$F2_R$: A TECHNIQUE FOR COLLAPSING $F2_{onset}$ AND $F2_{mid}$ INTO A SINGLE ACOUSTIC ATTRIBUTE

Daniel McCarthy and Jalal Al-Tamimi

Newcastle University, UK
d.mccarthy1@newcastle.ac.uk jalal.al-tamimi@newcastle.ac.uk

ABSTRACT

The formant information for identifying place of articulation in voiced plosives is conventionally represented using two acoustic attributes, namely $F2_{onset}$ and $F2_{mid}$ (e.g. [7, 9, 16]). This study compares the accuracy of such a technique with a new technique in which $F2_{onset}$ and $F2_{mid}$ are collapsed into a single attribute, termed $F2_R$. This method involves subtracting $F2_{onset}$ from $F2_{mid}$, multiplying this frequency difference by a constant (*c*), and subtracting this product from $F2_{onset}$, yielding $F2_R$. Results from a discriminant analysis (leave-one-out cross-validation) show that $F2_R$ can distinguish the place of articulation of /b d g/ at approximately the same rate as the conventional method using $F2_{onset}$ and $F2_{mid}$.

Given that this result accords well with the 1950s locus theory [3], it suggests that the locus theory held an important insight that was neglected in phonetic science following Öhman's [15] findings for VCV sequences.

Keywords: place of articulation, voiced plosives, formants, locus equations.

1. INTRODUCTION

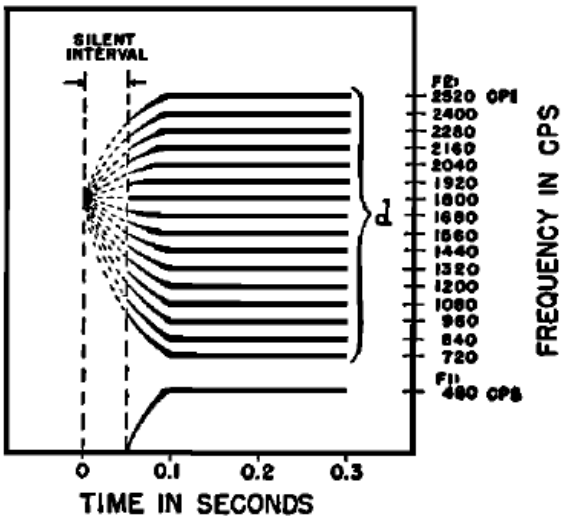
It is well established that the most important formant-based information for distinguishing voiced plosives' place of articulation lies in the second formant [16]. In particular, the frequencies of the second formant at vowel onset ($F2_{onset}$) and midpoint ($F2_{mid}$) have been recurrently used as attributes. One commonly used method of representing this information (e.g. [7, 9, 16]) is the locus equation, in which $F2_{onset}$ is represented along the vertical axis and $F2_{mid}$ along the horizontal axis for a variety of vowel contexts (with a line of regression fitted to the datapoints for each place of articulation).

The most striking finding of this research is that the regression lines for /b d/ show excellent fit to the datapoints [11]; for /g/ good fit is obtained if separate regression lines are plotted for front-vowel and back-vowel contexts [15].

The slopes of the regression lines for each place of articulation typically range from ca. 0.4 (for /d/) to 0.8 (for /b/) [16, p. 1314]. This indicates that $F2_{onset}$ and $F2_{mid}$ are moderately to highly correlated with each other, which suggests that some sort of collapsing of $F2_{onset}$ and $F2_{mid}$ into a single dimension might be feasible. Such an approach could minimize the number of features needed in speech recognition, an issue that has been noted by [14] and discussed lucidly by [4].

