THE DIVISION OF THE PERCEPTUAL VOWEL PLANE FOR DIFFERENT ACCENTS OF ENGLISH AND THE CHARACTERISTIC SEPARATION REQUIRED TO DISTINGUISH VOWELS

Ahmed Ghonim, Jeremy Lim, John Smith and Joe Wolfe
School of Physics, The University of New South Wales, Sydney NSW 2052

The results of an on-line study of vowel recognition by English speakers are analysed. A relatively unused region of the perceptual vowel plane is identified at about ($F2$, $F1$) = (1800 Hz, 350 Hz). The rest of the plane is divided among vowels in ways that differ somewhat for different countries and regions thereof. Vowel length is used in several cases to help distinguish vowels whose distributions overlap substantially in ($F2$, $F1$). When the fundamental frequency is higher, the values of $F1$ and $F2$ are also higher, though much less than proportionally. This is consistent with the observation that women's vocal tracts are usually shorter than men's. The characteristic separations required to distinguish vowels in the ($F2$, $F1$) plane were 115 Hz and 292 Hz in the $F1$ and $F2$ directions respectively, with similar values in different countries.

INTRODUCTION

This paper analyses results collected by an on-line study of vowel recognition. Its aims are to compare accents of English speakers in different provinces and countries by identifying the regions of the perceptual vowel plane that correspond to a given vowel. It also aims to quantify how far an intended vowel may be displaced on the vowel plane from its mean position before it ceases to be recognised. These aims could, in principle, be achieved in a laboratory study. On a large scale, however, such a study would be laborious and expensive. The advantage of this on-line study is that it is automated and that, following its launch five years ago, it has had large scale and wide-ranging international participation.

The method of the study was reported in detail by Ghonim et al. [1], where some preliminary results were reported. Briefly, the survey has a large set of synthesised sounds of the form h[vowel]d, chosen because nearly all such combinations are real English words. On-line volunteers listen to a synthesised sound and choose, from a list on their screen, the h[vowel]d word they think the sound most resembled, or else judge it unrecognisable. Their choice and the parameters used to synthesise that vowel are then recorded in a database. They then progress to the next sound. At the start of a session, each respondent gives information about their native language, their regions of birth and residence, their gender, age and some other details about their linguistic history and environments.

Much of the phonemic information in the vowels of English is contained in the first two formant frequencies, $F1$ and $F2$. These formants are broad peaks in the spectral envelope produced by the first two resonances in the vocal tract [2,3].

The study by Ghonim et al. [1] uses a synthesis method developed by the authors for the purpose [1]. It samples the vowel plane in 50 Hz steps between the boundaries shown in Figure 1. Two other parameters are varied: the vowels can have two different lengths (t = 120 and 260 ms, hereafter 'short' and 'long') and two different initial fundamental frequencies (f_0 = 126 and 260 Hz, hereafter 'low' and 'high'). The number of sounds identified by each subject depends on their good will and patience. However, over all subjects, points in the space ($F2$, $F1$, t, f_0) are presented in a pseudo-random order so that each point has a similar number of occurrences.

The present paper analyses the results from this study, shows how the perceptual ($F2$, $F1$) plane is divided among vowels and unrecognised regions, and how this division depends on vowel length and f_0. It then uses the data to determine how the chance of identifying a sound as having a particular vowel varies as a function of the distance from the sound having the mean values ($\overline{F2}$, $\overline{F1}$) for that vowel. Using this function for each vowel, the characteristic distances on the perceptual ($F2$, $F1$) plane that are required to distinguish different vowels are calculated. The vowel plane is usually plotted as ($F2$, $F1$) with the direction of the conventional axes reversed; this is to preserve a similarity to the phoneticians' plot of mouth opening versus position of the tongue constriction. This tradition has been followed in this work.

RESULTS

The data set

40.5% of respondents used headphones and 59.5% loudspeakers. The frequency range of $F1$ often lies in a range over which the gain of radiating loudspeakers varies strongly with frequency, so it was of interest to see whether this made a difference to results. Averaged over all vowels, the shift in mean frequency of ($F2$, $F1$) from headphones to loudspeakers was (5.5 Hz, –7.6 Hz) for the survey population. This shift is insignificant in comparison with the sampling interval on the ($F2$, $F1$) plane (±50 Hz) and consequently headphone and loudspeaker data are pooled in all the subsequent analysis.

