BEYOND NORTH AMERICAN ENGLISH: MODELLING VOWEL INHERENT SPECTRAL CHANGE IN BRITISH ENGLISH AND DUTCH

Daniel Williams[a], Jan-Willem van Leussen[b], Paola Escudero[c]

[a]University of Potsdam, [b]University of Amsterdam, [c]University of Western Sydney
daniel.williams@uni-potsdam.de, j.w.vanleussen@uva.nl, paola.escudero@uws.edu.au

ABSTRACT

Theories and methods modelling vowel quality in terms of *vowel inherent spectral change* (VISC) have been developed and tested overwhelmingly on North American English (AE) dialects, which raises the question of their generalisability in non-AE dialects and other languages. The present paper examines VISC as an aspect of vowel quality in Standard Southern British English (SSBE) and Northern Standard Dutch (NSD). Despite markedly different VISC patterns, SSBE vowels are analysable along the same lines as in AE. While the same mostly holds for NSD, VISC is found to be more important for determining SSBE vowel quality, especially for SSBE nominal diphthongs. Additionally, a pair of NSD diphthongs presents a challenge for current theories and methods as they are acoustically similar. In line with studies on AE, theorising vowel quality in terms of VISC aids descriptions of vowels and removes the need to treat nominal monophthongs and diphthongs in different ways.

Keywords: vowels, acoustics, Dutch, English

1. INTRODUCTION

In order to represent individual vowel qualities acoustically, first, second and third (F1, F2, F3) formant information is often obtained from a single time point for nominal monophthongs, e.g., midpoint, or two time points for nominal diphthongs, e.g., onset and offset, as in [1]. However, it has long been recognised that formants of nominal monophthongs as well as of diphthongs change over time [9].

Although vowel formants may vary to some extent according to phonetic context factors, e.g., flanking consonants or speaking style [13, 11], some formant movement may occur due to *vowel inherent spectral change* (VISC), i.e., the "relatively slowly varying changes in formant frequencies associated with vowels themselves, even in the absence of consonantal context" [8].

The role of VISC in speech perception is not trivial; it has consistently been found to improve North American English (AE) listeners' vowel identification compared to when it is excluded [4], though its perceptual importance varies, as some AE vowels can be well identified without formant movement [3]. VISC can thus help to signal vowel contrasts in production and perception, e.g., Standard Southern British English (SSBE) /i:/-/u:/ [2].

As for what particular aspects of VISC are crucial, Morrison [5] reviews three hypotheses which agree vowel onsets are perceptually important but disagree with respect to subsequent formant change. The first (*onset + slope*) states that the rate of change is most critical, while the second (*onset + direction*) posits that only the general direction of change is important. He concludes that the third hypothesis (*onset + offset*), which assumes formant frequencies both towards the beginning and end are important, provides a superior account, as evidence from speech perception, e.g., [8, 6], and production, e.g., [5, 7], best supports it. Methods for modelling VISC are also discussed in [5] and the onset + offset hypothesis can be modelled simply by sampling formant frequencies from two separate time points (at the beginning and end of vowel), though denser sampling provides a more detailed rendering of a given formant trajectory.

Importantly, Morrison [5] models VISC, both theoretically and methodologically, based on evidence only from AE dialects. Non-AE dialects or different languages may provide differing support for the onset + offset hypothesis, e.g., VISC patterns in SSBE vowels are generally very different from phonologically equivalent AE vowels [14].

The present study therefore tests the onset + offset hypothesis of VISC on vowels in SSBE and in another language, Northern Standard Dutch (NSD), and compares the VISC modelling methods of sampling formants from just a few time points versus using more detailed formant representations.