Has such a collapsing of $F2_{onset}$ and $F2_{mid}$ been proposed before? Indeed it has, and it is known as the locus theory [3, 8]. This theory posited that if a formant transition were traced backwards in time to approximately 50 ms prior to the beginning of the observed transition, then it would yield a frequency ($F2_{locus}$) that is specific to a given place of articulation, as shown in Figure 1:

Figure 1: Schematic diagram of the locus theory for a /d/ paired with a range of vowels that vary in backness. The F2 transitions for all the vowels begin at the same frequency of 1,800 Hz, at least if one traces their trajectory to an unobserved point in time approximately 50 ms prior to the vowel onset. This point is known as the F2 locus frequency ($F2_{locus}$). $F2_{locus}$ is posited to lie at a different frequency for the three places of articulation. Source: [3], p. 771.



However, confidence in the $F2_{locus}$ idea was shaken by Öhman's [15] investigation of V_1CV_2 sequences, which found that coarticulation from V_1 changed the formant transition in V_2 such that the transition in no

way pointed to an invariant frequency. For example the V_2 formant transition in [ybo] pointed upward whereas the one in [obo] pointed slightly downward (p. 160). This undermined the locus-theory belief that the transitions for a given place of articulation (in this example, bilabial) should point to the same frequency regardless of the vowel.

There are, however, a few observations to note about Öhman's study. The study was relatively small-scale (N = 225, all from a single speaker) and artificial: the study's author repeated nonce VCV sequences three times in a monotone with the vowels stressed equally. Because of this, it remains something of an open question whether the acoustic coarticulatory pattern Öhman found is also found in more naturalistic speech. A recent study by McCarthy [12] (summarized in [13]), using a much larger dataset (N = 758) from 20 speakers reading real speech found that, unlike Öhman's study, V_1 had only a modest acoustic influence on the $F2_{onset}$ of V_2 (the regression line between V_1F2_{mid} and V_2F2_{onset} had shallow slopes, to wit 0.12, 0.10 and 0.14 for /b d g/ respectively). Indeed, Lindblom and Sussman [11, p. 18] have argued that the widespread abandonment of the locus theory following Öhman's VCV findings was unfortunate.

The second point is that the 'locus' in the locus theory does not have to be an exact, invariant, pinpoint frequency: rather, we can loosen the definition of the $F2_{locus}$ to encompass a frequency *zone*, not a frequency *point*. Under this conception the preoccupation with finding an $F2_{locus}$ for each place of articulation that is perfectly invariant is bypassed in favour of finding $F2_{locus}$ zones for each place of articulation that are reasonably distinct from each other, sufficiently distinct to distinguish place of articulation at a decent rate. This conception is in the spirit of Lindblom's [10] championing of 'sufficient discriminability' in favour of invariance.

With all the above in mind it seems a revival of the (reframed) locus theory is warranted. The rest of this paper tests the locus theory by comparing its ability to distinguish place of articulation with the ability of $F2_{onset}$ (and $F2_{mid}$) to distinguish place.

2. METHODOLOGY

2.1. A formula for $F2_R$

We begin by presenting a formula for exploring $F2_{locus}$. Figure 1 illustrates the following generalization about all the /d/'s F2 transitions: the larger the difference in frequency between $F2_{onset}$ and $F2_{mid}$, the larger the difference in frequency between $F2_{onset}$ and $F2_{locus}$. This means that if we want to change $F2_{onset}$ into $F2_{locus}$ using $F2_{mid}$, the degree to which $F2_{onset}$ will have to change will depend on how large the difference in frequency is between $F2_{onset}$ and $F2_{mid}$. In other words, the first part of our technique is to subtract $F2_{onset}$ from $F2_{mid}$:

$$(1) \quad F2_{difference} = F2_{mid} - F2_{onset}$$

The second part of the technique is to subtract this $F2_{difference}$ from $F2_{onset}$:

