FORMANT MEASUREMENT ERRORS FROM SYNTHETIC SPEECH

P Harrison University of York & J P French Associates, York, UK

1 INTRODUCTION

Formants are resonances within the vocal tract that are associated principally with vowel sounds. The measurement and subsequent analysis of formants is an important aspect of many areas of phonetics including socio-phonetics and forensic speech analysis. The most common task undertaken by forensic speech scientists involves the comparison of speakers in criminal recordings with reference recordings to determine whether the speech samples are consistent with having come from the same speaker.[1] Following a ruling in the Court of Appeal of Northern Ireland[2] it is now effectively compulsory for formant analysis to figure within such comparisons done within the UK.

However, the measurement and analysis of formants is susceptible to many different sources of variation and error which so far have been the subject of little research. These include various differences within and between speakers, variation across software packages and the measurement techniques employed, as well as the analysis settings used. One study of note does provide some insight into the errors produced by different analysis settings and briefly examines some of the effects for different speakers but the investigation and results are limited.[3] The research described below attempts to shed further light on the errors and measurement variation produced by using different analysis settings.

2 FORMANT MEASUREMENT PROCESSES

2.1 Measurement Methods

The measurement of formant frequencies can be undertaken in several different ways. The potentially most straightforward method involves the visual inspection of spectrograms, computer generated plots of speech energy across frequency over time. The formants appear on the plots as dark bands which correspond to the regions of highest energy, i.e. the resonances. A cursor can be placed in the centre of the band and the value read from the vertical frequency axis. The top-left panel of Figure 1 below shows a spectrogram in which the formants of the vowel in the word ‘he’ are visible. The time aligned waveform plot is in the bottom-left panel. Note that the formant with the lowest frequency is referred to as F1, the next lowest is F2 and so on. The vowel quality (whether it is, for example, an /a/ /o/ or /e/) is determined by the formant frequencies, with the greatest contribution to its perception provided by F1 and F2. This gives rise to a concept known as ‘vowel space’, which refers to a region in the F1-F2 plane where vowels can occur and their position in this space is related to the position of the tongue in mouth when producing a specific vowel sound.

A generally more accurate method of measuring formants is to examine a spectral slice or average spectrum across the region of interest and measure the relevant spectral peaks. The top-right panel of Figure 1 shows the average spectrum across the vowel represented in the spectrogram.

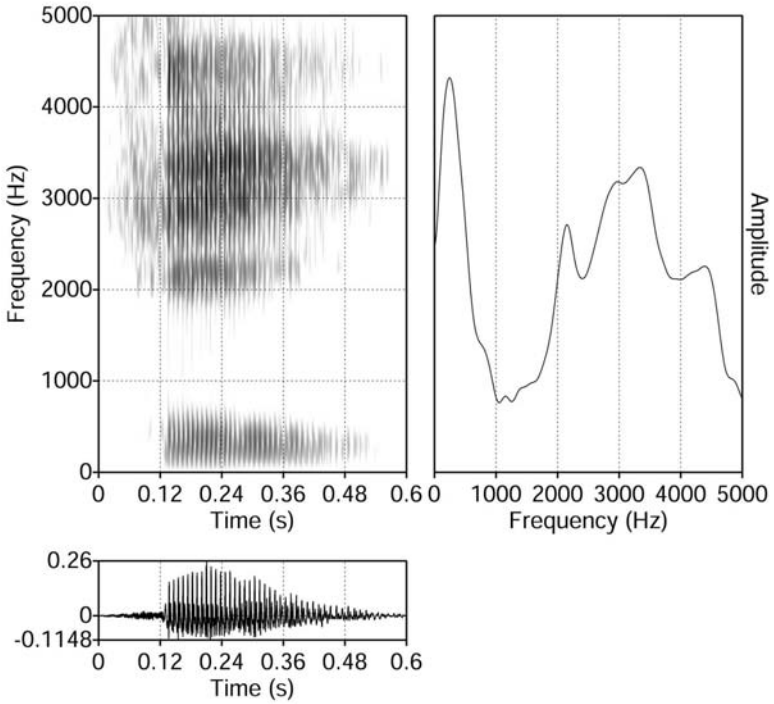

Figure 1 – Bottom-left panel: waveform of word ‘he’, top-left panel: spectrogram of word ‘he’, top-right panel: average spectrum of vowel in word ‘he’.

