MULTIDIMENSIONAL VARIATION IN ENGLISH DIPHTHONGS

Daniel Williams[a,d], Jaydene Elvin[b,d], Paola Escudero[c,d], Adamantios Gafos[a]

[a]University of Potsdam, [b]California State University, Fresno, [c]Western Sydney University,
[d]ARC Centre of Excellence for the Dynamics of Language
daniel.williams@uni-potsdam.de, jaydene@mail.fresnostate.edu,
paola.escudero@westernsydney.edu.au, gafos@uni-potsdam.de

ABSTRACT

The present study investigated the extent to which variation along formant trajectory dimensions – considered both separately and simultaneously – manifests in the English diphthongs CHOICE, FACE, MOUTH, GOAT and PRICE. The sources of variation were phonemic category, flanking consonants, speaker's gender and speaker's regional background. Formant trajectory dimensions were overall and time-varying first (F1) and second (F2) formant trajectories as parameterised by their means and time-varying slopes and curvature using the discrete cosine transform. Phonemic category was robustly predicted by variation in the combination of F1 and F2 means and slopes, whereas flanking consonants were not. Gender was strongly predicted by variation in F1 and F2 means, but not in F1 and F2 time-varying aspects. Regional background emerged in the variation of both F1 and F2 means, slopes and curvature to roughly similar extents. Given the acoustic multidimensionality of vowels, variation is most appropriately viewed with a multivariate approach.

Keywords: English dialects, vowel acoustics, spectral change, gender

1. INTRODUCTION

Vowel segments, like many phonetic phenomena, are acoustically *multidimensional*. That is, an individual vowel token may differ from others along several acoustic dimensions simultaneously, e.g., duration, fundamental frequency as well as aspects of first (F1) and second (F2) formant trajectories. Although one dimension may be more informative than another in speech perception [4], several dimensions are nonetheless used by listeners at the same time. Thus, should acoustic dimensions be examined independently when their values represent a single phonetic unit, such as a vowel token? This question is pertinent for variation, as differences on one dimension may be related to those on another [9].

Characterising time-varying formant trajectories relies on sampling frequencies at discrete points throughout the course of a vowel. Using this information in conjunction with explanatory variables, e.g., about the speakers themselves, is a technical challenge, and it is a matter of debate how best to represent formant trajectories. Ultimately, a balance needs to be struck between preserving acoustic detail without sacrificing the utility of explanatory variables that may give rise to variation.