$$(2) \quad F2_{reconstructed} = F2_{onset} - F2_{difference}$$

The output of (2) can be imagined as extrapolating the F2 transition backwards in time. Remember, however, that because we do not observe $F2_{locus}$, there is nothing to tell us exactly *how* far back in time we should go to obtain $F2_{locus}$. In Figure 1 above, the amount of time required (i.e. the part labelled "silent interval") is 50 ms, but this is a schematic diagram of artificial stimuli, not an empirical fact. Thus it seems wise to run a variety of $F2_{locus}$ formulae in which the degree to which $F2_{difference}$ modifies $F2_{onset}$ is varied by using a constant. Let us rewrite (2) as follows:

$$(3) \quad F2_{reconstructed} = F2_{onset} - (F2_{difference} \times c)$$

We shall refer to the family of attributes derived from (3) as "$F2_{reconstructed}$", or "$F2_R$" for short. The value of *c* will vary in increments of 0.2 from 0 to 3 to explore the space thoroughly, yielding 16 variants of $F2_R$.

2.2. Materials and Analysis

Speakers: 10 male and 10 female speakers of different varieties of British English were recruited. Their ages ranged from 18 to 38 at the time (2016). Accents represented included Yorkshire, Mancunian, Scouse, Geordie, Cockney, RP, and north Wales.

Recording: the material was read in an anechoic booth using a Roland Edirol R44-4 4-channel portable recorder, linked to a Roland Edirol CS-50 microphone (settings: 'lo cut' and 'focus'). Sampling frequency was 44.1 kHz with 16-bit quantization.

Material: 84 sentences were presented one by one on a screen to be read aloud. The subject matter was various everyday topics and the sentences were designed to contain as many plosives as was reasonable (ca. 14 per sentence). This yielded a corpus of 7,147 tokens. Some of these tokens were excluded (e.g. [ʔ] was not examined). Note also that the present paper is concerned only with the voiced prevocalic tokens in the corpus (N = 1,535). Vowels: front = 826, central = 280, back N = 429. Schwa tokens are not included in the present analysis.

Segmentation: each plosive, along with the preceding and following segment, was segmented

manually in Praat [2]. Five tiers were used: attribute, allophone, phoneme, word, and comment.

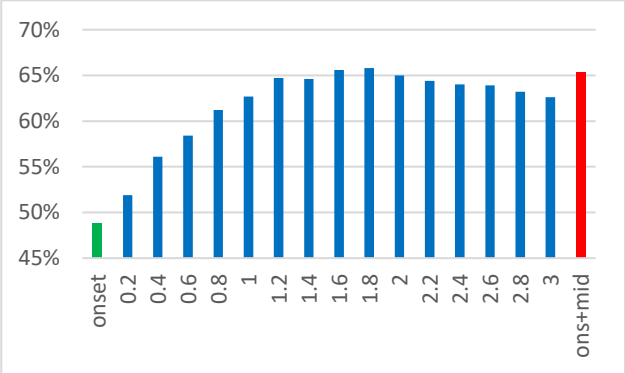
Measurements: F1, F2, and F3 frequencies were extracted from the onset and midpoint of the following segment and the offset and midpoint of the preceding segment. All data were extracted using a Praat script created by the second author.

Statistics: discriminant analyses (leave-one-out cross-validation) [5] were run in which /b d g/ were the three outcome variables and each variant of the $F2_R$ attribute was the predictor. The statistic quantifies the percentage of tokens classified correctly when each token is classified using all the dataset other than that token.

3. RESULTS

We begin with the results when $F2_R$ is used without any speaker normalization.

Figure 2: Cross-validated classification accuracy of $F2_R$ for distinguishing prevocalic /b d g/. The green bar shows the classification accuracy of $F2_{onset}$; the red bar ('ons+mid') shows the accuracy when $F2_{onset}$ and $F2_{mid}$ are separate attributes; and the blue bars represent the variants of $F2_R$, namely c = 0.2 to 3 increasing in increments of 0.2. N = 1,535.