Probably the most commonly used method is to rely on automatic formant trackers which are found within speech analysis software. The trackers attempt to determine the resonance frequencies by way of an LPC (linear-predictive coding) analysis.[4] This technique assumes that speech is produced by a source-filter process and derives the resonance peaks of the filter, which correspond to the formant frequencies. The formant frequencies for each frame of analysed speech are usually overlaid on a spectrogram which allows a simple visual check to be made to determine if the tracker has produced potentially correct values.

2.2 Sources of Error and Variation

All three methods are susceptible to many forms of variation which can result in widely differing formant measurements both within and across methods. In all cases the analyst has many decisions to make which will affect the measurements.

2.2.1 Preliminary Decisions for the Analyst

The analyst must decide where in time within the vowel sound to make the measurements, since formant frequencies do not remain constant across a vowel. In the case of the spectrogram only a spot value can be obtained so the point chosen must be representative of the whole vowel. In the case of spectral slices and formant trackers an average measurement can be made over a selected portion of the vowel. However, changing the boundaries of the selected region by a small amount can result in significant differences in the measured formant frequencies.

When taking measurements from a spectrogram the cursor should be positioned at the point of highest energy within the visible dark band. This is not always easy to achieve and the highest point is not always located in the centre of the band. Small changes in the vertical location of the cursor

result in differences in the measured formants. This problem can be resolved by examining spectra since the peaks are generally more clearly defined but issues of double peaks and very broad peaks may still make the process problematic. In the case of formant trackers the results cannot be relied upon blindly. The analyst must decide whether the tracks shown actually correspond to formants or if they are erroneous.

2.2.2 Analysis Settings

A further decision that must also be made is what analysis settings to use. Formant frequencies cannot be obtained directly from the digitised speech signal or by an alternative measurement or transduction technique. Therefore, the formant frequencies obtained are effectively a by-product of the analysis process. As a consequence the analysis settings selected can significantly alter the measured values.

In the case of spectrograms and spectral slices, formants only become apparent when the spectrum is subject to smoothing. If the degree of smoothing is too small the formants are not well defined and too many peaks are present. Conversely, if the smoothing is too great then definition is lost and formants become merged. Changing the degree of smoothing by even a small amount can alter the location of the peaks corresponding to the formants.

The most important parameter for formant trackers, which causes the greatest degree of variation in measurements, is the LPC order.[3] This parameter determines the complexity of the model used to represent the filtering effect of the vocal tract. If the model is too simple formants will not be resolved, and if the model is over-specified then spurious formants will appear.

A further factor which is linked to the analysis settings is the specific implementation of the analysis method within the software. This is most significant for the formant trackers since hidden 'post-processing' decision-making can be involved in the process before the formant values are presented to the user. Moreover, this varies between software packages.

2.3 Understanding Variation & Errors

As can be seen from the discussion above there are many factors within the measurement process alone that can cause variability and errors in formant measurements. When considering formant values, for example in a forensic speaker comparison case, it is important to be aware of the extent of the potential variations. This research aims to further investigate these issues and provide a better understanding of the extent of the variability.

3 VARIATION WITHIN & BETWEEN FORMANT TRACKERS: PREVIOUS STUDY

An initial study by the author focused on the variability of measurements from automatic formant trackers caused by different analysis parameters across different programs.[5]

3.1 Methodology

The source material consisted of a word list of 30 words with a CV (consonant vowel) or CVC structure containing vowels from one of five vowel categories that represent the four extremes of the 'vowel space' plus a central vowel. The list was read three times by two male speakers and was recorded both with a microphone and at the distant end of a telephone line.