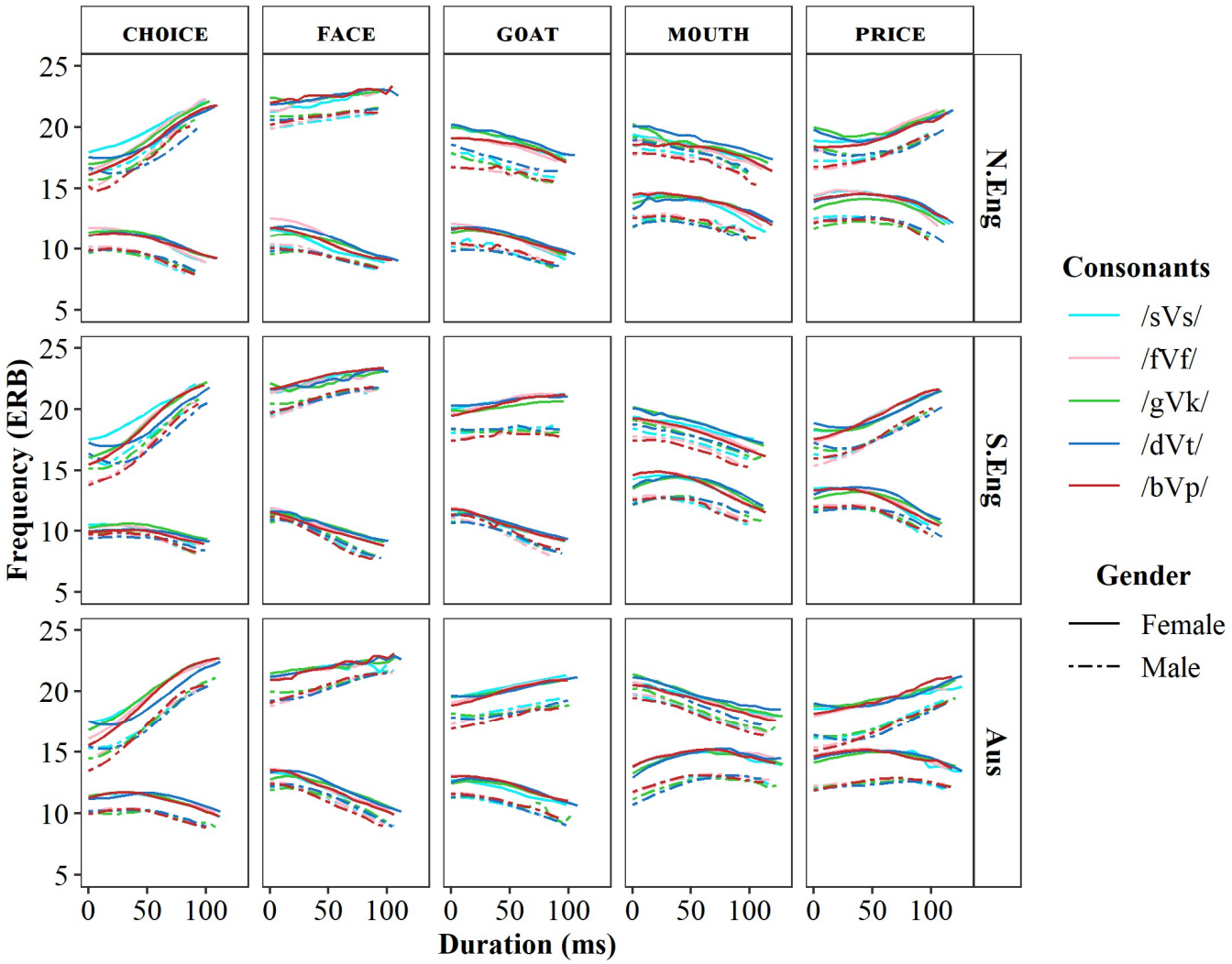
"Ad-hoc" representations of formant trajectories, e.g., generalised additive models [11], involve fitting curves to the sampled formant frequencies over time. They have the advantage of not forcing a parameterisation on trajectory shapes. Unfortunately, they offer limited scope – for the time being, at least – for modelling trajectories along with multiple and often related and/or interacting explanatory variables. Furthermore, the "ad-hoc" parameters of formant trajectories do not generalise across different datasets.

"Predefined" representations of formant trajectories, on the other hand, involve fitting curves with known parameters. In its simplest form, a formant sampled at vowel onset and offset represents a linear trajectory [7]. Despite the apparent crudeness, it is remarkably effective for classifying vowels according to phonemic categories, and such basic time-varying spectral information is perceptually very relevant, especially for English [2, 13, 14]. More intricate "predefined" representations involve a greater number of samples. A popular method is the discrete cosine transform (DCT) whose curve parameterisation is based on a trajectory's mean plus ½ cosine multiples, each with amplitudes representing deviations from this mean [6]. A formant trajectory is represented by a set of DCT coefficients (DCTCs): the 0^{th} is the mean across samples, the 1^{st} is a ½ cosine (the slope's magnitude and direction) and the 2^{nd} is a full cosine (overall curvature). Further DCTCs represent more complex shapes.

Importantly, DCT representations can be readily used in a plethora of linear modelling procedures which permit complex factorial and/or multivariate designs, overcoming some shortcomings of "ad-hoc" approaches. Effectively representing formant trajectories along these dimensions has been demonstrated for variation between phonemic vowel categories and also for explaining variation relating to speakers' regional background and gender [3, 12].

The present study investigates the extent to which four potential sources of variation in English

Figure 1: Mean F1 and F2 trajectories in central 60% portions.



diphthongs manifest along several dimensions, namely log-duration, overall (i.e., 0th DCTCs) and time-varying F1 and F2 frequencies (i.e., 1st DCTCs and 2nd DCTCs). The sources of variation are (1) phonemic category, (2) flanking consonants, (3) speaker's gender and (4) speaker's regional background. First, we examine how variation along these formant trajectory dimensions – when considered separately and simultaneously – can be attributed to the four potential sources. Second, we show why it is advisable to consider variation along all acoustic dimensions of interest at the same time.

