AN AUTOMATED WEB TECHNIQUE FOR A LARGE-SCALE STUDY OF PERCEIVED VOWELS IN REGIONAL VARIETIES OF ENGLISH

Ahmed Ghonim, John Smith and Joe Wolfe

School of Physics, The University of New South Wales, Sydney NSW 2052

J.Wolfe@unsw.edu.au

Because vowels in English are largely distinguished by the frequencies of their first two formants ($F1$, $F2$), the division of the ($F2$, $F1$) plane is an important and quantifiable component of accents. We report results of a web-based study into some of the many accents of English. Participants identified the vowel in h[vowel]d words produced by synthesis from a large set of possible values of $F1$, $F2$ and $F3$, using two different fundamental frequencies and two different durations. Compared to analysing spoken utterances, this approach has a number of obvious disadvantages, which we discuss. It has the significant advantages, however, of low cost, large scale and wide-ranging international participation. It is then possible to use the same experimental protocol to characterise the (perceptual) vowel plane of a substantial number of subjects and accents, thus allowing simple comparisons. From the large data base thus acquired, we present four examples of vowel maps for different Anglophone countries and regions therein. Knowledge of local variations in the perceptual ($F2$, $F1$) map, and the way in which these depend on fundamental frequency f_0, is not only of phonetic interest, but may be useful to those who use synthetic speech in automated communication systems.

INTRODUCTION

In Western languages, the vowels are chiefly distinguished by the frequencies of the low frequency formants, mainly the first two ($F1$, $F2$). The formants or peaks in the spectral envelope arise from acoustical resonances of the vocal tract, which increase the power of the radiated speech at frequencies near those of the resonances. The articulatory and acoustic origins of formants and their properties and roles in phonetics are important and well studied. Fine reviews are given by Fant [1] and Clark *et al*. [2].

The division of the ($F2$, $F1$) plane into vowels is one identifying feature of different accents and one that is readily and objectively quantified. There are many different regional and cultural accents of English, especially if one includes those of regions in which it is spoken as a foreign language. In principle, the different divisions could be determined by recording samples of speakers of each accent under similar conditions and analysing the recordings. This would, however, be difficult and expensive for a single research group. Collating the work of many groups is also a large task, and it could encounter variations in experimental technique.

Here we report an automated routine on a web site that uses synthetic speech to sample the vowel plane and to determine the perceptual vowel plane of volunteer subjects, rather than the produced vowel plane. It has gathered (and continues to gather) a large database of divisions of the vowel plane from regional varieties of English.

The perceptual division of the ($F2$, $F1$) plane is in principle different from the divisions in the space of produced vowels, but this does not make it less interesting. Indeed, in the field of synthesised speech, one is especially interested in how vowels produced with particular values of ($F2$, $F1$) will be perceived among the target listening group. Manell [3] has used perception of synthesised words to study vowel drift over time. Hay *et al*. [4] have used forced-choice perceptual studies of vowels to investigate the effects of age and social class.

The advantages of the method reported here are that it is automated and is available to volunteers around the world at times of their convenience. This has allowed us to accumulate a large and growing data set from about a thousand volunteer subjects. Because the data set is large, this paper includes just a few vowel planes as examples of regional variation, but leaves detailed analysis for other studies.

MATERIALS AND METHODS

Vowels and carrier words

The vowels are presented in the h[vowel]d context because, in English, all utterances thus produced are real words, with the exception of hud, whose pronunciation is reasonably obvious because of anticipated rhymes with the words bud, cud, dud, mud and sud.

The sounds produced in this study are all pure vowels rather than diphthongs. Subjects were, however, permitted to identify these pure vowel words as words that are usually spoken as either pure vowels or diphthongs and to identify sounds as one of the words h[written vowel]rd. This decision was made after some preliminary trials suggested that some respondents might decide that a long version of an utterance on the plane near ‘head’ sounded like ‘haired’, or identify a pure vowel with a word spoken in some regions using a diphthong. We could think of no reason to disallow such a choice: we were, after all, interested in their perception. We can justify this decision in retrospect: the h[written vowel]rd and h[diphthong]d words were indeed chosen by some respondents, though less often than the h[written vowel]d words. Conversely, it is possible that speakers of some variants of English might not identify the word 'hard' in this survey because they expect the word to contain a rolled 'r'.