2. VISC IN SSBE AND NSD VOWELS

2.1. SSBE and NSD vowel corpora

The SSBE and NSD vowel tokens have previously been described in [14] and [13]. Briefly, vowel tokens were produced by 20 NSD speakers (10 female) and 17 SSBE speakers (10 female) in CVC words embedded in a sentence frame. For SSBE, V was one of the nominal monophthongs /iː, ɪ, ɛ, ɜː, a, ɑː, ʌ, ɒ, ɔː, ʊ, uː/ or one of the diphthongs /eɪ, əʊ, aʊ, aɪ, ɔɪ/. For NSD, V was one of the nominal monophthongs

/i, y, ɪ, ʏ, ɛ, aː, ɑ, ɔ, u/ or one of the nominal potential diphthongs /eː, øː, oː/ or nominal true diphthongs /ɛi, œy, ʌu/. The five CVC contexts were /bVp/, /dVt/, /gVk/, /fVf/ and /sVs/ for SSBE and /pVp/, /tVt/, /kVk/, /fVf/ and /sVs/ for NSD[1]. Duration values and F1, F2 and F3 frequencies were obtained with the procedures in [14].

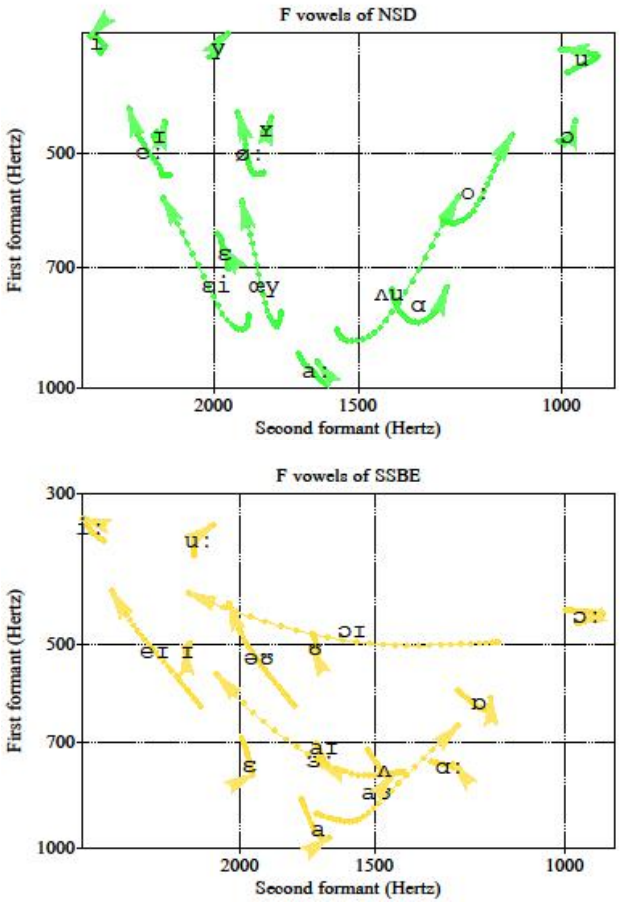
Figures 1-4 display means (from medians over each speaker's pool of vowel tokens) of F1 and F2 values from 30 equally spaced time points in the central 60% portion of each token fitted with 2nd-order discrete cosine transform (DCT) curves.

As can be seen, many of the NSD and SSBE nominal monophthongs exhibit spectral change, though nominal diphthongs tend to show greater amounts of spectral change.

