# A PARAMETRIC MODEL OF AUSTRALIAN ENGLISH VOWELS IN FORMANT SPACE

Michael Barlow & Frantz Clermont
{spike,frantz}@cs.adfa.edu.au
School of Computer Science,
University College, University of NSW

ABSTRACT – This paper concerns the development of a parametric model for characterising the formant (F1-F3-F3) space of three sociolinguistic varieties (Broad, General and Cultivated) of spoken vowels in Australian English. The vowel-formant space is modelled as a quadratic surface, which captures the non-linearity in the F3 dimension and yields a parametric formulation for F3 as a weighted combination of F1 and F2. Differences between the surfaces, together with their application for prediction/classification on a per-speaker basis provide a holistic quantification of the differences between the varieties of Australian English.

## INTRODUCTION

In connection with a long-range goal of defining vowel categories in the acoustic domain, Broad and Wakita (1977) and Broad (1981) proposed a bi-planar model for accommodating phonetic-speaker interactions that extend beyond the traditional F1-F2 plane into the F3 dimension. Although a two-plane (one for the back and one for the front vowels) formulation was shown to provide a good fit to their formant data, no claim is made by these authors about the uniqueness of the mathematical formulation. Nonetheless, the resulting 3-D representation has yielded some unprecedented insights into between-speaker differences and their expected manifestations within the vowel-formant space. In particular, these studies have underscored the effectiveness of a vowel-aggregate approach to speaker characterisation and the importance of F3 in elucidating vowel-speaker interactions.

Following Broad's accommodating representation of lower and upper formants, we sought to characterise the formant space of the three sociolinguistic varieties, or idiolects as per Mitchell and Delbridge (1967) and Bernard (1967), of spoken vowels in Australian English. It was found, however, in a pilot investigation that the potential discontinuity along the two-plane intersection can render unreliable the visualisation or the comparison of surfaces across individual speakers or groups of speakers. Consequently, a single quadratic surface is proposed with the distinct advantage that it does not depend on any *a priori* partitioning of the formant space as it would be required in the two-plane case.

The next sections of the paper detail the corpus of vowel formants employed, the quadratic formulation adopted and the model parameters obtained via regression analysis. Vocal-tract (VT) length normalisation is also introduced in an attempt to secure some homogeneity amongst speakers and to reduce unnecessary biases in the quadratic fits. Separate surfaces are finally derived for the three idiolects, followed by a quantitative description of between-surface differences.

## VOWEL-FORMANT CORPUS

The corpus used for this study is based on Bernard's (1967) spectrographic measurements of the Australian English vowels, and on the subsequent restoration (Clermont, 1996) that yielded a 36-speaker set of formant frequencies for the vowels' respective steady-state frame. There are 14 Broad, 11 General and 11 Cultivated, adult-male speakers, whose {F1, F2, F3} -patterns are *completely* specified for each of the 11 monophthongal vowels uttered in citation and in a carrier phrase.

## PARAMETRIC MODEL

The formulation used for our quadratic surfaces (see Equation 1 below) is similar to that which Kazuya and Yoshizawa (1992) applied to the Japanese vowels, and affords the prediction of F3 from weighted combinations of the F1 and F2 values. Using the entire corpus described above, regression analysis was first employed to solve for the parameters of the equation, yielding weight values as given in Equation 2 and a surface such as that which is shown in Figure 1. The resulting model was found to predict the F3 data with a high degree of accuracy. The global RMS error between predicted F3 and measured F3 was 38 Hz, as per Equation 2 and Figure 1. The per-vowel, signed differences listed in Table 1 indicate that only the vowels in "herd" and "who'd" are the most distant from the fitted surface.