Formant parameters

The ($F2$, $F1$) plane is sampled at a spacing of 50 Hz in both directions. This value was chosen as a compromise between resolution on the plane and the required number of samples. The choice of the boundaries for ($F2$, $F1$) was difficult. It varies among accents [2] and perhaps also according to measurement technique. We used values that include the limits shown in plots of ($F2$,$F1$) for spoken vowels, e.g. [2]. We set:

$$300 \text{ Hz} \le F1 \le 800 \text{ Hz},$$
$$800 \text{ Hz} \le F2 \le 2200 \text{ Hz}, \text{ and}$$
$$F1 \le F2 - 200 \text{ Hz} \quad \text{and} \quad F2 \le 3100 \text{ Hz} - 2F1$$

These boundaries are shown in Fig. 1. $F3$ was determined using the empirical relationship $F3 = 2100 \text{ Hz} + 0.42{*}F2$ that had been determined by fitting a linear regression to values of $F2$ and $F3$ collected from a range of sources. The bandwidth of all formants is set as a function of $F1$, $F2$ and $F3$ using the equations of Hawks and Miller [5].

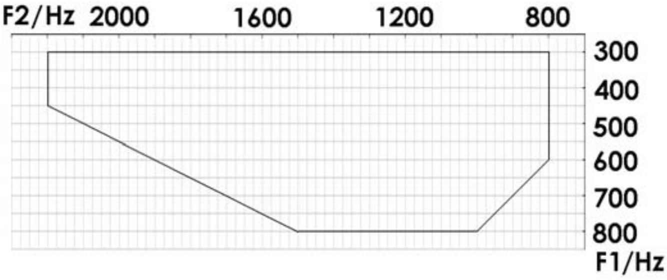

Figure 1. The chosen boundaries of the ($F2$, $F1$) plane investigated resemble those for speech. The plane is sampled at intervals of 50 Hz. The reversed axes are traditional in phonetics.

Jitter and shimmer were applied using the values of Minematsu *et al*. [6]. For each sampling of the vowel plane, tokens were synthesised with two values of initial f_0: 126 Hz and 260 Hz (hereafter 'low' and 'high') and two values of vowel duration: 120 and 260 ms (hereafter 'short' and 'long'). f_0(t) was decreased slightly (by 20 Hz) during each token. The limitation to two values of f_0 and duration was to limit the size of the data base in these two dimensions. Higher resolution of the effects of these parameters may be easily measured in studies that do not aim for such large data sets. Pragmatically, therefore, we chose values of f_0 that were very likely to be identified as man and woman, and durations likely to be identified as short or long vowels in isolated utterances.

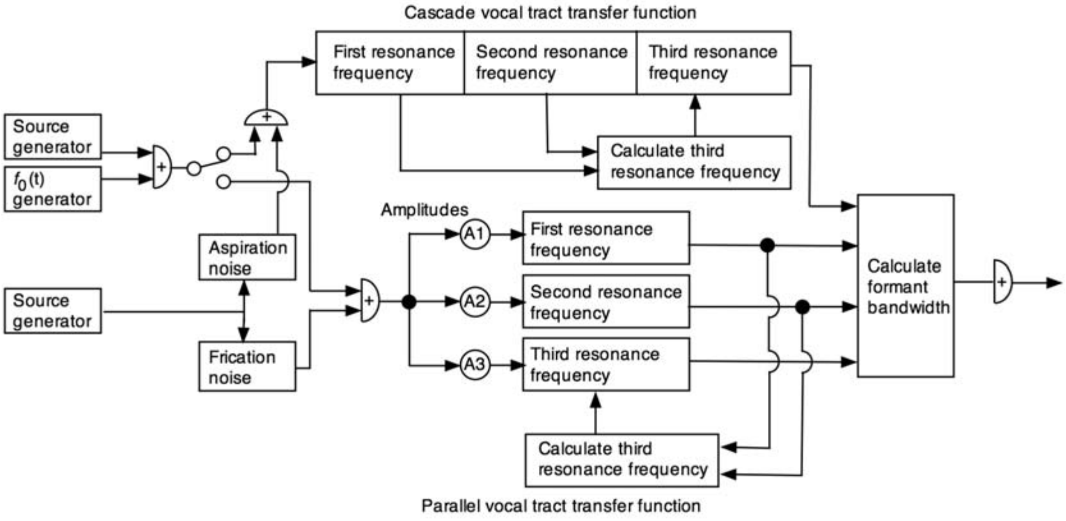
Cascade vocal tract transfer function
First resonance frequency
Second resonance frequency
Third resonance frequency
Source generator
f_0(t) generator
Calculate third resonance frequency
Amplitudes
Aspiration noise
A1
A2
A3
First resonance frequency
Second resonance frequency
Third resonance frequency
Calculate formant bandwidth
Source generator
Frication noise
Calculate third resonance frequency
Parallel vocal tract transfer function

Figure 2. A schematic of the software used to generate the tokens.