Since the main interest was the implications of the outcomes of the study in the forensic context, a survey of forensic analysts was undertaken to discover which programs were being most widely used in the forensic community. These turned out to be Praat,[7] Multispeech[8] and Wavesurfer.[9] Formant measurements were made and logged using the formant trackers within these programs whilst systematically altering the LPC order' frame/analysis width and pre-emphasis settings.

The study was concerned only with the variability in measurements caused by altering the analysis settings. The other decisions normally made by analysts such as where to take the measurements from and whether to reject certain values were removed by automating the process using the scripting capabilities of the software. The start and end points for the measurements were pre-determined for each token and then used each time by the scripts.

3.2 Results

The initial consideration of the raw measurements and the subsequent comparative analysis failed to show any general patterns across the different settings, vowel categories, speakers, recording conditions, programs and parameters. Certain trends were apparent but they were only present under specific circumstances. However, it was clear that the greatest source of variability was caused by altering the LPC order setting.

As mentioned above, there is no independent way to measure the formant values to be able to check the accuracy of the measurements. Therefore the measurements were simply compared across analysis settings and the differences were recorded. However, this provided no indication of the accuracy of the measurements but simply the differences present.

In an attempt to provide some indication of accuracy, the measurements were subsequently reanalysed. The measurement process employed was entirely automated and involved no intervention from an analyst. As a consequence, the results contained many measurements that would have been rejected as obviously inaccurate by an analyst if presented as formant tracks overlaid on a spectrogram. So the results were considered in terms of what percentage of the measurements fell within a 300 Hz acceptable band for each token. The band was determined by examining a spectrogram of each utterance and specifying an upper and lower limit for each formant. Considering the data in this way did reveal some patterning within the results showing that certain LPC orders systematically produced more reliable measurements (i.e. those that fell within the 300 Hz band) than others. However, the ultimate accuracy or error of these in-band measurements was still unknown.

4 INSIGHTS FROM SYNTHESISED SPEECH: CURRENT STUDY

The previous study highlighted the potential variability that could be caused by altering analysis settings, but to better understand and quantify the effects an alternative approach was required. The main problem with the study was the inability to determine the true formant values and therefore calculate the measurement errors.

The current study attempts to address this issue by using synthesised speech as the source material. This seemed to be the best way to know the true formant values since they are specified during the synthesis process. This approach allowed much greater control of the source material, including being able to specify the speaker's pitch, and to represent the entire vowel space instead of the limited regions covered by the five vowel categories previously employed.

4.1 Synthesis Method

The relatively straightforward 'source-filter' synthesis method was chosen since this allows the formant values to be explicitly stated. This synthesis method is derived from the source-filter theory of speech production which models speech production as consisting of a sound source, the vibration of the vocal folds, and a filter, the vocal track above the vocal folds. The pitch (fundamental frequency, F0) specifies the frequency of the vocal fold vibrations and the formants determine the peaks within the filter response. The bandwidths of the formants must also be specified in the synthesis process.

The F1 and F2 frequencies were chosen to represent a typical vowel space, with F1 ranging from 200 to 1000 Hz at 10 Hz intervals and F2 from 800 to 2500 Hz at 20 Hz intervals. Constraints were then applied to remove unrealistic formant combinations. The resulting 6,034 tokens representing the vowel space are shown in Figure 2 below. For each token F3, F4 and F5 together within bandwidths for each formant were generated using a combination of empirically derived formulae and specified values[9].