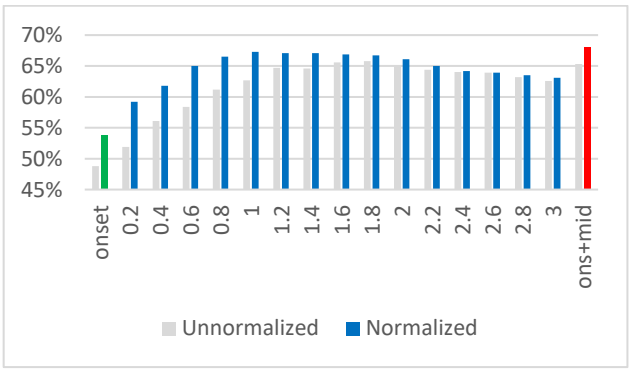


The classification accuracy of all variants of $F2_R$ surpasses that of $F2_{onset}$. This suggests that the 1950s locus theory held an important insight about formant transitions: when $F2_{mid}$ is higher in frequency than $F2_{onset}$, $F2_{onset}$ is dragged up whereas when $F2_{mid}$ is lower in frequency than $F2_{onset}$, $F2_{onset}$ is pulled down, and the size of this shift is proportional to the size of the frequency difference between the two.

Perhaps more importantly, the classification accuracy of $F2_R$ at its strongest (for values of c = 1.8) is as large as that of $F2_{onset}$ and $F2_{mid}$ (65.8% versus 65.3%).[i] This suggests that the collapsing of these two attributes into a single dimension can be achieved without compromising the classification accuracy.

We now quantify how much the above result is improved by normalizing the above formant frequencies for each individual speaker. The normalization consists of subtracting a speaker's mean F3 frequency from each token of F2. The theoretical reasoning behind this style of normalization is that, on a logarithmic scale (and the Bark scale is logarithmic in the F2 and F3 regions), the only difference in the formant pattern of a given vowel between two speakers of differing vocal tract length is in the location of the pattern along the frequency axis [18, p. 2375] (see [12] for details).

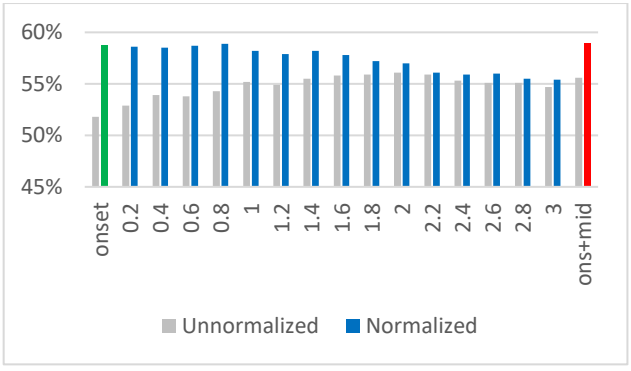
Figure 3: Comparison of the classification accuracy of $F2_R$ when normalized ($F2_R$ – $\mu F3_{individual}$) with the unnormalized data from Figure 2. N = 1,535.



Unsurprisingly the normalization improves the classification accuracy, by 2 to 3 percentage points. More interestingly we again see that the classification accuracy for normalized $F2_R$ at its strongest (for c = 1, 67.3%) is very similar to that of $F2_{onset}$ and $F2_{mid}$ (68.1%). It seems again, then, that $F2_{onset}$ and $F2_{mid}$ can be collapsed into a single attribute with little compromise of classification accuracy.

F3 is the other formant that provides information on place of articulation [17, pp. 250-251]. Here are the results when the F_R procedure is applied to F3:

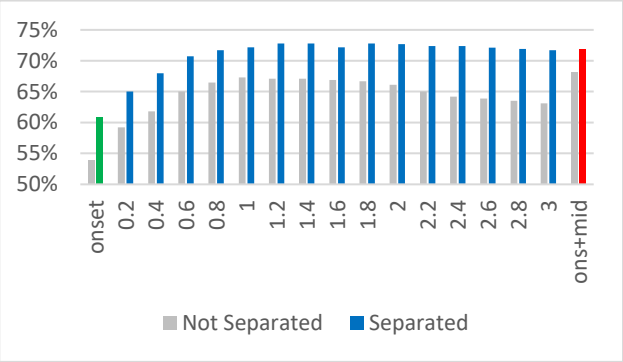
Figure 4: Comparison of the classification accuracy of $F3_R$ with and without normalization by individual speaker ($F3_R$ – $\mu F3_{individual}$). N = 1,535.



