DISCRIMINATION OF SPEAKERS USING THE FORMANT DYNAMICS OF /uː/ in BRITISH ENGLISH

Kirsty McDougall and Francis Nolan

Department of Linguistics, University of Cambridge
kem37@cam.ac.uk, fjn1@cam.ac.uk

ABSTRACT

A study of speaker-distinguishing properties of the formant dynamics of /uː/ is presented. Measurements at equidistant intervals along the F1 and F2 contours of /uː/ are compared with polynomial characterisations of the contours. Approximating the contours with quadratic and cubic polynomials allows more speaker-distinguishing information to be conveyed with fewer parameters. Based on discriminant analysis, the best value per predictor is provided by a quadratic approximation of F2.

Keywords: speaker characteristics, formant frequencies, formant dynamics, SSBE, /uː/

1. INTRODUCTION

While research in speaker characteristics has traditionally focused on 'static' properties of the speech signal (e.g. a vowel's formant frequencies at its midpoint), more recent work shows that dynamic (time-varying) features of speech carry important information about a speaker. Static features demonstrate differences among speakers since they are related to speakers' anatomical dimensions, e.g. formant frequencies reflect the length and configuration of the vocal tract [11]. Dynamic features of speech, on the other hand, offer greater scope for variation among speakers, since they reflect the *movement* of the individual's speech organs as well as anatomical dimensions. If speech is conceived of as a series of linguistically determined targets (canonically thought of as the 'centres' of segments) linked by transitions, we can hypothesise that the targets are highly constrained by the shared language system, and that the transitions present greater potential for individual variation. In moving between targets in the stream of speech, the speaker has a large number of degrees of freedom available, and is likely to adopt an individual articulatory solution [6, 7, 9].

Formant frequency dynamics have been shown to offer speaker-specific information in a number of studies [1, 3, 5-7]. For example, McDougall [5] demonstrates considerable differences among five speakers' F1-F3 dynamics for the sequence /aɪk/, using measurements at 10% intervals along its contours. However further work is needed to develop efficient ways to capture and utilise this information. McDougall [6, 7] attempts to characterise the formant dynamics of /aɪk/ with fewer parameters by fitting quadratic and cubic polynomials to each contour. This technique discriminated speakers almost as well as the direct measurements.

This study examines the formant dynamics of the vowel /uː/ in Standard Southern British English (SSBE) for a larger group of speakers (n = 20). The motivation for studying /uː/ was as follows. /uː/ has been observed to be undergoing change in SSBE [2] so it was hypothesised that, in addition to the reasons for analysing dynamic speech features above, variability among speakers in the production of this vowel could be expected. Further, as part of investigations undertaken in the DyViS project, when taking static formant measurements of the 'target' of various vowels, wide between-speaker variation was observed in the formant dynamics of /uː/. Speakers differed both in terms of the temporal location of the 'target' for /uː/ and in their transitions to and from this target. /uː/ rarely exhibited a steady state and, as such, 'monophthong' is not really an appropriate description for the vowel. The present work thus aims to quantify between-speaker variation in the formant dynamics of /uː/. Measurements of its formant frequencies are taken along each formant contour, then fitted with polynomial equations. The original measurements and polynomial fittings are then compared with respect to their accuracy in distinguishing speakers.

2. METHOD

2.1. Subjects

Subjects were twenty male speakers of SSBE, aged 18-25 years (denoted S1, S2, …). The recordings were taken from the DyViS database [see 10].

2.2. Materials and elicitation

The data analysed are six repetitions per speaker of the vowel /uː/ receiving nuclear stress in the hVd context 'who'd'. The target word was included in capitals in the sentence:

He hates contracting words, but he said a WHO'D today.

Six instances of this sentence were arranged randomly among a number of other sentences. The sentences were presented to subjects one at a time using *PowerPoint*. Subjects were asked to read aloud each sentence at a normal speed, in a normal, relaxed speaking style, emphasising the word in capitals. They practised reading a few sentences before the experimental items were recorded. Subjects were encouraged to take their time between sentences and asked to re-read any sentences containing errors.