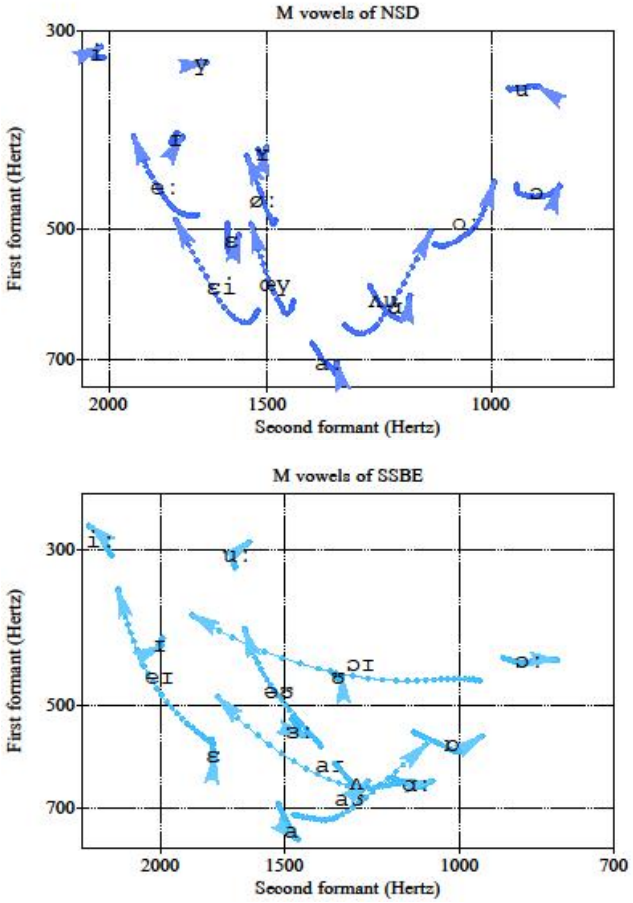
2.2. Method

To test the onset + offset hypothesis on the present data, Discriminant Analyses (DAs) were trained on various combinations of acoustic input variables. For this hypothesis to be well supported, the DAs modelling it should generate the most accurate classifications of NSD and SSBE vowel tokens.

Figures 1-2: Average F1 and F2 trajectories of the central 60% of NSD (upper) and SSBE (lower) female speakers' vowel tokens.



Figures 3-4: Average F1 and F2 trajectories of the central 60% of NSD (upper) and SSBE (lower) male speakers' vowel tokens.



In addition, two VISC modelling methods are tested. *Method A*, using formant frequencies from just a few separate time points, incorporates duration and F1, F2 and F3 values sampled from 20%, 50% and 80% duration to represent onset, midpoint and offset. *Method B*, using more detailed formant information, is based on duration and coefficient values representing aspects of F1, F2 and F3 trajectories. Specifically, formant values sampled from 30 time points in the central 60% portion of every vowel token were fitted with 2nd-order DCT curves and the resulting DCT coefficients were used to characterise aspects of formant trajectories, namely the 0th DCT coefficient represents its mean, the 1st its slope (direction and magnitude of deviation from the mean) and the 2nd its curvature [5].

All DAs were run using the cross-validation (jack-knife) approach and performed separately on SSBE and NSD and male and female speakers' tokens. The percentage of correct classifications for the different combinations of input variables (averaged across genders) are reported in Tables 1-4.

2.3. Results for SSBE

Looking first at the results from Method A (Table 1), using F1, F2 and F3 values from both onset and offset results in considerable improvements compared to using values from just one time point (13.8-22.4%). Unsurprisingly, the most dramatic improvements are for the nominal diphthongs (32.7-36.2%). Including formant values from three time points results in the greatest classification accuracy, though its improvements over using onset and offset are modest.

Table 1: *Method A.* % correct classifications of SSBE vowels from DAs using duration and F1, F2 and F3 values from one, two and three time points.

Model	All	Monoph-thongs	Diph-thongs
onset	67.3	72.7	55.3
offset	73.9	83.1	53.6
midpoint	75.9	84.4	57.1
onset + offset	89.7	89.7	89.8
onset + midpoint + offset	91.0	90.4	92.3

Turning to the Method B (Table 2), the model corresponding to the onset + offset hypothesis (0^{th} + 1^{st}) performs best – around 12.6% more accurate than 0^{th} DCT coefficient values. Unsurprisingly, the 1^{st} DCT coefficient and, to a lesser extent the 2^{nd} DCT coefficient, better classify nominal diphthongs than monophthongs, as these coefficients correspond to aspects of formant trajectory shapes.

Table 2: *Method B.* % correct classifications of SSBE vowels from DAs using duration, 0^{th}, 1^{st} and 2^{nd} DCT coefficients of F1, F2 and F3 trajectories.

Model	All	Monoph-thongs	Diph-thongs
0^{th}	79.3	86.7	63.1
1^{st}	46.7	34.3	73.9
2^{nd}	29.8	25.9	38.5
0^{th} + 1^{st}	91.9	92.5	90.7
0^{th} + 1^{st} + 2^{nd}	91.7	91.8	91.6

2.4. Results for NSD

With Method A (Table 3), using onset and offset formant values results in higher classification accuracy than using those from single time points, though the amount of improvement is lower than for SSBE (Table 1). It appears that the NSD tokens can be fairly well differentiated with duration and midpoint F1, F2 and F3 values (85.0%), though this is 13.3% lower for nominal true diphthongs. As with SSBE, three time points results in very little improvement over two.