$$F_3 = \alpha_0 + \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_1^2 + \alpha_4 F_2^2 + \alpha_5 F_1 F_2 \qquad (1)$$

$$F_3 = 2417 + 1.322F_1 - 0.747F_2 - 4.718x10^{-4} F_1^2 + 3.644x10^{-4} F_2^2 - 2.993x10^{-4} F_1 F_2 \qquad (2)$$

| Vowel | Actual | Predicted | Difference |
|---|---|---|---|
| heed | 2734 | 2706 | 28 |
| hid | 2682 | 2681 | 1 |
| head | 2603 | 2614 | -11 |
| had | 2580 | 2571 | 9 |
| hard | 2490 | 2492 | -2 |
| hud | 2499 | 2494 | 5 |
| hod | 2452 | 2481 | -29 |
| hoard | 2434 | 2426 | 8 |
| hood | 2383 | 2384 | -1 |
| who'd | 2320 | 2385 | -65 |
| herd | 2518 | 2425 | 94 |

Table 1: Per-vowel difference between mean F3 for the entire speaker population and that predicted by equation 2.

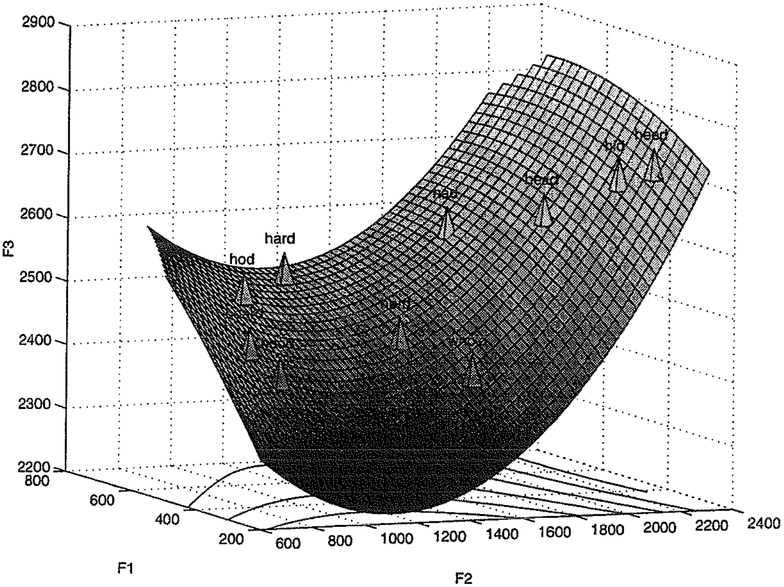

Figure 1: One view of the quadratic surface in F3 as a function of F1 and F2 for 36 adult-male speakers of Bernard's vowel-formant corpus for Australian English; the mean formant values for each vowel are represented as "pyramids" on the surface. This is the plane defined by Equation 2.

## VT LENGTH NORMALISATION

A well-known influence (Fant, 1960; Stevens, 1971) upon mean formant values is the speaker's VT length, such that mean formant values are inversely proportional to VT length. Clearly, variations in length will influence the coefficients of the plane equation and hence the generality of the formula. For a large population of speakers the influence of differing or unusual lengths may be assumed to normalise out through an averaging process. By contrast, for smaller populations, variations in VT length could be a significant factor and therefore, prior to regression analysis, an attempt was made to normalise the formant values using length-derived scaling factors.

The normalisation itself was performed by first estimating the VT length for each speaker and for the entire population using a formant-dependent formula given by Paige and Zue (1970). Individual speakers' VT lengths were then normalised to that of the entire population by scalar multiplication, and their formant values altered correspondingly. It is worth noting that, in order to estimate the mean VT length for an individual speaker, only the vowels in "had", "hard", "hud", and "heard" were used (Mokhtari, 1998), owing to their relative immunity to lip rounding and larynx height variations. The VT lengths obtained on a per-speaker basis, were found to range between 14.92cm and 17.15cm, with a population mean of 15.71cm, thus yielding scaling factors ranging from 0.916 to 1.053.