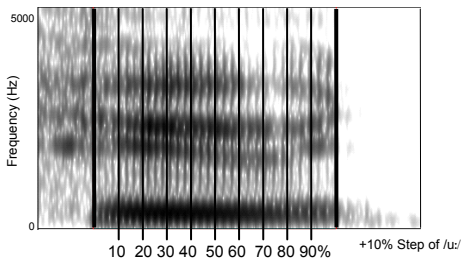
2.3. Recording

Subjects were recorded in the sound-treated booth in the Phonetics Laboratory at the University of Cambridge. Each subject was seated with a Sennheiser ME64-K6 cardioid condenser microphone positioned approximately 20 cm from his mouth. The recordings were made with a Marantz PMD670 portable solid state recorder using a sampling rate of 44.1 kHz.

2.4. Measurements

Using *Praat*, a wide-band spectrogram was produced for each utterance and two vertical markers placed by hand to demarcate the /uː/ segment. Measurements of the centre frequencies of the F1 and F2 contours of each /uː/ token were made with the aid of the *Praat* formant tracker in the following way. For each token, the accuracy of the formant tracking overlay was checked visually. The default number of formants picked by the algorithm was five, but this was adjusted to six or seven in cases where lower formants were skipped. A script was used to calculate the total duration of each /uː/ segment and divide it into ten equal intervals, as in Figure 1. The script measured the centre frequencies of F1 and F2 at each +10% step, thus time-normalising each formant contour. The resulting measurements were checked graphically and in cases where an implausible measurement arose, the formant was remeasured manually using the spectrogram and the spectral slice provided by *Praat*.

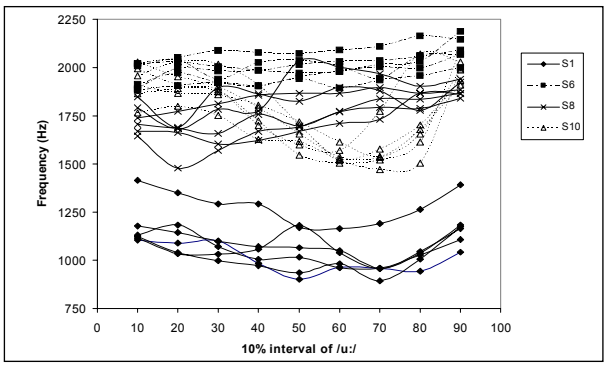
Figure 1: Spectrogram of a token of /uː/ produced by S10, showing +10% steps at which F1 and F2 frequency measurements were made.



3. RESULTS AND ANALYSIS

As an example, the F2 contours of /uː/ for the six tokens produced by four of the twenty speakers are shown in Figure 2. Differences among speakers are clearly evident: individuals differ considerably in the shape and relative frequency of their formant contours. This is the case for all speakers, for F1 and even more so for F2. Some speakers assume a relatively straight trajectory (e.g. S6 in Figure 2) while others exhibit a great deal of movement, (e.g. S10). Within a speaker, positions and shapes of the formant contours are relatively consistent.

Figure 2: F2 frequency contours of /uː/ for the six tokens produced by each of S1, S6, S8 and S10.



To quantify the dynamic differences observed between speakers, polynomial equations were fitted to the formant contours in the following way using *Matlab* (see [7] for further elaboration of this method). Using linear regression the F1 and F2 contours of each token were fitted with each of a quadratic, a cubic and a quartic polynomial of the forms:

$$y = a_0 + a_1 t + a_2 t^2$$
$$y = b_0 + b_1 t + b_2 t^2 + b_3 t^3$$
$$y = c_0 + c_1 t + c_2 t^2 + c_3 t^3 + c_4 t^4$$

where y represents the formant frequency, t represents time on a normalised scale from 1 to 9 (for the nine +10% steps), and a_0, a_1 and a_2, b_0, b_1, b_2 b_3, c_0, c_1, c_2, c_3 and c_4 are the coefficients which define the quadratic, cubic and quartic

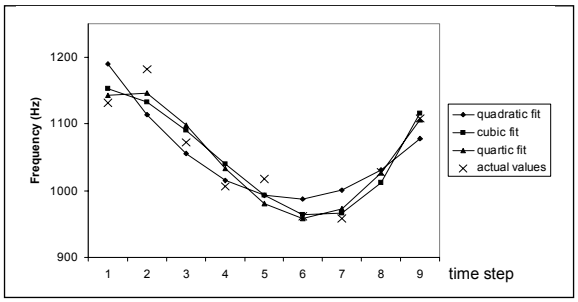
respectively.