Table 3: *Method A.* % correct classifications of NSD vowels from DAs using duration and F1, F2 and F3 values from one, two and three time points.

Model	All	M	D	TD	PD
onset	77.1	82.1	69.7	51.0	88.3
offset	82.3	83.3	80.8	73.0	88.7
midpoint	85.0	86.4	82.8	71.7	94.0
onset + offset	87.2	87.9	86.2	77.3	95.0
onset + midpoint + offset	87.9	89.3	85.8	77.0	94.7

M = monophthongs; D = diphthongs (both true and potential diphthongs); TD = true diphthongs; PD = potential diphthongs

In Method B, using 0^{th} DCT coefficient values classifies NSD vowels relatively well. Using 1^{st} or 2^{nd} DCT coefficients, on the other hand, does not result in high levels of classification accuracy and adding 2^{nd} DCT coefficients to the 0^{th} + 1^{st} model provides only very modest improvement.

Table 4: *Method B.* % correct classifications of NSD vowels from DAs using duration, 0^{th}, 1^{st} and 2^{nd} DCT coefficients of F1, F2 and F3 trajectories.

Model	All	M	D	TD	PD
0^{th}	86.8	87.8	85.3	76.7	94.0
1^{st}	38.1	32.4	46.5	45.7	47.3
2^{nd}	28.5	26.3	31.7	37.7	25.7
0^{th} + 1^{st}	88.4	90.2	85.7	77.3	94.0
0^{th} + 1^{st} + 2^{nd}	88.5	90.3	85.8	78.3	93.3

Interestingly, the NSD nominal true diphthongs are consistently classified with the lowest accuracy with either Method A or B and in all models (cf., classification of NSD potential diphthongs).

3. DISCUSSION

In line with [5], the present study provides support for the onset + offset hypothesis with non-AE vowels – models that lacked information relating to both vowel onsets and offsets were generally less accurate at classifying NSD or SSBE vowel tokens, e.g., formant frequencies from a single time point (Method A) or using a single DCT coefficient (Method B).

Focusing first on Method A, midpoint formant frequencies generated fairly accurate classifications, though NSD fared better, especially as 25.7% more NSD diphthong tokens were correctly classified than SSBE diphthongs. Using formant frequencies sampled at both onset and offset generated more accurate classifications than at a single time point, representing improvements of 13.8-22.4% for SSBE

and 2.2-10.1% for NSD; most noteworthy is the 32.7-36.2% improvement for SSBE diphthongs compared to only 3.4-19.5% improvement for NSD diphthongs. For both languages, using formant frequencies sampled from three time points generally performed slightly better than two time points (-0.3-2.5%).

Turning to Method B, using more than one DCT coefficient resulted in more accurate classifications, e.g., using both 0th and 1st DCT coefficients resulted in an overall improvements of 12.6% than when using just 0th DCT coefficients for SSBE but only 1.6% for NSD; the difference in improvement between NSD and SSBE is most striking for nominal diphthongs, which was 27.6% for SSBE and 0.4% for NSD. Using 2nd DCT coefficients (corresponding to a formant trajectory's curve) along with the 0th and 1st DCT coefficients (representing formant mean and slope, respectively) made little difference (-0.7-1.0%).

Throughout the results, a cross-linguistic difference has emerged: NSD vowel quality can be much better specified than SSBE without specific reference to spectral change, i.e., by using formant frequencies from a single time point (Method A) or by using 0th DCT coefficients (Method B), and this is especially the case for nominal diphthongs.

Overall, the two methods for modelling the onset + offset hypothesis of VISC produced comparable results. That is, Method B with 0th and 1st DCT coefficients resulted only in slightly greater accuracy than Method A with formant frequencies from onset and offset, i.e., 1.2% for NSD and 2.2% for SSBE.