Equation 3 describes the fitted surface obtained when every speaker's VT length is normalised to a mean of 15.71cm. By contrasting the coefficients of Equation 3 with those of Equation 2 (no normalisation) it is found that there is very little change in weight values, with the exception of the coefficient for $F_1^2$ whose magnitude is nearly doubled.

$$F_3 = 2183 + 2.041F_1 - 0.671F_2 - 1.068x10^{-3}F_1^2 + 3.529x10^{-4}F_2^2 - 3.683x10^{-4}F_1F_2 \qquad (3)$$

Per-vowel and total RMS error for the planes defined by equation 2 (no normalisation) and equation 3 (VT length normalisation) were then calculated and contrasted as seen in Table 2. As can be observed from the table the approach does reduce overall (RMS) error by approximately 10%, with the most significant improvement being in the vowel "herd", but goodness of fit between the two approaches is clearly vowel dependent. For all hyper-planes modelled (idiolects and entire population), the RMS error between the plane and data was found to be smaller after VT length normalisation (Broad: from 29.8Hz to 26.8Hz, General: from 43.4Hz to 41.3Hz, Cultivated: from 49.7Hz to 49.0Hz) and hence the approach of VT length normalisation has been adopted for all subsequent sections.

| Vowel | Difference (Actual – Predicted) No VT Length Normalisation | Difference (Actual-Predicted) VT Length Normalisation |
|---|---|---|
| heed | 28 | 31 |
| hid | 1 | -5 |
| head | -11 | -27 |
| had | 9 | 4 |
| hard | -2 | 10 |
| hud | 5 | 17 |
| hod | -29 | -36 |
| hoard | 8 | 7 |
| hood | -1 | 3 |
| who'd | -65 | -63 |
| herd | 94 | 79 |
| RMS Error | 38 | 35 |

Table 2. Contrast of plane fit on a per-vowel basis between the 'raw' formant data (equation 2) and that normalised to have a vocal tract length of 15.71cm for all speakers (equation 3).

## VARIETIES OF AUSTRALIAN ENGLISH

Beyond its fitting adequacy, the quadratic-surface representation of the formant space implicates aggregates of vowels by definition and, therefore, could be interpreted to embody the accepted notion that the so-called idiolectal differences reflect group rather individual tendencies amongst speakers of a given language. That is, such a representation should be expected to yield useful insights of a holistic nature. In pursuit of this contention, the modelling paradigm described earlier was separately applied to the vowel-formant data for the 3 idiolects of Australian English. As a pre-processing step, VT length normalisation was attempted in order to eliminate some of the grosser speaker differences in the data. Following VT length normalisation, separate surfaces were fitted for the Broad (14), General (11) and Cultivated (11) speakers of the database. Table 3 shows the equations of those three planes.

| Variety | Intercept | $F_1$ | $F_2$ | $F_1^2$ | $F_2^2$ | $F_1.F_2$ |
|---|---|---|---|---|---|---|
| Broad | 2280 | 1.866 | -0.707 | $-9.095\times10^{-4}$ | $3.463\times10^{-4}$ | $-3.362\times10^{-4}$ |
| General | 2047 | 2.758 | -0.675 | $-2.213\times10^{-3}$ | $3.248\times10^{-4}$ | $-1.542\times10^{-4}$ |
| Cultivated | 2356 | 1.335 | -0.778 | $-1.174\times10^{-4}$ | $4.222\times10^{-4}$ | $-4.779\times10^{-4}$ |

Table 3: Coefficient values for the equations of the planes based on predicting F3 as a function of F1, and F2. Planes are calculated for the three different varieties (idiolects) of Aust. English subsequent to every speaker's formant values being normalised on the basis of mean vocal tract length.

Differences between surfaces were then examined as shown in Tables 4, 5 and 6. Table 4 represents the contrast of the planes using all values in the database. The RMS error between the actual F3 and that predicted by the three equations was calculated on a per-speaker basis. Those values were then grouped on the basis of idiolect, as well as being employed in a pseudo classification task in which the idiolect of the speaker was "selected" as being that which corresponded to the plane with the minimum RMS. A "leave-one-out" strategy was employed in all cases (Tables 4, 5, and 6): the data being "predicted" by the equations was not employed in the regression analysis which derived the equation coefficients. The results found in Table 4 once again illustrate the aggregate nature of the Australian English idiolects: when taken together the speakers of a particular idiolect are more similar than those of different idiolects. However, individually, there is a high degree of variability within an idiolect.

