forensischer Linguistik*. Tübingen: Niemeyer Verlag. 201-213.

French, J. P. & Harrison, P. T. (2004). Adapting the Praat speech analysis programme to the purposes of forensic phonetic casework and research. Presented at BAAP Colloquium, Cambridge.

French, P. & Harrison, P. (2006). Investigative and Evidential Applications of Forensic Speech Science. In A. Heaton-Armstrong, E. Shepherd, G. Gudjonsson & D. Wolchover (eds.), *Witness Testimony: Psychological, Investigative and Evidential Perspectives*. Oxford: Oxford University Press. 247-262.

French, J. P. & Harrison, P. (2007). Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech, Language and the Law*, 14(1), 137-144.

French, J. P., Nolan, F., Foulkes, P., Harrison, P. & McDougall, K. (2010). The UK position statement on forensic speaker comparison: a rejoinder to Rose and Morrison. *International Journal of Speech, Language and the Law*, 17(2), 143-152.

French, J. P., Foulkes, P., Harrison, P., & Stevens, L. (2012). Vocal tract output measures: relative efficacy, interrelationships and limitations. Presented at IAFPA 21st Annual Conference, Santander, Spain.

Fry, D. B. (1979). *The Physics of Speech*. Cambridge: Cambridge University Press.

García Laínez, J. E., Ribas González, D., Miguel Artiaga, A., Lleida Solano, E. & Calvo de Lara, J. R. (2012). Beam-search formant tracking algorithm based on trajectory functions for continuous speech. *CIARP 2012, LNCS 7441*, 749-756.

Gläser, C., Heckmann, M., Joublin, F., Goerick, C. & Groß, H. M. (2007). Joint estimation of formant trajectories via spectro-temporal smoothing and Bayesian techniques. *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 4, 477-480.

Gläser, C., Heckmann, M., Joublin, F. & Goerick, C. (2010). Combining auditory preprocessing and Bayesian estimation for robust formant tracking. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(2), 224-236.

Gold, E. & French, J. P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 18 (2), 293-307.

González, D. R., Laínez, J. E. G., Miguel, A., Gimenez, A. O., Lleida, E. & de Lara, J. R. C. (2012). Evaluation of a New Beam-Search Formant Tracking Algorithm in Noisy Environments. In *Advances in Speech and Language Technologies for Iberian Languages*. Springer Berlin Heidelberg. 40-48.

Gray, D. C. (2005). *Acoustic Pulse Reflectometry for Measurement of the Vocal Tract with Application in Voice Synthesis*. PhD Thesis, University of Edinburgh.

Griesbach, R., Esser, O. & Weinstock, C. (1995). Speaker identification by formant contours. In A. Braun, & O. Köster (eds.) *Studies in forensic phonetics*. Trier: Wissenschatlicher Verlag. 49-55.

Guillemin, B. J. & Watson, C. I. (2006). Impact of the GSM AMR Speech Codec on Formant Information Important to Forensic Speaker Identification. *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, 483-488.

Guillemin, B. J. & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *International Journal of Speech, Language and the Law*, 15(2), 193-218.

Hansen, G. F. & Pharao, N. (2006). Microphones and Measurements. *Proceedings from Fonetik 2006*, Lund, 49-52.

Harrington, J. & Cassidy, S. (1999). *Techniques in Speech Acoustics*. Dordrecht: Kluwer Academic Publishers.

Harrison, P (2004). *Variability of Formant Measurements*. MA Dissertation, University of York.

Harrison, P. (2006). *Variability of formant measurements – Part 2*. Presented at IAFPA 15th Annual Conference, Göteborg, Sweden.

Harrison, P (2007). *Formant Measurement Errors: Preliminary Results from Synthetic Speech*, Presented at IAFPA 16th Annual Conference, Plymouth, UK.

Harrison, P (2008a). *Formant Errors from Synthetic Speech*. Poster at BAAP Colloquium, Sheffield, UK.

Harrison, P (2008b). Formant Measurements Errors from Synthetic Speech. *Proceedings of the Institute of Acoustics*, 30(2), 638-645.

Harrison, P (2008c). *A Method for Reducing Formant Measurement Errors in Synthetic Speech*. Presented at IAFPA 17th Annual Conference, Lausanne, Switzerland.

Hawks, J. W. & Miller, J. D. (1995). A formant bandwidth estimation procedure for vowel synthesis. *Journal of the Acoustical Society of America*, 97(2), 1343-1344.

Hillenbrand, J. M., Clark, M. J. & Nearey, T. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, 109(2), 748-763.

Home Office (2003). *Advice on the use of voice identification parades*. UK Home Office Circular 057/2003 from the Crime Reduction and Community Safety Group, Police Leadership and Powers Unit.

Howard, D. M. (1998). Practical voice measurement. In T. Harris, S. Harris, J. S. Rubin & D. M. Howard (eds.) *The Voice Clinic Handbook*. London: Whurr Publishers Ltd. 323-382.

Howard, D. M., Hirson, A., French, J. P. & Szymanski, J. E. (1993). A survey of fundamental frequency estimation techniques used in forensic phonetics. *Proceedings of the Institute of Acoustics*, 15(7), 207-215.