2. METHOD

The diphthong tokens came from corpora reported in [3] and [12]. 55 speakers aged between 18 and 30 at the time of recording participated and they had different regional backgrounds: 19 (10 female) were from Northern England (N.Eng), 17 (10 female) were from Southern England (S.Eng) and 19 (12 female) were from Australia (Aus). Each speaker produced the syllables /sVs/, /fVf/, /gVk/, /dVt/ and /bVp/ where /V/ is one of the five English closing diphthongs CHOICE, FACE, GOAT, MOUTH and PRICE (according to Wells' [10] Lexical Sets). N.Eng and S.Eng speakers produced each syllable in a sentence twice, while Aus speakers produced each syllable once in isolation and once in the same sentence frame used by N.Eng and S.Eng speakers. This yielded two repetitions of each unique syllable per speaker. Using the default settings for male and female speakers in *Praat* [1], F1 and F2 frequencies were sampled from 19 equally spaced intervals in the central 60% of each vowel token in order to remove obvious transitions with flanking consonants. F1 and F2 Hz values were converted to ERB and transformed using the DCT.

3. RESULTS

Plots of F1 and F2 trajectory means (across the 55 speakers' pools of tokens) are displayed in Figure 1. All analyses were carried out in *R* [8] using the *MCMCglmm* [5] package which fits generalized

linear mixed-effects models employing Markov chain Monte Carlo (MCMC) sampling for Bayesian statistics. For every model, weak priors were set for the fixed-effects residuals and degree of belief parameter was set to the lowest bound for fixed and random effects. Single Markov chains sampled from the posterior distribution for each model; the initial 10,000 iterations were discarded and a further 100,000 iterations were run and then thinned, leaving 1,000 samples per model.

3.1. Formant trajectory variation as a predictor

Mixed-effects multinomial logistic regressions were run with each potential source of variation as the categorical dependent variable. For each source, we ran five models with different combinations of standardized F1 and F2 DCTCs as predictors; standardized log-duration was included as a control variable in all models because we were interested in variation specifically along acoustic dimensions relating to F1 and F2. To account for token imbalances (e.g., 58% of tokens were produced by female speakers), which might affect the likelihood of a member of a particular category occurring, random intercepts were added for item which covered all categorical variables other than the dependent variable and whether syllables were said in isolation or in a sentence. To assess how well the four sources of variation can be predicted by different formant trajectory dimensions (as represented by DCTCs) separately as well as simultaneously, we report predicted probabilities of correct classifications averaged across tokens in Table 1. Probabilities of incorrect classifications (confusions) are not reported.

For phonemic category, both 0th and 1st DCTCs were good predictors, whereas 2nd DCTC performed worse. Most striking is that 0th and 1st DCTCs together resulted in accurate separation, which is remarkable given that there are different syllables produced by 55 speakers of different regional backgrounds and genders. Adding 2nd DCTCs to the combination did not improve classification.

Predictions of flanking consonants were poor overall, indicating that this source of variation does not manifest consistently in F1 and F2 trajectories within the central 60% portion of diphthong tokens.

For gender, classification is substantially above chance with only 0th DCTCs, but around chance with 1st or 2nd DCTCs, and classification does not improve when these latter measures are combined with 0th DCTCs. Interestingly, females' diphthongs were classified more accurately, suggesting less between-speaker variation in females' use of time-varying formant dimensions compared to those of males.

For regional background, classifications were reasonably above chance with 0th, 1st or 2nd DCTCs separately. There was not much improvement with both overall and time-varying trajectory dimensions together. Overall, these results indicate a fair amount of overlap in variation across the three varieties. Nevertheless, Aus tokens were much more accurately identified than the other two varieties, suggesting formant trajectory variation in Aus tokens is more distinct from that found in N.Eng and S.Eng tokens.

In summary, all sources of variation, except for flanking consonants, can be predicted above chance levels based on overall (0th DCTCs) or time-varying (1st and 2nd DCTCs) F1 and F2 trajectory dimensions. Importantly, using these two formant dimensions is useful only for predicting phonemic category and is negligible for flanking consonants, gender or regional background. This is likely because the acoustic manifestations of these sources of variation co-vary, e.g., with phonemic category (cf., Figure 1).