| | | Compared-To / Classified-As | | |
|---|---|---|---|---|
| | | Broad | General | Cultivated |
| Originating Speaker's Idiolect | Broad | 154.8 [4] | 157.3 [5] | 157.2 [5] |
| | General | 146.8 [4] | 145.4 [5] | 157.7 [2] |
| | Cultivated | 166.5 [5] | 178.0 [1] | 166.4 [5] |

Table 4: Combined distance-table and confusion-matrix resulting from using the equations found in Table 3 for scoring and classification on the basis of all data. The first figure in each cell represents the mean (across all speakers of an idiolect) RMS error between speaker's F3 and that predicted by the planes of Table 3. The bracketed figures represent a confusion-matrix based on classifying the speaker as belonging to the idiolect that corresponded to the plane with minimum RMS error. All figures correspond to a "leave-one-out" strategy in which the originating speaker's utterances were not employed in calculating the plane coefficients.

| | | Compared-To | | |
|---|---|---|---|---|
| | | Broad | General | Cultivated |
| Originating Speaker's Idiolect | Broad | 130.6 | 127.8 | 136.8 |
| | General | 111.2 | 122.8 | 128.2 |
| | Cultivated | 151.6 | 157.4 | 147.5 |

Table 5: Distance table resulting from using the equations found in Table 3 for scoring on the basis of front vowels alone. The values represent the mean (across all speakers of an idiolect) RMS error between speaker's F3 and that predicted by the planes of Table 3.

Tables 5 and 6 show a more detailed analysis of the between-surface differences: Table 5 shows differences on the basis of front vowels alone, while Table 6 shows that for the back vowels alone. Distinct differences can be seen from the values with Cultivated speakers being clearly differentiable in the front-vowel regions of the surfaces, while General speakers are clearly differentiable in the back-vowel regions. Classification accuracies range from 38.9% for all data, through 41.7% for the front vowels alone, to 58.3% for the back vowels alone.

|  |  | Compared-To |  |  |
|---|---|---|---|---|
|  |  | Broad | General | Cultivated |
| Originating Speaker's Idiolect | Broad | 173.3 | 181.7 | 173.5 |
|  | General | 162.9 | 151.6 | 172.0 |
|  | Cultivated | 154.9 | 177.8 | 155.3 |

Table 6: Distance table resulting from using the equations found in Table 3 for scoring on the basis of back vowels alone. The values represents the mean (across all speakers of an idiolect) RMS error between speaker's F3 and that predicted by the planes of Table 3.

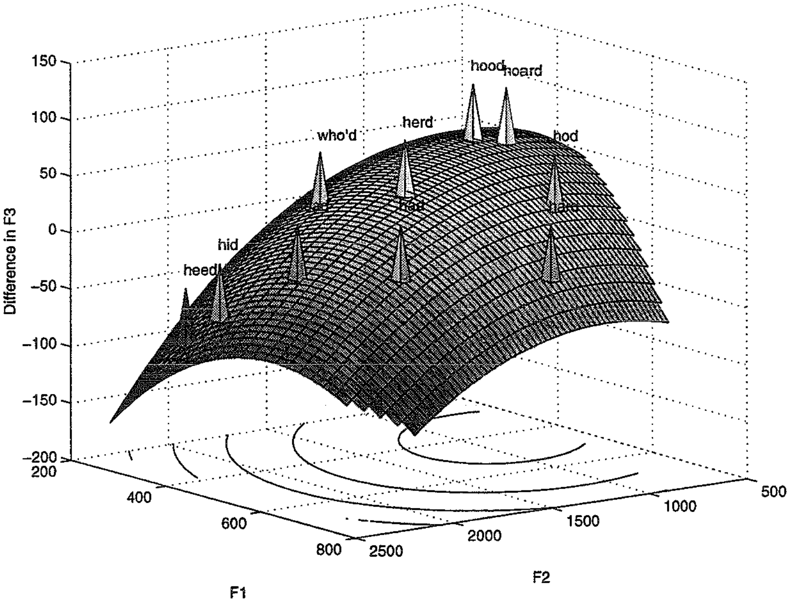

Figure 2: Difference in hyper-planes between that of Broad and Cultivated speakers. The "difference surface" represents the main region of F3 (and by implication F1 and F2) differences between Broad and Cultivated idiolects. The indicative pyramids correspond to the mean F1, and F2 locations of the monophthongs for the two varieties.