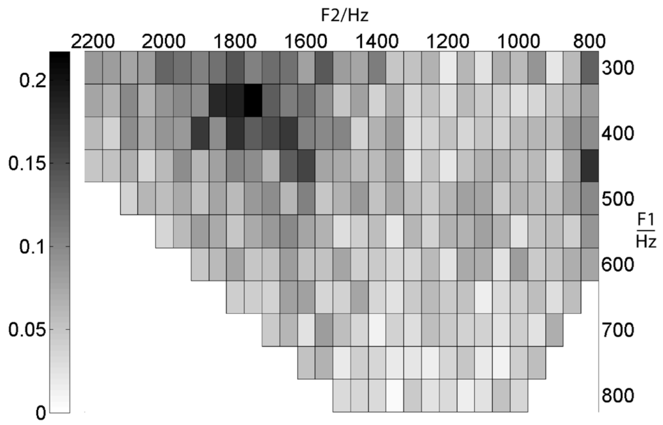


Figure 1. Distribution of unrecognised sounds on the perceptual (*F*2, *F*1) plane as a fraction of all choices by all respondents. The grey scale on the left indicates the fraction of sounds that were not recognised as any word.

Unrecognised sounds

The grey scale in Figure 1 shows the fraction of sounds that were not recognised as any of the listed words. Over the whole parameter space and for all respondents, the fraction of sounds that were not recognised as any of the words is 6.5%. In one area of the plane, near (1800 Hz, 350 Hz) or between 'heed' and 'who'd' in US, Australian or UK English, the proportion rises to 15-20%, suggesting that this area of the plane is not so much used in the accents of English most represented in this study, which are American, Australian and British. Other local areas of low recognition occur on the right, at very low values of *F*2.

Figure 2 shows the percentage of tokens over all of the parameter space that were recognised as each of the listed words by respondents born in the US (202 male respondents, 193 female), Australia (54 male, 49 female) and the UK (49 male, 18 female). They are grouped into words with monophthongs without the letter *r*, words containing the letter *r*, words that are often pronounced with diphthongs and those which were unrecognised. In what follows the effects of diphthongs and the letter *r* are discussed.

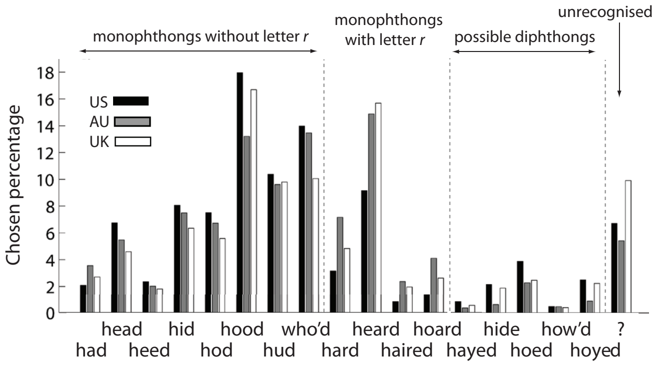


Figure 2. Percentage of words chosen by respondents born in US, AU and UK. '?' indicates that the vowel was unrecognised.

Diphthongs and *r*

The sound samples did not include any words synthesised with a diphthong. Nevertheless, the survey allows respondents to chose words that would, in most Australian speech, be pronounced as h[diphthong]d: *hayed*, *hide*, *hoed*, *how'd*, *hoyed*. They were included because it is conceivable that some of these words might be pronounced as monophthongs in some accents. In practice, few respondents chose these words: over all sounds, *hayed* (0.8%, 0.3%), *hide* (2.1%, 0.6%), *hoed* (3.9%, 2.2%), *how'd* (0.5%, 0.5%), *hoyed* (2.5%, 0.9%) where the two values are respectively for respondents born in the US and Australia. These results suggest that more Americans than Australians recognise these words as monophthongs.