Table 1: Probabilities of correct classifications by phonemic category (chance = 0.20), flanking consonants (chance = 0.20), gender (chance = 0.50) and regional background (chance = 0.33) according to different combinations of DCTCs as predictors.

Phonemic category	0th	1st	2nd	0th, 1st	0th, 1st, 2nd
CHOICE	0.59	0.86	0.23	0.95	0.95
FACE	0.89	0.50	0.32	0.90	0.90
GOAT	0.47	0.45	0.31	0.80	0.81
MOUTH	0.52	0.80	0.39	0.92	0.93
PRICE	0.47	0.63	0.37	0.88	0.89
Mean	*0.59*	*0.65*	*0.32*	*0.89*	*0.89*
Flanking consonants	0th	1st	2nd	0th, 1st	0th, 1st, 2nd
/sVs/	0.21	0.22	0.21	0.22	0.23
/fVf/	0.24	0.24	0.24	0.26	0.24
/gVk/	0.21	0.23	0.23	0.24	0.37
/dVt/	0.23	0.24	0.26	0.25	0.20
/bVp/	0.18	0.20	0.20	0.21	0.19
Mean	*0.22*	*0.23*	*0.23*	*0.24*	*0.25*
Gender	0th	1st	2nd	0th, 1st	0th, 1st, 2nd
Female	0.90	0.62	0.62	0.90	0.91
Male	0.86	0.45	0.45	0.86	0.87
Mean	*0.88*	*0.55*	*0.55*	*0.89*	*0.89*
Regional background	0th	1st	2nd	0th, 1st	0th, 1st, 2nd
N.Eng	0.42	0.50	0.41	0.50	0.50
S.Eng	0.46	0.43	0.38	0.53	0.52
Aus	0.72	0.68	0.62	0.77	0.77
Mean	*0.54*	*0.54*	*0.48*	*0.61*	*0.61*

3.2 Predicting formant trajectory variation

The previous section showed that variation *simultaneously* in the F1 and F2 0th and 1st DCTCs performed best overall at predicting membership of phonemic category, regional background or gender because also including variation from F1 and F2 2nd DCTCs did not improve performance. Thus, it seems reasonable to define formant trajectories along these most informative dimensions at the same time.

Typically, variation is studied in the reverse way, namely, explanatory variables are used to predict acoustic variation – often separately for more than one dependent variable, even though these all describe the same tokens. We illustrate that failing to account for acoustic multidimensionality may limit explanations of acoustic variation. We model variation in the five diphthongs using a multivariate mixed-effects linear regression. Phonemic category, regional background and gender served as interacting predictors and log-duration, F1 and F2 0th and 1st DCTCs were the dependent variables. Also entered were by-speaker random intercepts and slopes for phonemic category, flanking consonants and whether syllables were said in isolation or a sentence. Crucially, a covariance matrix was estimated for the dependent variables at the level of tokens.

Table 2: Covariance matrix of dependent variables at the level of tokens (expressed as correlations of posterior medians) from a mixed-effects multivariate regression model. Significant correlations – those with 95% credible intervals (in brackets) not crossing zero – are displayed in bold.

		F1		F2	
		0th	1st	0th	1st
F1	1st	**–0.17** (–0.20, –0.13)			
F2	0th	**–0.09** (–0.12, –0.04)	0.01 (–0.02, 0.05)		
	1st	0.00 (–0.04, 0.03)	–0.02 (–0.05, 0.02)	**–0.05** (–0.09, –0.02)	
Log-duration		**–0.19** (–0.23, –0.15)	**0.27** (0.22, 0.31)	**–0.06** (–0.11, –0.01)	**–0.14** (–0.18, –0.09)

Table 2 presents the covariance matrix from this model. It is clear there are some modest correlations, suggesting that some variation is indeed shared across dimensions at the level of tokens. Most striking is that higher log-duration values are significantly correlated with higher F1 1st DCTC values, suggesting that tokens with longer durations also display greater falling F1 trajectory change. A further example is that tokens with longer log-durations show lower F1 0th DCTCs, i.e., longer durations are correlated with lower F1 means. Additionally, tokens with higher F1 1st DCTCs exhibit lower F1 0th DCTCs, indicating that greater falling F1 trajectories are associated with lower F1 trajectory means.