Differences between paired surfaces (for instance Figure 2 illustrates the difference between Broad and Cultivated idiolects) were also calculated and analysed. As expected the "difference surfaces" showed regions of formant-space in which significant differences were found between idiolects (such as the high-front vowels shown in Figure 2). The difference-planes not only provide an alternate to the numeric representation of tables 4, 5, and 6 as well as confirm some known phenomena but arguably give a more complete view of similarity and disparity between idiolects.

## DISCUSSION

We have described and evaluated a parametric model for the Australian English vowels, which yields a quadratic-surface representation of their {F1-F2-F3}-patterns. The model parameters are obtainable via regression analysis and have been found to provide a very good fit to the spectrographic measurements made by Bernard. Applying the methodology to the three varieties if Australian English it was found that differences between the models (planes) not only confirmed well-known phenomena about idiolectal differences but arguably gave a more complete, holistic view of said differences.

An alternate, dynamic three dimensional view (Barlow & Clermont 2000) showing the animated transition between the three planes defined in Table 3 (together with the underlying formant data itself) can be found at http://www.cs.adfa.edu.au/~spike/Research/VisualSpeech/index.html#aPlane.

Considerable potential remains for further investigation and application of the methodology. The current model is built (and tested) using the steady-state portions of the eleven monophthongs of Australian English. Incorporation of dynamic sounds include onset, offset and diphthongs would doubtless lead to a more accurate and descriptive model of the varieties of Australian English.

Examining the goodness of fit of the models on a per vowel basis it is clear that the models do not accurately describe certain sub-regions of formant-space; in particular that region about the vowels "who'd" and "herd" (low-F1 and mid-F2). Clearly there is scope for examining alternate formulations of the parametric model and interpreting the significance of those better fitting models.

The true power of the model for both prediction and classification remains to be explored. Can the model be used for speaker clustering on the basis of differences between the planes; can the planes be used to predict (and hence, perhaps adopt models) as yet unheard phones for a speaker on the basis of those already processed; how does the model generalise to dialect and language differences and can it be used for such tasks as language identification? These are all, as yet, unanswered questions.

## REFERENCES

Broad, D.J., and Wakita H., (1977), "Piecewise-planar representation of vowel formant frequencies", J. Acoust. Soc. Am. 62(6), 1467-1473.

Broad, D.J., (1981), "Piecewise-planar vowel-formant distributions across speakers", J. Acoust. Soc. Am. 69(5), 1423-1429.

Bernard, J.R.L., (1967), "Some measurements of some sounds of Australian English", PhD Thesis, Sydney University.

Clermont, F., (1996), "Multi-speaker formant data on the Australian English vowels: A tribute to J.R.L. Bernard's (1967) pioneering research", Proc. 6th Australian Int. Conf. Speech Science and Technology, Adelaide, Australia, 145-150.

Fant G., (1960), *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands).

Kasuya, H., and Yoshizawa, S., (1992), "Geometric Representation of Speaker Individualities in Formant Space and its Application to Speech Synthesis", Proc. International Congress on Acoustics, Beijing, China, G3-10.

Mitchell, A.G. & Delbridge, A. (1965), *The Speech of Australian Adolescents*, Angus & Robertson, Sydney.

Mokhtari, P. (1998), "An Acoustic-Phonetic and Articulatory Study of Speech-Speaker Dichotomy", PhD Thesis, The University of New South Wales.

Paige, A., and Zue V.W. (1970), "Calculation of Vocal-Tract Length", IEEE Trans. on Audio and Electroacoustics 18, 268-170.

Stevens K.N., (1971), "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds", Proc. Seventh Int. Congr. Phonetic Sciences: 206-232.