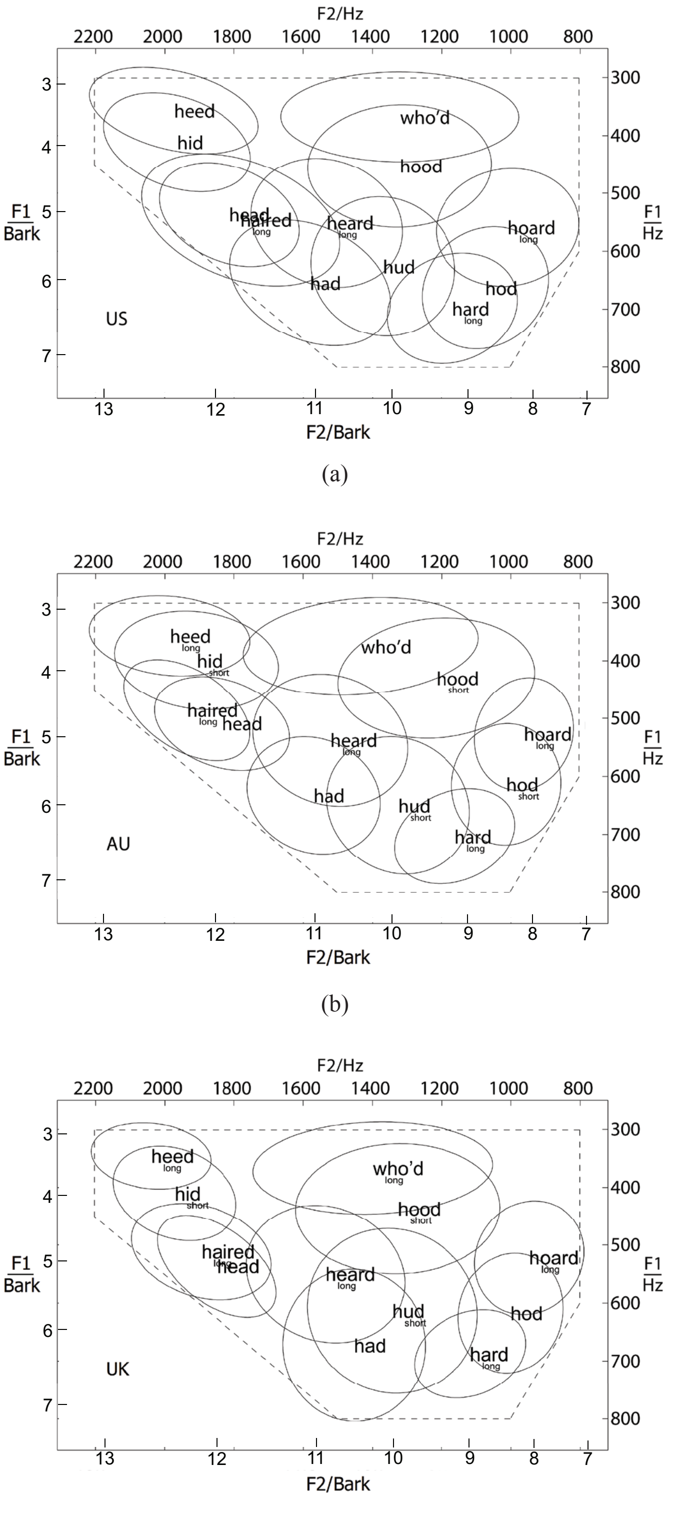
The sound samples did not include any rhotic *r* sounds. Nevertheless, the survey allows respondents to choose the words *haired*, *hard*, *heard* and *hoard*. In Australian English, and in some other varieties, these words are often pronounced without the rhotic *r* as h[vowel]d, but this is less frequent in the US. Figure 2 shows that each of these words was chosen by a higher proportion of respondents born in Australia than the proportion of residents born in the US. The proportions of Australians who chose these words when the vowel was long and short were: *haired* (98.1%, 1.9%), *hard* (93.4%, 6.6%), *heard* (92.1%, 7.9%) and *hoard* (96.8%, 3.2%). This is consistent with the observation that *r* in Australian English has the effect of lengthening the preceding vowel [4].