Unlike what we saw for F2, the classification accuracy of (normalized) $F3_{onset}$ is not increased by the inclusion of $F3_{mid}$. Because of this, the classification of $F3_R$ does not exceed that of $F3_{onset}$.

Segregating the classification of back and non-back vowels improves the classification of $F2_R$ considerably:

Figure 5: Classification accuracy of normalized $F2_R$ before and after separation by vowel backness. The results for non-back and back were run separately and summed. N = 1,535.



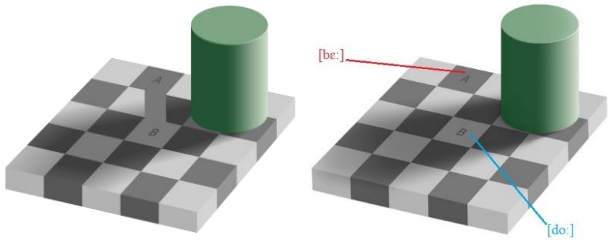
The peak classification accuracy of $F2_R$ improves by 6 percentage points with the separation by vowel backness. Presumably this improvement is a result of the fact that velars' F2 transitions have long been known to point to different locus frequencies before front vowels and back vowels [3]. Hence separating by vowel backness presumably prevents the two velar loci from being mixed together.

4. DISCUSSION

Our main finding is that the reduction of $F2_{onset}$ and $F2_{mid}$ to a single dimension is possible and produces an attribute with about the same classification accuracy as $F2_{onset}$ and $F2_{mid}$. This accords well with the 1950s locus theory which, as we saw in the Introduction, posited [3] that despite the smearing together of two phonemes' information in a formant transition, some semblance of invariance can be extracted from the transition if the imaginary $F2_{locus}$ frequency is used as output rather than the observed $F2_{onset}$ and $F2_{mid}$ frequencies. Nevertheless, we have cautioned against imagining $F2_R$ (or $F2_{locus}$) as yielding a pinpoint of a locus that is entirely free from vowel-induced coarticulation; $F2_R$ *mitigates* coarticulation, it doesn't *remove* it. This eschewing of absolute invariance is along the lines of Lindblom's [10] notion of sufficient discriminability.

One might wonder how plausible it is that imaginary frequencies be used in speech recognition. An analogy from vision might help. In Figure 6 the tiles labelled A and B are physically of identical intensity. Perceptually, however, A looks dark grey whereas B looks whitish. This is due to a perceptual mechanism in the visual system known as colour constancy [19], which separates out the distortion of lighting conditions on the colours of objects. Because of colour constancy, medium grey is perceived as dark grey if the surrounding context is bright (tile A in Figure 6) but the very same shade of grey is perceived as whitish if the surrounding context is dark (tile B) [1]. Analogously, $F2_{onset}$ is 1,750 Hz in both the syllables [bɛː] and [doː] but is perceived as bilabial in one and alveolar in the other. Just as the perceptual phenomenon of colour constancy makes a given light intensity darker in bright contexts and brighter in dark contexts, $F2_R$ makes $F2_{onset}$ lower in frequency in high-frequency contexts and higher in low-frequency contexts.

Figure 6: Analogy of colour constancy and $F2_{onset}$ variation. [1]



The results of the present study suggest that the locus theory held an important insight about how to mitigate the redundancy between $F2_{onset}$ and $F2_{mid}$.