Speech synthesis

The synthesis follows the principles of Klatt [7] and Boersma and Weenink [8]. The software is represented schematically in Fig. 2 and details are given elsewhere [9]. A total of 22,488 monaural files in the .wav format were generated and stored with 16 bit precision and sampled at 11 kHz.

The user interface and data acquisition

The web interface is written in PHP and Java and is described in detail elsewhere [9]. Initially, a page asks the user to specify the type of loudspeakers used: headphones, internal speakers or external speakers (subjects are encouraged, but not obliged, to use headphones to improve the frequency response). The software then acquires demographic data on the subject: country and region of origin, country and region of current residence, first and second languages, age and gender. Subsets of data may be subsequently plotted using these demographic data. Any differences attributable to the type of loudspeakers used for the test (headphones, internal or external speakers) may also be examined. Differences between vowels generated with high and low pitch may also be distinguished.

Once those data are recorded, the software displays the data acquisition page, an example of which is shown in Fig. 3. A sound plays three times and the user chooses one of 17 possible words or ‘Vowel unrecognisable’. Additional repeats are available by clicking a ‘play’ arrow. Following this choice, another sound is played three times and the user can either make another choice or go to the ‘Results’ section.

The parameter space is sampled in a pseudo-random routine that repeats once all points in the space ($F2$, $F1$, duration, f_0) have been sampled. Subjects may continue for as long as they wish. At any stage, they may stop and view the results for their own data and return to continue either immediately or later.

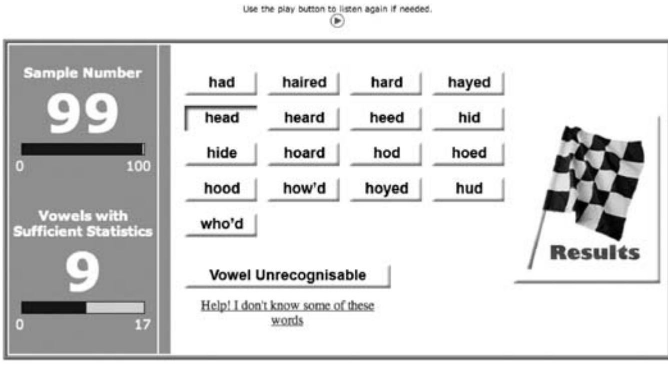

Figure 3. Part of the data acquisition page. The user hears a sound, clicks on a choice and either requests a repeat, proceeds to the next sound or proceeds to the ‘results’ page.

Initially, subjects were recruited by announcing the URL (www.phys.unsw.edu.au/swe) on our own speech and music web sites (www.phys.unsw.edu.au/speech) and by inviting colleagues and friends to participate. An announcement in *Echoes*, a newsletter published by the Acoustical Society of America, also recruited subjects.

RESULTS AND DISCUSSION

At the time of writing, 302 American residents, 112 Australian residents and 71 residents of the UK had been surveyed, along with subjects living in 63 other countries.

On all the displays, $F1$ is plotted in the negative y direction and $F2$ in the negative x direction. This presentation is traditional in phonetics, because it roughly corresponds to the vowel maps in which jaw height is plotted in the y direction and tongue fronting is plotted in the negative x direction [10]. For the benefit of those unfamiliar with phonetics, this rough correspondence is indicated on the display of results in our study and thus also on Figs. 4 and 5.

The coordinates plotted for any word are the mean values of ($F2$, $F1$) for all sounds identified as that word. ‘Short’ printed with a vowel means that more than 75% of our subjects’ selections of that word were from sounds of the short duration class and similarly for ‘long’.

Figure 4 displays the data collected from 346 subjects born in the USA and Australia, selected by origin, but with no constraint on sex or age. (The default display includes ellipses whose semi-axes are the standard deviations in the directions of greatest and least correlation, but these have been omitted here for clarity.)

There are, of course, considerable similarities between the maps for these two countries: Americans and Australians can usually understand each other. Figure 4 confirms that there are, however, differences in detail: for instance, when an American says ‘Bob’ (short for Robert), an Australian may hear ‘Barb’ (short for Barbara).