Distribution of vowels in different regions

Figure 3 shows the distribution of vowels on the perceptual (*F*2, *F*1) plane for respondents born in the US, Australia and the UK. For each group of respondents, the centre of each ellipse shows the mean values, the direction of the major axis is the line of regression and the semi-axes show the standard deviations in that direction and the direction at right angles to the line of regression. The gap between 'heed' and 'who'd' (mentioned above in the context of unrecognised vowels) is less noticeable in the Australian than in the American or UK data.

'Short' or 'long' printed below one of the words in Figure 3 means that more than 75% of the selections of that word were from the short or long sound samples, respectively. (On the average, each respondent should have received equal numbers of short and long sounds). This difference explains the overlap of some of the vowels: in the Australian and UK data, the distinction between *heed* and *hid* is largely made by vowel length, rather than position in the perceptual (*F*2, *F*1) plane. It is also important in distinctions between *hud* and *hard*, *hod* and *hoard*, and *head* and *haired*. This effect is smaller in the US data.

If we consider only the words that do not contain r and that are not possible diphthongs, Figure 2 shows that the words that are least chosen by both US and Australia respondents are *had* (2.1%, 3.5% respectively) and *heed* (2.4%, 2.0%). For *heed*, the alternative choice in that region is either *hid* or 'unrecognised'. For *had*, there is no nearby peak in the 'unrecognised' choice, but there is competition for much of that region of vowel space from several other vowels. Conversely, Figure 3 shows that there are few vowels at the top right of the plane, so *hood* and



(a)

(b)

(c)

Figure 3. Vowel distribution and standard deviation ellipses for (a) US, (b) Australian and (c) UK respondents to this survey. The dashed line shows the limit of the perceptual ($F2$, $F1$) plane sampled in these two dimensions.

who'd have high values in Figure 2. The word with the neutral vowel, *heard*, is also chosen frequently.

Figure 4 compares the results of New South Wales (38 respondents) and Queensland (18 respondents). For these populations, seven of the vowels showed differences significant at the 95% level (Figure 4). Averaged over all vowels, $F1$ was larger for NSW by 17 Hz, and $F2$ by 32 Hz. Both $F1$ and $F2$ increase with increasing mouth aperture so, on its own, this suggests that Queenslanders, on average, open their mouths less widely than New South Welshmen. However, the Queensland means are usually closer to the edges of the vowel plane, suggesting that Queenslanders use more of the vowel plane and thus have larger differences between vowels. It should be remembered, however, that 32 Hz is still smaller than the separation between harmonics in this study.

Among the three US states with the largest number of respondents – California (47), New York (33) and Ohio (20) – the differences were smaller than those between New South Wales and Queensland. At the 95% level, significant differences were found for only three vowels between California and Ohio (*had*, *hod*, *who'd*), three vowels between New York and Ohio (*had*, *heed*, *who'd*) and two between California and New York (*had*, *heed*).

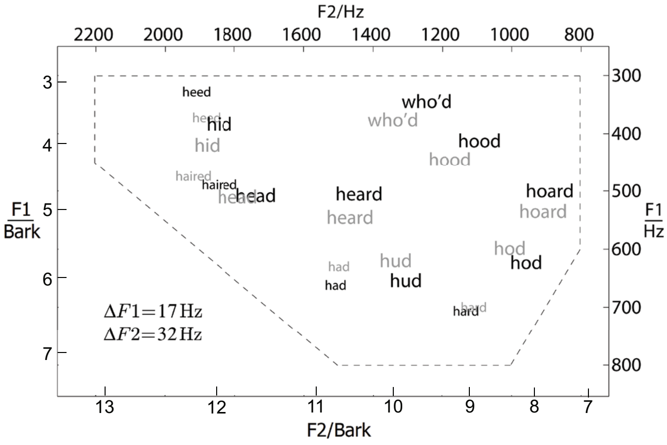


Figure 4. Shift in mean formant frequencies from Queensland (black) to New South Wales (grey). Those in large font are significantly different at the 95% level. For each word, $F1$ lies on the mid-line of the word, and $F2$ immediately to the left of the word.