5. CONCLUSION

Given the moderate to high correlation between $F2_{onset}$ and $F2_{mid}$ that has long been documented by locus-equation studies [8, 6, 14] this paper has collapsed $F2_{onset}$ and $F2_{mid}$ into a single acoustic attribute ($F2_R$), drawing inspiration from the 1950s locus theory [3]. It has been shown that such collapsing of $F2_{onset}$ and $F2_{mid}$ yields an attribute with approximately as strong a classification accuracy as $F2_{onset}$ and $F2_{mid}$. Given the abstract similarity of $F2_R$ to how colour constancy functions in vision, it is not implausible that an analogous mechanism could be found in speech perception.

6. ACKNOWLEDGEMENTS

Much of present paper is adapted from the first author's PhD thesis [12]. This research was funded by the United Kingdom's Economic and Social Research Council (award reference 1506273).

7. REFERENCES

[1] Adelson, E. H. 1995. Checker-Shadow Illusion. Retrieved August 27 2018, from http://persci.mit.edu/gallery/checkershadow

[2] Boersma, P., Weenink, D. 2014. *Praat: doing phonetics by computer*. Retrieved February 15, 2015, from http://praat.org.

[3] Delattre, P. C., Liberman, A. M., Cooper, F. S. 1955. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, *27*(4), 769–773.

[4] Huckvale, M. 1996. Learning from the experience of building automatic speech recognition systems. *UCL Working Papers in Speech, Hearing and Language*, *9*, 1–14.

[5] IBM. 2013. *Statistical Package for the Social Sciences*. Armonk, NY: IBM Corporation.

[6] Hasegawa-Johnson, M. A. 1996. *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification*. PhD thesis, Massachusetts Institute of Technology.

[7] Krull, D. 1987. Second formant locus patterns as a measure of consonant-vowel coarticulation. *Phonetic Experimental Research at the Institute of Linguistics, Stockholm University*, *V*, 43–61.

[8] Liberman, A. M. 1996. *Speech: A Special Code*. Cambridge, MA: MIT Press.

[9] Lindblom, B. 1963. Spectrographic Study of Vowel Reduction. *Journal of the Acoustical Society of America*, *35*(11), 1773–1781.

[10] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In *Speech Production and Speech Modelling*, pp. 403–439.

[11] Lindblom, B., Sussman, H. M. 2012. Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, *40*(1), 1–19.

[12] McCarthy, D. Forthcoming. *The Acoustics of Place of Articulation in English Plosives.* PhD thesis, Newcastle University.

[13] McCarthy, D. and Al-Tamimi, J. Submitted. The Acoustic Influence of V_1 on the Onset of V_2 in Intervocalic Voiced Plosives. *INTERSPEECH 2019 20th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria*, Proceedings.

[14] Morgan, N., Wegmann, S., Cohen, J. 2013. *What's Wrong With Automatic Speech Recognition (ASR) and How Can We Fix It?* Berkeley, CA: International Computer Science Institute.

[15] Öhman, S. E. G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*(1), 151–168.

[16] Sussman, H. M., McCaffrey, H. A., Matthews, S. A. 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, *90*(3), 1309–1325. https://doi.org/10.1121/1.401923

[17] Sussman, H. M., Fruchter, D., Hilbert, J., Sirosh, J. 1998. Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, *21*(2), 241–259.

[18] Turner, R. E., Walters, T. C., Monaghan, J. M., Patterson, R. D. 2009. A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *Journal of the Acoustical Society of America*, *125*(4), 2374-2386.

[19] Walsh, V., Kulikowski, J. J. 1998. *Perceptual Constancy: Why Things Look as They Do*. Cambridge: Cambridge University Press.

[i] Hasegawa-Johnson [6, p. 25] notes that studies of /b d g/ have typically yielded classification accuracies of 65% to 70%. The present result is thus well within the normal range of classification accuracy. The higher accuracy (77%) found by one study [16] may be due to the highly controlled nature of their stimuli, viz. /CVt/ words repeated five times in a carrier phrase.