Howard, D. M. (2002). The real and the non-real in speech measurements. *Medical Engineering & Physics*. 24, 493-500.

ISO/IEC 17025 (2005). *General requirements for the competence of testing and calibration laboratories*. Geneva: ISO.

Jemaa, I., Rekhis, O., Ouni, K. & Laprie, Y. (2009). An evaluation of formant tracking methods on an Arabic database. *Proceedings of Interspeech 2009*, Brighton, UK, 1667-1670.

Jessen, M. (2008). Forensic Phonetics. *Language and Linguistics Compass*, 2(4), 671-711.

Johnson, K. (1997). *Acoustic and Auditory Phonetics*. Oxford: Blackwell.

Kassin, S. M., Dror, I. E. & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, 2 (1), 42-52.

Kasuya, H. & Yoshizawa, S. (1992). Geometric representation of speaker individuality in formant space and its application to speech synthesis. *Proceedings of 14th ICA*, Beijing, China.

Kasuya, H., Tan, X. & Yang, C. (1994). Voice source and vocal tract characteristics associated with speaker individuality. *Proceedings of The 3rd International Conference on Spoken Language Processing, ICSLP 1994*, Yokohama, Japan, 1459-1462.

Kay Elemetrics, (2004). *Multi-Speech* [Computer Program]

Kent, R. D. & Read, C. (2002). *Acoustic analysis of speech. 2nd edn*. San Diego: Singular Publishing.

Klatt, D. H. & Klatt, L. C. (1990). Analysis, synthesis and perception of voice quality variations among male and female talkers. *Journal of the Acoustical Society of America*, 87(2), 820–856.

Künzel, H. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics*, 8(1), 80-99.

Karlsson, I. (1975). Formant measurements on female voices. *STL-QPSR*, 16(4), 21-26.

Kewley-Port, D. & Watson, C. S. (1994). Formant-frequency discrimination for isolated English vowels. *Journal of the Acoustical Society of America*, 95(1), 485-496.

Ladefoged, P. (1996). *Elements of Acoustic Phonetics*. 2nd edn. London: The University of Chicago Press.

Law Commission (2011). *Expert Evidence in Criminal Proceedings*. London: The Stationery Office.

Lindblom, B. (1960). Spectrographic measurements. *STL-QPSR*, 1(2), 5-6.

Lindblom, B. E. F. & Sundberg, J. E. F. (1971). Acoustical Consequences of Lip, Tongue, Jaw and Larynx Movement. *Journal of the Acoustical Society of America*, 50(4), 1166-1179.

Lindblom, B., Öhman, S. & Risberg, A. (1960). Evaluation of spectrographic data sampling techniques. *STL-QPSR* 1(1), 11-13.

Lindblom, B. (1961). Sona-graph measurements. *STL-QPSR* 2(3), 3-5.

Livijn, P. (2004). A comparison between four common ways of recording and storing speech: Implications for forensic phonetics. *In Proceedings of FONETIK 2004*, 104-107.

Loakes, D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins*. PhD Thesis, University of Melbourne, Australia.

Makhoul, J. I. & Wolf, J. J. (1972). *Linear Prediction and the Spectral Analysis of Speech*. AD-749066, BBN Report No. 2304, Bolt Beranek and Newman Inc., Cambridge, Mass.

Makhoul, J. (1975). Spectral Linear Prediction: Properties and Applications. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(3), 283-296.

Markel, J. D. (1972). Digital inverse filtering - a new tool for formant trajectory estimation. *IEEE Transactions on Audio and Electroacoustics*, 20(2), 129-137.

Markel, J. D. & Gray Jr., A. H. (1976). *Linear Prediction of Speech*. Berlin, Springer.

MathWorks, The (2007). *MATLAB* (Version 7.4.0) [Computer program]

McDougall, K. (2005). *The Role of Formant Dynamics in Determining Speaker Identity*. PhD Thesis. University of Cambridge.

McDougall, K. (2011). Acoustic correlates of perceived voice similarity: a comparison of two accents of English. Presented at IAFPA 20th Annual Conference, Vienna, Austria.

McDougall, K. (2013). Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech, Language and the Law*, 20(2), 163-172.

Mehta, D. D. (2011). Email to F Clermont re. VTR database, 25th October 2011.

Mehta, D. D., Rudoy, D. & Wolfe, P. J. (2012). Kalman-based autoregressive moving average modelling and inference for formant and antiformant tracking. *Journal of the Acoustical Society of America*, 132(3), 1732-1746.

Monsen, R. B. & Engebretson, A. M. (1983). The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction. *Journal of Speech and Hearing Research*, 26(1), 89–97.

Morris, R. J. & Brown, W. S. (1996). Comparison of various automatic means for measuring mean fundamental frequency. *Journal of Voice*, 10(2), 159-165.

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Communication*, 53, 242-256.

Morrison, G. S., Lindh, J. & Curran, J. M. (2014). Likelihood ratio calculation for a disputed-utterance analysis with limited available data. *Speech Communication*, 58, 81-90.