Figure 5 displays the data for subjects born in two different Australian states; 29 from New South Wales (NSW) and 17 from Queensland. Here, again, there are differences.

Are the differences great enough to lead to confusion? Dowd *et al.* [11] measured a characteristic separation on the vowel plane beyond which vowel sounds cease to be confused. This corresponds to about 170 Hz in the $F1$ direction and 450 Hz in the $F2$ direction, and Pythagorean combinations in any other direction. Some pairs of vowels that fall within this distance for NSW fall beyond it for Queensland (e.g. ‘heard’ and ‘had’) and *vice versa* (‘heed’ and ‘hayed’).

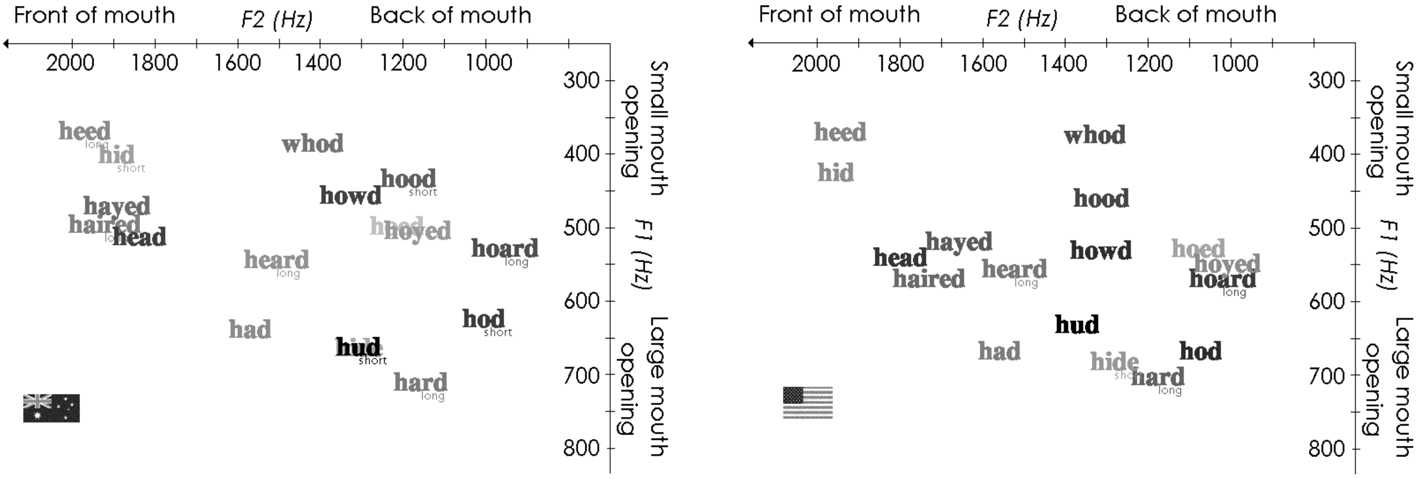

Figure 4. The data for 78 subjects born in Australia and 268 born in the USA. The words are printed so that their centres lie over the mean ($F2$,$F1$). Because this allows printed words to obscure one another, we note that, ‘hud (short)’ coincides with ‘hide’ (on average) for these Australian subjects. The words appear in different colours on the web.