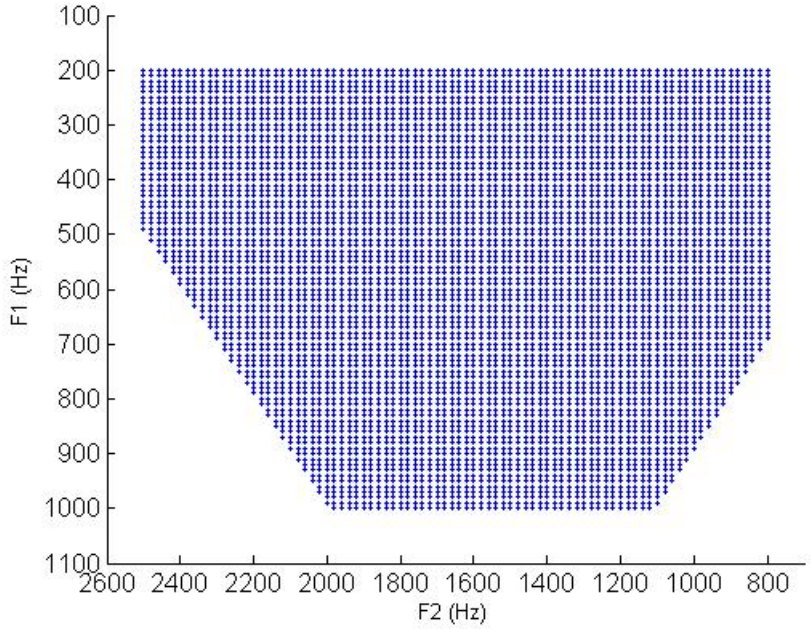

Figure 2 – F1 and F2 values for the 6,034 tokens used in the synthesis.

A simple pulse train was chosen to represent the glottal sound source. This was generated with fundamental frequencies from 70 to 190 Hz at 5 Hz intervals.

4.2 Measurement Process

The entire synthesis and analysis process was automated using the built-in scripting capabilities of Praat. A script generated all of the formant values and bandwidths for the vowel space. These values were saved to a text file which was then used repeatedly by the main script. This script did the majority of the work synthesising the speech using the formant values from the text file across the various pitches and measuring the formants using Praat's tracker and finally logging values. The measurements were made with LPC orders from 6 to 14. The final process for the script was to calculate the measurement error. This was calculated for each of the 6,034 tokens by subtracting the measured value from the real value specified in the text file. A positive error is the measured value showed it was higher than the true value and whilst a negative error showed the measured value was lower.

4.3 Error Analysis

The ideal way of examining the huge amount of generated measurements was to create three-dimensional plots of the various measurement errors over the F1-F2 vowel space. Figure 3 below is an example of the error surface for an LPC order 10 and a fundamental frequency of 100 Hz. Animations of the error surfaces were also created to observe the effects of altering pitch and LPC order. These techniques provided much greater insight to the data than that which could be achieved by simply looking at tables of figures.

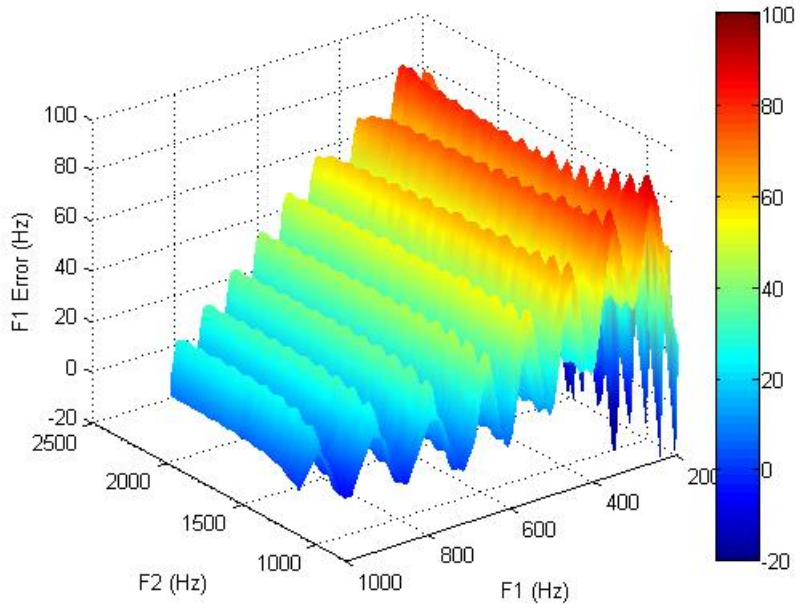

Figure 3 – 3D error surface for F1 measurements with LPC order of 10 and F0 of 100 Hz.