Vowel length

In several cases, vowels whose ellipses overlap significantly when plotted as in Figure 3 were, in part, distinguished by vowel length. Table 1 plots, for each word and each of the US, UK and Australia data, the fraction of choices that were long vowels. Thus *hard* was usually chosen when the sound sample had a long vowel, which distinguished it from *hud* and *hod*, which are nearby on the vowel plane for all these countries. *heed* and *hid* are distinguished by length in all these countries, though the difference is slightly less in the US.

We looked for patterns in the displacement on the perceptual vowel plane between the long and short versions of the same chosen word. $F1$ increases with mouth aperture

Table 1. The percentage of choices that were for a long vowel, for each word and for each of three countries. Bold font highlights values above 75% or less than 25%, that is, words that are classed as long or short in Figure 3.

%long	had	head	heed	hid	hod	hood	hud	who'd	hard	heard	haired	hoard
US	61.0	41.9	74.5	27.9	54.3	29.1	26.2	**72.4**	**87.6**	**90.5**	**80.0**	**87.7**
AU	45.7	31.3	**80.9**	**15.8**	**18.4**	**15.4**	**5.3**	**63.6**	**93.4**	**92.1**	**98.1**	**96.8**
UK	55.9	29.5	**88.2**	**11.7**	28.1	**15.1**	**5.4**	**79.7**	**98.9**	**98.0**	**95.9**	**98.0**

and so, to a lesser extent, does $F2$. Perhaps sustained vowels give the speaker more time to open the mouth. If so, one would anticipate the longer vowels to be displaced down and to the left on the perceptual plane. There was no such effect, nor any other consistent pattern in the US, UK and Australian data, and the average shift for these pairs was only several Hz. These results differ from an earlier perceptual study [5], where shifts in $F1$ and $F2$ were recorded for vowel lengths similar to those studied here.

Dependence on f_0

The reason for including high and low f_0 was to simulate the difference between male and female voices. Acoustic measurements of the vocal tract resonances of young Australian men [6] and women [7] showed that the resonant frequencies used by women for a given vowel are typically higher than those used by men, which is traditionally explained by observing that women, on average, have shorter vocal tracts than men. Positive shifts on the perceptual ($F2$, $F1$) plane for synthetic vowels have been reported previously [8,9] so one might expect a similar result for formants in this perceptual study. Figure 5 shows the displacements of the vowels (low to high) for the Australian data. The displacements are positive in $F1$ and $F2$, as expected.

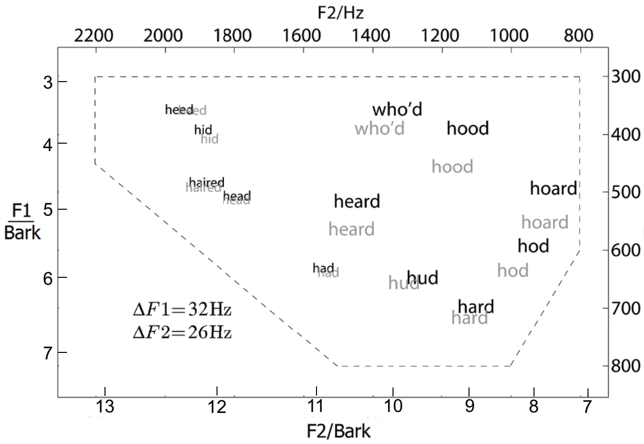


Figure 5. Shift in mean formant frequencies from the 'male voice' (f_0 = 126 Hz), printed in black to the 'female voice' (260 Hz, grey) in the Australian data. A large font indicates that the difference is significant at the 95% level.