Lastly, a closer look at the relatively poor classification of NSD true diphthongs suggests this arises from high confusion rates (> 20%) involving /ɛi/-/œy/ (cf., Figures 1 and 3). This has also been found in NSD listeners' vowel identification, as Van Leussen *et al*. [12] report 39% of /ɛi/ tokens (from the same corpus in this study) were misidentified as /œy/.

4. CONCLUSION

On the basis of non-AE data, modelling vowel quality in terms of VISC improves classification accuracy and the onset + offset hypothesis is supported, in line with [5]. Additionally, VISC modelling methods incorporating formant information from many points perform only marginally better than those using information from two or three time points, as also found for AE [5]. Notably, the cross-linguistic scope of the present study points to the relative importance of VISC for determining vowel quality being somewhat language-dependent, and the acoustic similarity of NSD /ɛi/-/œy/ presents a challenge to the tested VISC theories and methods.

Finally, this study could be extended with listener data testing the perceptual relevance of VISC in SSBE and NSD, and its interaction with other aspects of vowel production, e.g., fundamental frequency [10].

5. REFERENCES

[1] Adank, P., Van Hout, R., Smits, R. 2004. An acoustic description of the vowels of Northern and Southern Standard Dutch. *J. Acoust. Soc. Am.* 116, 1729–1738.

[2] Chládková, K., Hamann, S., Williams, D., Hellmuth, S. Under review. F2 trajectory as a perceptual cue for the front-back contrast in Standard Southern British English.

[3] Hillenbrand, J. M. 2013. Static and dynamic approaches to vowel perception. In: Morrison, G. S., Assmann, P. F. (eds), *Vowel Inherent Spectral Change*. Berlin-Heidelberg: Springer Verlag, 9–30.

[4] Hillenbrand, J. M., Nearey, T. M. 1999. Identification of resynthesized /hVd/ utterances: effects of formant contour. *J. Acoust. Soc. Am.* 105, 3509–3523.

[5] Morrison, G. S. 2013. Theories of vowel inherent spectral change. In: Morrison, G. S., Assmann, P. F. (eds), *Vowel Inherent Spectral Change*. Berlin-Heidelberg: Springer Verlag, 31–47.

[6] Morrison, G. S., Nearey, T. M. 2007. Testing theories of vowel inherent spectral change. *J. Acoust. Soc. Am.* 122, EL19–EL22.

[7] Nearey, T. M. 2013. Vowel inherent spectral change in vowels in North American English. In: Morrison, G. S., Assmann, P. F. (eds), *Vowel Inherent Spectral Change*. Berlin-Heidelberg: Springer Verlag, 49–85.

[8] Nearey, T., Assmann, P. 1986. Modeling the role of inherent spectral change in vowel identification. *J. Acoust. Soc. Am.* 80, 1297–1308.

[9] Peterson, G. E., Barney, H. L. 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Am*. 24, 175–184.

[10] Sims, M., Tucker, B. V., Nearey, T. M. 2012. Modelling vowel inherent spectral change in spontaneous speech. *Can. Acoust.* 40, 36–37.

[11] Stack, J., Strange, W., Jenkins, J., Clarke, W., Trent S. 2006. Perceptual invariance of coarticulated vowels over variations in speaking style. *J. Acoust. Soc. Am.* 119, 2394–2405.

[12] Van Leussen, J.-W., Escudero, P., Williams, D. 2011. The interrelation between the production and perception of Dutch vowels: real and modeled listeners. *9th ISSP* Montreal.

[13] Van Leussen, J.-W., Williams, D., Escudero, P. 2011. Acoustic properties of Dutch steady-state vowels: contextual effects and a comparison with previous studies. *Proc. 17th ICPhS* Hong Kong, 1194–1197.

[14] Williams, D., Escudero, P. 2014. A cross-dialectal acoustic comparison of vowels in Northern and Southern British English, *J. Acoust. Soc. Am.* 136, 2751–2761.

[1] The /tVk/ environment from the NSD corpora was excluded from the present study as there was not an equivalent environment in the SSBE data.