4.3.1 Pitch Influence

One of the most obvious features of the error plots is the cyclic or repetitive nature of the error surfaces. This basically shows that there are alternating regions of higher errors and lower errors within the vowel space. These are clearly related to the pitch of the speech because the location and regularity of the cycles changes as the pitch is altered.

4.3.2 Minimum Errors

By considering the results obtained across all of the LPC orders it was possible to determine the minimum errors that could be achieved for each formant. Again, periodicity was present in these results. The minimum errors were reassuringly small, which is encouraging and shows that the LPC analysis method can give relatively accurate formant measurements. The smaller errors become apparent in Figure 4 below when compared with Figure 3 above. The LPC order used to obtain these minimum errors was also plotted across the vowel space and again periodicity was present.

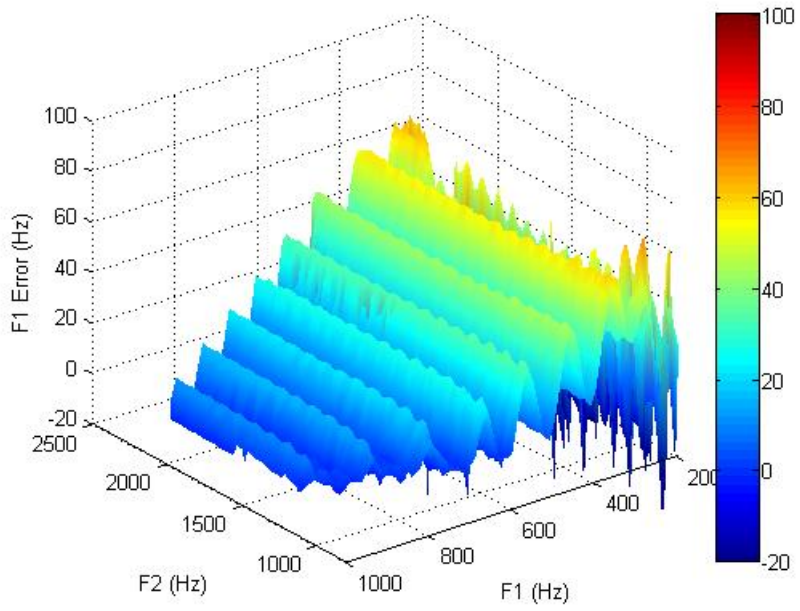

Figure 4 – 3D minimum error surface for F1 measurements across all LPC orders with F0 of 100 Hz

4.3.3 Bunching Effect

An alternative way of viewing the measurement errors is to look at where the measured formant values lie in the vowel space. The F1 and F2 values specified in the synthesis process were spaced at regular intervals in the space (this can be seen in Figure X above). However, when the measured F1 and F2 values are plotted they tend to bunch in certain areas of the space and distort its regular pattern. This is clearly visible in Figure 5 below.