Characteristic displacements for vowel recognition

A sound with values of $F1$ and $F2$ corresponding to the centre of one of the ellipses in Figure 3 has a high chance of being recognised as containing the vowel indicated on that ellipse: it has the mean values of $F1$ and $F2$ for that vowel identified by all respondents from that country. For sounds displaced significantly from that point, the chance of being thus identified falls. How far can a vowel 'stray' on the vowel plane before it ceases to be recognised? To answer this, the chance of being thus recognised as a function of distance on the vowel plane from its mean value was plotted. Distance on the perceptual vowel plane could be measured in Hz, but this would over-represent displacements in the $F2$ direction, because $F2$ is distributed over a larger range of frequencies. In a previous paper [10], a non-dimensional displacement d on the vowel plane was defined. The Pythagorean distance between two points a and b on the plane was scaled by the standard deviations σ_{F1} and σ_{F2} of all vowels in the $F1$ and $F2$ directions to give the dimensionless separation:

$$d = \sqrt{\frac{(F1_b - F1_a)^2}{\sigma^2_{F1}} + \frac{(F2_b - F2_a)^2}{\sigma^2_{F2}}} \qquad (1)$$

where σ_{Fi} is the standard deviation in Fi over all vowels, which in this case is σ_{F1} = 147 Hz and σ_{F2} = 374 Hz. So, for a particular vowel v, whose mean value on the recognition plane occurs at ($\overline{F2}$, $\overline{F1}$), the fraction f_v of vowels recognised as v is plotted as a function of the radial distance d from ($\overline{F2}$, $\overline{F1}$).

The ellipses in Figure 3 show that the spread of vowel recognition is large and that there is considerable overlap. It is therefore interesting to ask how much of this spread is due to variation among respondents and how much to variation in the choices made by each individual respondent.

Figure 6 shows $f(d)$ for the respondents born in Australia. It also shows $f(d)$ for one Australian-born respondent who had a relatively large number of sample responses, and thus gave reasonably good statistics. At $d = 0$, the rate of recognition by the single subject was about 60% while for the population it was about 25%. Of course, the plot shows that, even for one subject, a vowel occupies a finite area on the plane. For a large population, which may have and be familiar with different accents, the distribution for each vowel is larger than for an individual.

Dowd et al. [10] fitted both exponential ($a_0e^{-d/\lambda}$) and Gaussian functions ($b_0e^{-d^2/2\sigma^2}$) to $f(d)$, so we fit those functions here, to give two characteristic, non-dimensional distances, λ and σ respectively. In the present study, the Gaussian appears to be a rather better fit (Figure 6). For the Australian individual and the Australian sample data, the values of a_0 are respectively 0.63 and 0.27, b_0 are 0.56 and 0.23, values of λ are 0.94 and 1.00, while those of σ are 0.62 and 0.74 respectively. These values of σ correspond to 86 and 229 Hz in the $F1$ and $F2$ directions respectively for the individual, and 109 and 277 Hz for the population. The population values are surprisingly similar to those of [10], who reported 105 and 279 Hz. Direct comparison between them is not advised, however: in the Dowd et al. [10] study, we used acoustic measurements of the tract resonances, not formants, we used real human speech, not synthesis, and the language studied was French, not English.

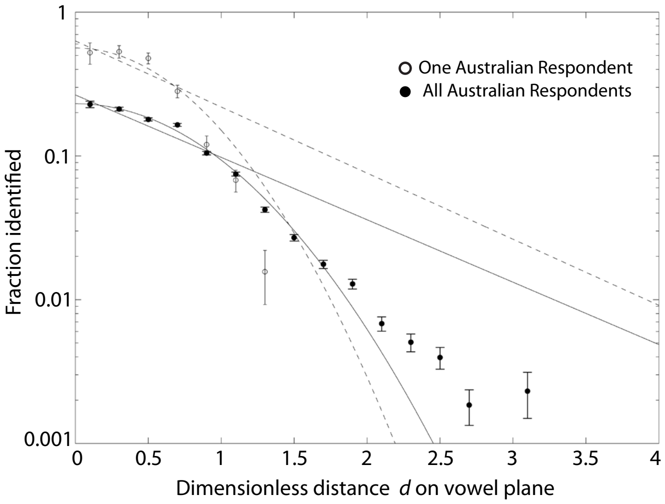


Figure 6. The fraction of sounds identified as a having a particular vowel plotted as a function of the dimensionless Pythagorean distance (*d*) from the mean position for that vowel. The data are averaged for all vowels. Solid lines and black points are for the Australian population. Dashed lines and grey points are for one Australian respondent with a large data set. The straight and curved lines indicate the results of an exponential or Gaussian fit respectively to the data. Error-weighted fits are used, hence points with large values of *d* that were chosen infrequently do not contribute strongly to the fit.