Mustafa, K. & Bruce, I. C. (2006). Robust formant tracking for continuous speech with speaker variability. *IEEE Transactions on Audio, Speech and Language Processing*, 14(2), 435-444.

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: National Academies Press.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, 85(5), 2088-2113.

Nolan, F. & Oh, T. (1996). Identical twins, different voices. *Forensic Linguistics*, 3(1), 39-49.

Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, 2(2), 143-173.

Peterson, G. E. & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2), 175-184.

Qi, Y. & Bi, N. (1994). A simplified approximation of the four-parameter LF model of voice source. *Journal of the Acoustical Society of America*, 96(2), 1182-1185.

Özbek, I. Y. & Demirekler, M. (2008). Vocal tract resonances tracking based on voiced and unvoiced speech classification using dynamic programming and fixed interval Kalman smoother. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 4217-4220.

Reynold, D. A., Quatieri, T. F. & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10, 19-41.

Rose, P. (2002). *Forensic speaker identification*. London: Taylor-Francis Ltd.

Rose, P. & Morrison, G. S. (2009). A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law*, 16(1), 139-163.

Rudoy, D., Spendley, D. N. & Wolfe, P. J. (2007). Conditionally linear Gaussian models for estimating vocal tract resonances. *Proceedings of Interspeech 2007*, Antwerp, Belgium, 526-529.

Schiller, N. O. & Köster, O. (1995). Comparison of Four Widely Used F0-Analysis-Systems in the Forensic Domain. In A. Braun & J.-P. Köster (Eds.), *Studies in Forensic Phonetics*. Trier: WVT Wissenschaftlicher Verlag Trier, 146-158.

Sjölander, K. (1997). *The Snack Sound Toolkit* [Computer program]. Available http://www.speech.kth.se/snack/

Sjölander, K. (2004). *Snack v2.2.8 manual* [Online] KTH. Available at: http://www.speech.kth.se/snack/man/snack2.2/tcl-man.html#sformant [Accessed 28th June 2014]

Sjölander, K. & Beskow, J. (2006a). *WaveSurfer* [Computer program] Available http://www.speech.kth.se/wavesurfer/

Sjölander, K. & Beskow, J. (2006b). *WaveSurfer User Manual* [Online] KTH. Available at: http://www.speech.kth.se/wavesurfer/man.html [Accessed 28th June 2014]

Smit, T., Türckheim, F. & Mores, R. (2012). Fast and robust formant detection from LP data. *Speech Communication*, 54, 893-902.

Stevens, K. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Talkin, D. (1987). Speech formant trajectory estimation using dynamic programming with modulated transition costs. *Journal of the Acoustical Society of America*, 82, S55.

Traunmüller, H. & Eriksson, A. (1997). A method of measuring formant frequencies at high fundamental frequencies. *Proceedings of EuroSpeech'97,* vol.1, 477-480.

Tukey, J. W. (1970). *Exploratory Data Analysis*. Reading, Mass: Addison-Wesley Publishing Co.

Vallabha, G. K. & Tuller, B. (2002). Systematic errors in the formant analysis of steady-state vowels. *Speech Communication,* 38, 141-160.

Vallabha, G. K. & Tuller, B. (2004). Choice of Filter Order in LPC Analysis of Vowels. *From Sound to Sense,* MIT, 203-208.

van Son, R. J. (2005). A study of pitch, formant, and spectral estimation errors introduced by three lossy speech compression algorithms. *Acta acustica united with acustica,* 91(4), 771-778.

Vemula, N. R., Engebretson, A. M., Monsen, R. B. & Lauter, J. L. (1979). A speech microscope. *Journal of the Acoustical Society of America,* 65(1), S22.

Vermeulen, J. F. M. (2009). *'Beware of the distance': Evaluation of spectral measurements of synthetic vowels re-recorded at different distances*. MSc Dissertation, University of York.

Wells, J. C. (1982). *Accents of English* (3 vols.), Cambridge: Cambridge University Press.

Woehrling, C. & de Mareüil, P. B. (2007). Comparing Praat and Snack formant measurements on two large corpora of northern and southern French. *Proceedings of Interspeech 2007,* Antwerp, Belgium, 1006-1009.

Wood, S. (1989). The precision of formant frequency measurement from spectrograms and by linear prediction. *STL-QPSR*, 30(1), 91-94.

Yegnanarayana, B. & Reddy, R. (1979). A distance measure derived from the first derivative of linear prediction phase spectrum. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 744-747.

Zhang, C., Morrison, G. S., Enzinger, E. & Ochoa, F. (2012). Laboratory Report: Human-supervised and fully-automatic formant-trajectory measurement for forensic voice comparison – Female voices. *FVC, EE&T, UNSW Laboratory Report*.

Zue, V., Seneff, S. & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9(4), 351-356.

Court Cases

The Queen v Anthony O'Doherty 19/4/02 ref: NICB3173 Court of Criminal Appeal Northern Ireland.

Regina v Ronald Flynn and Joe Philip St John [2008] EWCA Crim 970 2nd May 2008.