4. DISCUSSION AND CONCLUSION

The present study sought to explain four potential sources of variation – phonemic category, flanking consonants, gender and regional background – in the F1 and F2 trajectories of the five English diphthongs CHOICE, FACE, GOAT, MOUTH and PRICE. The acoustic multidimensionality of vowels was also focused on.

The main findings are that F1 and F2 trajectory variation contributes most strongly to phonemic category. Importantly, it is the *simultaneous* variation in overall and time-varying trajectory dimensions which leads to the most successful predictions of phonemic category. On the other hand, gender manifested strongly in variation of formant means, reflecting expected differences in vocal tract sizes, but made little contribution to variation in the time-varying dimensions of formant trajectories. Regional background was predicted with moderate accuracy and was apparent to roughly equal extents in overall and time-varying aspects of trajectories, though both formant dimensions together did not result in large improvements. Flanking consonants were poorly predicted, suggesting that their acoustic influence cannot be captured well in the central 60% portion.

It has been shown that using DCTCs clearly extends to capturing formant trajectory variation other than phonemic category. In line with [9], variation in 2nd DCTCs – corresponding to a formant trajectory's curvature – predicted phonemic category poorly and did not improve classification performance in combination with other dimensions of formant trajectory shapes. However, 2nd DCTCs did perform just as well (or poorly) as 1st DCTCs – corresponding to a formant trajectory's slope – for predicting other potential sources of variation. Nevertheless, 1st and 2nd DCTCs together did not result in improvements, suggesting two time-varying dimensions at the same time are perhaps redundant.

Finally, the amenability of DCTCs in established linear modelling techniques is noteworthy. Variation in different formants, including aspects of their time-varying trajectories, along with variation in vowel duration, can only be examined simultaneously in multivariate analytical approaches. In this way, the manifestation of different explanatory variables can be more accurately modelled on relevant acoustic dimensions, thereby better revealing the rich structures underlying phonetic variation.

5. REFERENCES

[1] Boersma, P., Weenink, D. 2018. *Praat: Doing Phonetics by Computer*. Version 6.0.43. Retrieved 2nd October 2018 from http://www.praat.org/.

[2] Chládková, K., Hamann, S., Williams, D., Hellmuth, S. (2017). F2 slope as a perceptual cue for the front–back contrast in Standard Southern British English. *Lang. Speech* 60, 377–398.

[3] Elvin, J., Williams, D., Escudero, P. 2016. Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *J. Acoust. Soc. Am.* 140, 576–581.

[4] Goudbeek, M., Cutler, A., Smits R. 2008. Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Comm.* 50, 109–125.

[5] Hadfield, J. D. 2010. MCMC methods for multi-response generalized mixed models: The MCMCglmm R Package. *J. Statistical Software* 33, 1–22.

[6] Morrison, G. S. 2013. Theories of vowel inherent spectral change. In: Morrison, G. S., Assmann P. F. (eds.), *Vowel Inherent Spectral Change*. Berlin: Springer, 31–48.

[7] Morrison, G. S., Nearey, T. M. 2007. Testing theories of vowel inherent spectral change. *J. Acoust. Soc. Am.* 122, EL15–EL22.

[8] R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Version 3.5.1. Retrieved 2nd July 2018 from https://cran.ma.imperial.ac.uk/.

[9] Roettger, T. B. 2019. Researcher degrees of freedom in phonetic research. *Lab. Phon.: J. Assoc. Lab. Phon.* 10, 1.

[10] Wells, J. C. 1982. *Accents of English*. Cambridge: Cambridge University Press.

[11] Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *J. Phonetics* 70, 86–116.

[12] Williams, D., Escudero, P. A cross-dialectal acoustic comparison of vowels in Northern and Southern British English. *J. Acoust. Soc. Am.* 136, 2751–2761.

[13] Williams, D., Escudero, P., Gafos, A. 2018. Perceptual sensitivity to spectral change in Australian English close front vowels: an electroencephalographic investigation. *Proc. Interspeech 2018* Hyderabad, 1442–1446.

[14] Williams, D., Escudero, P., Gafos, A. 2018. Spectral change and duration as cues in Australian English listeners' front vowel categorization. *J. Acoust. Soc. Am.* 144, EL215–221.