An example of the three polynomials fitted to the F2 contour of a token of /u:/ produced by S1 is given in Figure 3. The original formant measurements are shown, together with the quadratic, cubic and quartic fittings. The equations of these curves are:

$y = 1280 - 102t + 8.82t^2$
$y = 1140 + 37.7t - 24.4t^2 + 2.21t^3$
$y = 1040 + 169t - 76.5t^2 + 10t^3 - 0.391t^4$

The quadratic provides a reasonable fit to the contour (R = 0.7439), but is improved on by the cubic (R = 0.8899) and slightly further improved by the quartic (R = 0.9087). Similar improvements in goodness of fit are exhibited across the data set for both F1 (mean R: quadratic 0.8819, cubic 0.9343, quartic 0.9668) and F2 (mean R: quadratic 0.8258, cubic 0.9232, quartic 0.9560).

Figure 3: Graph of the measurements of the F2 contour of the first token of /u:/ produced by S1. Quadratic, cubic and quartic polynomials are fitted.



The next step was to evaluate the effectiveness in distinguishing speakers of the dynamic information captured by each of the polynomials. Direct discriminant function analyses were performed testing the polynomial coefficients as predictors of 'membership' of the twenty speakers, S1, S2, … (k = 20). Separate analyses were run for each formant. For comparison, a second discriminant analysis with the same number of predictors was run for each analysis, as shown in Table 1. The second set of predictors was a subset of the original measurements of the same formant, selected at equidistant intervals along the contour.

All analyses achieve much greater discrimination of speakers than chance level (1/20 = 5%). The F2 analyses consistently provide higher levels of classification than F1 analyses. This is compatible with visual observation of graphs of the contours in which F2 exhibits greater between-speaker variation. It is also consistent with findings from previous research [e.g. 4, 5, 8] where higher formants have tended to yield greater levels of individual variation than F1.

Table 1: Discriminant analyses run on the data set. The first column shows p, the number of predictors, the second lists the predictors, the third gives the classification rate resulting (CR) and the fourth gives an estimation of the 'worth' of each predictor, CR/p.

p	Predictors	CR (%)	CR/p
3	F1 – 20%, 50%, 80%	28.3	9.4
3	F1 – a_0, a_1, a_2	28.3	9.4
4	F1 – 20%, 40%, 60%, 80%	32.5	8.1
4	F1 – b_0, b_1, b_2, b_3	39.2	9.8
5	F1 – 10%, 30% 50%, 70%, 90%	36.7	7.3
5	F1 – c_0, c_1, c_2, c_3, c_4	30.0	6.0
3	F2 – 20%, 50%, 80%	41.7	13.9
3	F2 – a_0, a_1, a_2	45	15
4	F2 – 20%, 40%, 60%, 80%	35	8.8
4	F2 – b_0, b_1, b_2, b_3	49.2	12.3
5	F2 – 10%, 30% 50%, 70%, 90%	49.2	9.8
5	F2 – c_0, c_1, c_2, c_3, c_4	34.2	6.8
5	F1 – b_0, b_3, F2 – a_0, a_1, a_2	51.7	-

The analyses based on the quadratic and cubic polynomial coefficients provide the same or more speaker-distinguishing information than those with the same number of predictors based on the original measurements. For F1, the quadratic-based predictors (p = 3) performed equally well as the original measurements, and the cubic-based predictors (p = 4) were 6.7% better. For F2, improvements of 3.3% and 14.2% were made by quadratic-based and cubic-based predictors respectively. In other words, fitting these curves with quadratic or cubic polynomials enables the speaker-specific information conveyed by formant dynamics to be captured more efficiently. The quartic fitting (p = 5), however, does not improve the classification achieved: a decrease of 6.7% for F1 and of 15% for F2 is observed.

In general, the more predictors included in a discriminant analysis, the greater the classification achieved, depending on the degree of speaker-specific information contributed by each predictor. This can be seen for the analyses based on 3, 4 and 5 original measurement predictors for F1 and F2 (except for the discrepancy of F2 – 20%, 40%, 60%, 80%) [cf. 5]. However there is a limit to the number of predictors which can be included in a discriminant analysis. It is not possible simply to include a large number of direct measurements of the contours as predictors, partly because the predictors will exhibit a great deal of correlation amongst themselves, and also because discriminant analysis requires the number of predictors to be smaller than the sample size of the smallest group.

In the present study, since $n = 6$ the maximum number of predictors is 5.