The values of σ from the Gaussian fits are listed in Table 2 for all respondents and for the five countries having the greatest numbers of respondents (Australia, US, UK, Canada and France). There is little variation among these.

Table 2. The characteristic distance required to distinguish vowels (Gaussian model). σ is the dimensionless separation defined by equation (1) and σ_1, σ_2 the separations in $F1$ and $F2$ respectively.

Population	AU	US	UK	CA	FR	ALL
σ	0.74	0.79	0.76	0.70	0.79	0.78
σ_1/Hz	109	116	112	103	116	115
σ_2/Hz	277	295	284	262	295	292

Future use

One possible use of the data gathered by this survey might be voice synthesis that is tailored for different regions. To obtain finer detail, it would be necessary to advertise the survey in the required geographical regions.

The survey [11] has run for only a few years, so it is too early to look for evidence of vowel drift with time. It would be interesting, however, to study changes on a time scale of decades, as suggested by Mannell [12]. The authors are prepared to make data available to other researchers, subject to conditions that include the anonymity of the data being met.

CONCLUSIONS

This survey quantifies the vowel plane for several countries and regions thereof. A relatively unused region of the vowel plane is identified at about ($F2$, $F1$) = (1800 Hz, 350 Hz). In several cases, vowel length helps distinguish vowels that overlap on the plane. The values of $F1$ and $F2$ rise slightly when the fundamental rises from typical women's to men's range. Using a Gaussian model for vowel distribution, the characteristic separations required to distinguish vowels in the ($F2$, $F1$) plane were respectively 115 Hz and 292 Hz in the $F1$ and $F2$ directions.

ACKNOWLEDGEMENTS

We thank the ARC for support and the many volunteer subjects who responded to the survey.

REFERENCES

[1] A. Ghonim, J. Smith and J. Wolfe, "An automated web technique for a large-scale study of perceived vowels in regional varieties of English", *Acoustics Australia* **38**, 152-155 (2010)

[2] G. Fant, *Acoustic theory of speech production*, Mouton, The Hague, 1960

[3] J. Clark, C. Yallop and J. Fletcher, *An introduction to phonetics and phonology,* Blackwell, 2007

[4] ibid. p439

[5] W.A. Ainsworth, "Duration as a cue in the recognition of synthetic vowels", *Journal of the Acoustical Society of America* **51**, 648-65 (1972)

[6] J. Epps, J.R. Smith and J. Wolfe, "A novel instrument to measure acoustic resonances of the vocal tract during speech", *Measurement Science & Technology* **8**, 1112-1121 (1997)

[7] T. Donaldson, D. Wang, J. Smith and J. Wolfe, "Vocal tract resonances: a preliminary study of sex differences for young Australians", *Acoustics Australia* **31**, 95-98 (2003)

[8] R.L. Miller, "Auditory tests with synthetic vowels", *Journal of the Acoustical Society of America* **25**, 114-121 (1953)

[9] W.A. Ainsworth, "Perception of synthesized isolated vowels and h-d words as a function of fundamental frequency", *Journal of the Acoustical Society of America* **49**, 1323-1324 (1971)

[10] A. Dowd, J.R. Smith and J. Wolfe, "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time", *Language & Speech* **41**, 1-20 (1998)

[11] A. Ghonim, *Sounds of world English* (2008) project. phys.unsw.edu.au/swe/main.php

[12] R.H. Mannell, "Perceptual vowel space for Australian English lax vowels: 1988 and 2004", P*roceedings of the 10th Australian International Conference on Speech Science and Technology*, Sydney, 8-10 December 2004, pp. 221-226