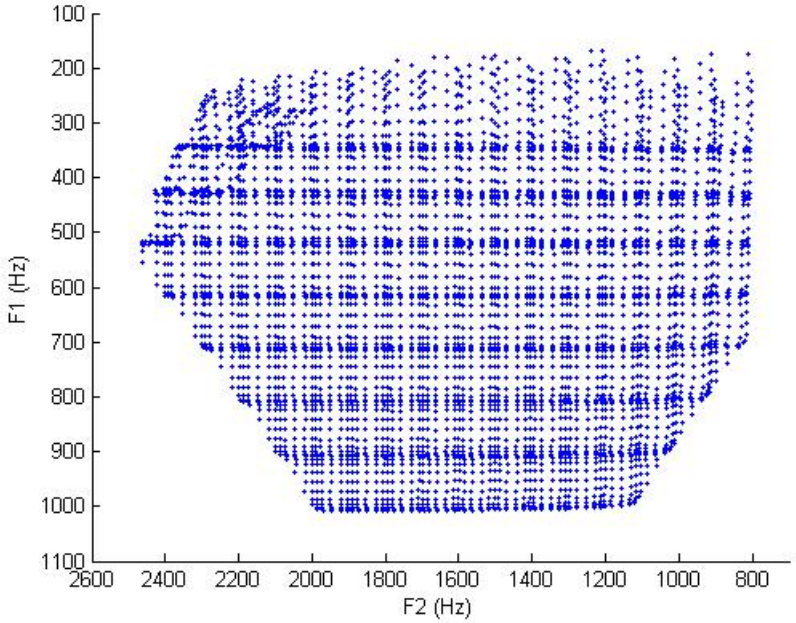

Figure 5 – F1 and F2 measurements obtained with LPC order 10 and F0 of 100 Hz showing the distortion of the vowel space

Again, regularity is present in the bunching of the measurements and is related to pitch but the apparent focus of each bunch is centred on a harmonic of the fundamental frequency.

4.4 Summary of Findings

The use of synthesised speech as the source material provided greater insight into the magnitude of the measurement errors than that which was achieved in the previous study. One of the most prominent features was the periodicity in the error surfaces that is apparently linked to pitch. Again, it is clear that the LPC order has a pronounced effect on the measurement errors. The minimum errors produced over all the LPC orders tested were reassuringly small. Perhaps the most interesting finding was the regular bunching of the measurements in the vowel space, however, the underlying reasons for this are not yet clear.

5 WAY FORWARD

The ultimate aim of this work is to develop a method for consistently obtaining the most accurate formant measurements possible. This may involve determining the ideal settings for specific vowel categories or speakers but there are clearly many more steps that need to be taken before this can be achieved. One important step is to understand the relationship between the periodicity in the errors and the fundamental frequency. Hopefully, it will then be possible to compensate for this effect and therefore reduce the errors.

Further testing needs to be done using more realistic and varied glottal sources since the one used so far is an idealised source. This will be helpful in determining, to some extent, how stable the errors may be across different speakers. Obviously, more work needs to be done with real speech but a great deal can still be learnt using synthetic material.

A further strand of investigation is the effect of higher LPC orders on the measurements. So far the analysis has concentrated on taking the results from the tracker at face value, i.e. the F2 returned by the tracker is considered as being F2. However, by over specifying the LPC order it may well be that other peaks actually produce a more accurate measure of the formants.

6 REFERENCES

1. French, J.P. & Harrison, P. ‘Investigative and evidential applications of forensic speech science’. In A. Heaton-Armstrong, E. Shepherd, G. Gudjonsson and D. Wolchover (eds.) Witness Testimony: Psychological, Investigative and Evidential Perspectives. Oxford: Oxford University Press (2006).
2. The Queen v Anthony O’Doherty 19/4/02 ref: NICB3173 Court of Criminal Appeal Northern Ireland.
3. Vallabha, G. K. & Tuller, B. ‘Systematic errors in the formant analysis of steady-state vowels’, Speech Communication 38: 141-160. (2002).
4. Markel, J.D. & Gray Jr., A.H. Linear Prediction of Speech. Berlin, Springer (1976).
5. Harrison, P.T. Variability of Formant Measurements. MA Thesis: University of York (2004).
6. http://www.praat.org
7. http://www.kayelemetrics.com/Product%20Info/3700/3700.htm
8. http://www.speech.kth.se/wavesurfer/
9. Fant, G. Vocal tract wall effects, losses, and resonance bandwidths. STL-QPSR, 13(2-3), 028-052. (1972).