The efficiency brought about by expressing this formant dynamic information in terms of quadratic or cubic polynomial coefficients is very useful, but, given that the cubic fitting requires one more predictor than the quadratic, which approximation is the most efficient? The ultimate aim is to extract the optimal amount of speaker-distinguishing information for the number of predictors permissible; the fewer predictors needed to exploit the F1 and F2 dynamics, the greater the scope for including further predictors based on F3, F4, duration, etc. To assess the 'worth' of each predictor in the discriminant analyses, the classification rate resulting from each analysis was divided by the number of predictors, as shown in the rightmost column of Table 1.

For F2, and overall, the quadratic-based analysis gave the highest value per predictor, with each predictor having a worth of 15%, in contrast to 12.3% for the cubic-based analysis. For F1, the cubic-based analysis gave only slightly better value per predictor than the quadratic-based (9.8% versus 9.4%). So although the cubic polynomials provide a better fit to the F1 and F2 contours, it appears that a worthwhile amount of speaker-distinguishing information can be captured with the quadratic approximations, saving the expense of the additional predictor. A final discriminant analysis was thus carried out using five of the highest 'worth' predictors – the three F2 quadratic-based predictors, and two F1 cubic based predictors, F1b_0 and F1b_3. The two F1 predictors were chosen based on the structure matrix for the corresponding discriminant analysis, as those contributing most to the discrimination. This analysis achieved a classification rate of 51.7%, a further demonstration that selecting efficient formant dynamic predictors offers a powerful method for speaker discrimination.

4. CONCLUSION

Formant dynamics are an interesting source of speaker-discriminating information, reflecting both differences in speakers' vocal tract morphology and individual differences in the articulatory trajectories chosen to produce each sound. In this study formant dynamics of /u:/ were quantified for 20 male speakers of SSBE. Large between-speaker variation was observed in the shape and frequency of F1 and F2 contours, especially in F2. Characterising the formant contours with quadratic, cubic and quartic polynomials enabled speaker-distinguishing information to be captured with fewer numbers. In discriminant analysis, cubic approximations offered the best classification rates (F1: 39.2%, F2: 49%), but the quadratic approximations were the most efficient, carrying greater speaker-distinguishing information per predictor. Further work should build on these findings to develop efficient techniques for capturing a speaker's individuality using formant dynamics.

5. ACKNOWLEDGEMENTS

This research is supported by the UK Economic and Social Research Council as part of the project 'Dynamic Variability in Speech [DyViS]: A Forensic Phonetic Study of British English' [RES-000-23-1248]. Thanks are due to Gea de Jong and Toby Hudson for assistance with recording subjects and to Caroline Williams and Toby Hudson for writing the *Praat* script.

6. REFERENCES

[1] Greisbach, R., Esser, O., Weinstock, C. 1995. Speaker identification by formant contours. In: Braun, A., Köster, J.-P. (eds.), *Studies in Forensic Phonetics: BEIPHOL 64*. Trier: Wissenschaftlicher Verlag Trier, 49-55.

[2] Hawkins, S., Midgley, J. 2005. Formant frequencies of RP monophthongs in four age-groups of speakers. *JIPA* 35(2): 183-199.

[3] Ingram, J.C.L., Prandolini, R., Ong, S. 1996. Formant trajectories as indices of phonetic variation for speaker identification. *Forensic Linguistics* 3(1): 129-145.

[4] Jessen, M. 1997. Speaker-specific information in voice quality parameters. *Forensic Linguistics* 4(1): 84-103.

[5] McDougall, K. 2004. Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. *Int. Jnl. Speech, Language and the Law* 11(1): 103-130.

[6] McDougall, K. 2005. *The Role of Formant Dynamics in Determining Speaker Identity*. Ph.D. Dissertation, University of Cambridge.

[7] McDougall, K. 2006. Dynamic features of speech and the characterisation of speakers: towards a new approach using formant frequencies. *Int. Jnl. Speech, Language and the Law* 13(1): 89-126.

[8] Nolan, F. 1983. *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

[9] Nolan, F. 1997. Speaker recognition and forensic phonetics. In: Hardcastle, W.J., Laver, J. (eds.), *The Handbook of Phonetic Sciences*. Oxford: Blackwell, 744-767.

[10] Nolan, F., McDougall, K., de Jong, G., Hudson, T. 2006. A forensic phonetic study of 'dynamic' sources of variability in speech: the DyViS project. *Proc. 11th SST*, Auckland, 13-18. <http://www.assta.org/sst/2006/sst2006-17.pdf>.

[11] Stevens, K.N. 1971. Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. *Proc. 7th ICPhS*, Montreal, 206-232.