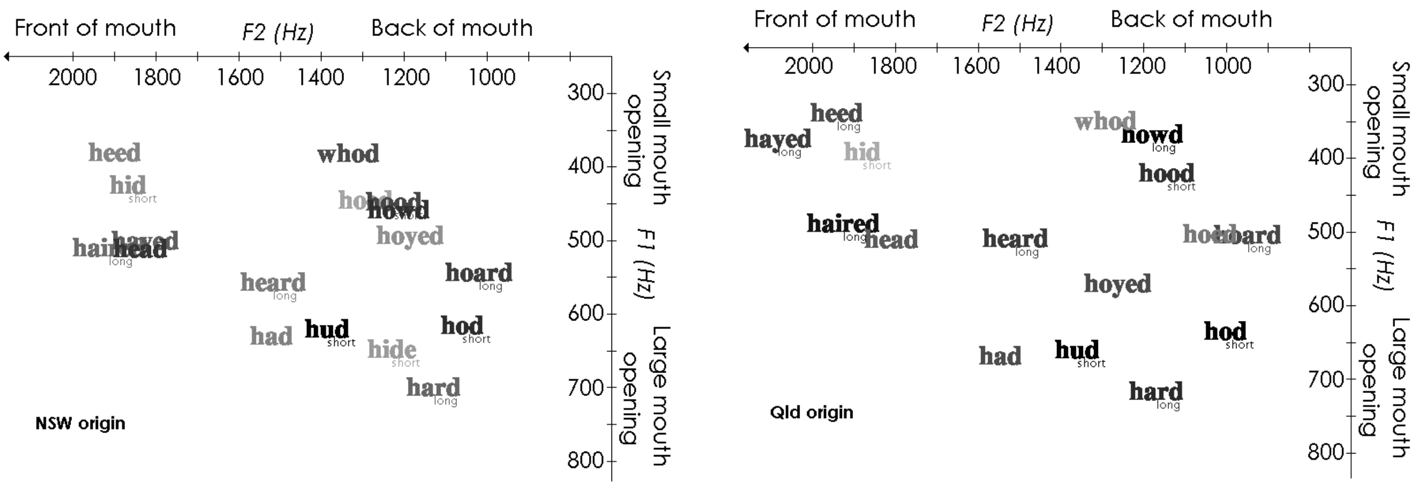

Figure 5. The data for 29 subjects born in New South Wales and 17 from Queensland. We note that ‘haired (long)’, ‘head’ and ‘hayed’ overlap for NSW, as do ‘hood’, ‘howd’ and ‘hoed’, while ‘hoard (long)’ and ‘hoed’ coincide for Queensland. (The sample is not yet large enough to give good statistics on ‘hide’ for the latter.).

It is possible, of course, to produce very many such plots and comparisons for different regions or for different sets of the experimental parameters. Subjects who have finished recording a set of responses are invited to look at their own vowel map, as well as those of various demographic groups, which may be sorted by country and province of birth, region of current residence and/or region in which the subject has previously resided, and/or by first, second or third language, by gender and age and by combinations of these.

CONCLUSIONS

The technique has been demonstrated over three years and the supporting technology proved reliable. It appears that it has not been noticeably vandalised by spurious entries. The data set is large and growing and samples four dimensions. This paper has given only simple examples, illustrating the expected regional variations. Quantitative analysis, however, is left for further studies, possibly involving experts in different areas. In the future, it may also be interesting to compare results gathered in different decades, as Mannell [3] is doing in another study. We do not propose allowing completely free searches of the database, because this might violate the privacy of the subjects. We do, however, propose to make the data available to interested investigators after discussion of any possible ethical issues involved.

ACKNOWLEDGMENTS

We thank the many volunteer subjects who have contributed to the project and the Australian Research Council for support.

REFERENCES

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

[2] J. Clark, C. Yallop and J. Fletcher. *An Introduction to Phonetics and Phonology*. Blackwell, 2007

[3] R.H. Mannell. “Perceptual vowel space for Australian English lax vowels: 1988 and 2004”, *Proc. 10th Australian Intl. Conf. Speech Sci. and Tech.*, Sydney, pp 221-226 (2004)

[4] J. Hay, P. Warren and K. Drager “Factors influencing speech perception in the context of a merger-in-progress”. *J. Phonetics*, 34; 458-484 (2006)

[5] J. W. Hawks and J. D. Miller “A formant bandwidth estimation procedure for vowel synthesis”. *J. Acoust. Soc. Am.* 97; 1343-1344 (1995)

[6] N. Minematsu, M. Sekiguchi and K. Hirose “Automatic estimation of one’s age with his/her speech based upon acoustic modeling techniques of speakers”. *Acoustics, Speech, and Signal Processing. Proceedings ICASSP’02, IEEE* (2002) pp 137-140 (2002)

[7] D. Klatt. “Software for a cascade/parallel formant synthesizer”, *J. Acoust. Soc. Am.* 67; 971–995 (1980)

[8] P. Boersma, and D. Weenink “Praat - Doing phonetics by computer (version 4.3.14) [Computer Program]”. Tech. Rep., Institute of Phonetic Sciences, University of Amsterdam, Netherlands. Viewed 3/10/2010 http://www.praat.org (2006)

[9] A. Ghonim. *Sounds of World English: a Web Investigation*. Undergraduate Thesis, The University of New South Wales, Sydney, Australia, 2007

[10] M. Joos “Acoustic phonetics” *Language* 24, 1–136 (1948)

[11] 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 and Speech*, 41, 1-20 (1998)