

Title:

Theories of vowel inherent spectral change: A review

Running title:

vowel inherent spectral change

Author:

Geoffrey Stewart Morrison¹

Department of Linguistics, University of Alberta, Edmonton, Alberta, T6G 2E7, Canada.

Version date:

30 March 2008

¹ Current Address: School of Language Studies, Australian National University, Canberra, ACT 0200, Australia.

Abstract:

In many dialects of North American English, in addition to vowels which are traditionally described as true and phonetic diphthongs, several traditional monophthongs also have substantial formant movement, and vowel inherent spectral change has been found to be an important factor in vowel identification. The present paper reviews literature pertinent to theories of the perceptually relevant aspects of vowel inherent spectral change. Of the three basic hypotheses, formant onset + offset, formant onset + slope, and formant onset + direction, the weight of evidence indicates that the offset hypothesis is superior. Models with fit curves to whole formant trajectories have, as yet, not been found to outperform simple models based on formant measurements taken at two points (onset and offset) in formant trajectories. The most successful curve-fitting model is interpretable as a parameterisation of the onset + offset hypothesis.

Keywords: review; theory; diphthong; vowel inherent spectral change;

1. Introduction

The English vowel system traditionally comprises true diphthongs, e.g., /aɪ, aʊ, ɔɪ/, so called phonetic diphthongs, /e, o/ (often transcribed as /eɪ, ou/), and nominal monophthongs, e.g., /i, ɪ, ε, æ/. However, a growing number of studies have observed that several nominal monophthongs are in fact diphthongised, At least in many dialects of North American English. For example, Alabama: Fox & McGory (2007). Alberta: Andruski & Nearey (1992); Assmann, Nearey, & Hogan (1982); Nearey & Assmann (1986). Indiana: Hargus Ferguson & Kewley-Port (2002). Michigan: Hillenbrand, Clark, & Nearey (2001); Hillenbrand et al. (1995); Hillenbrand & Nearey (1999). Ohio: Fox (1983); Fox & McGory (2007). Texas: Assmann & Katz (2000). This also appears to be true for Dutch (Adank, van Hout, & Smits, 2004; Adank, van Hout, & van de Velde, 2007). Figure 1 illustrates the extent of *vowel inherent spectral change* (VISC) from the beginning to the end of phonetic diphthongs and nominal monophthongs produced by speakers of Alberta English. Given that traditional monophthongs can have substantial VISC, the present paper simply treats all traditional monophthongs, phonetic diphthongs, and true diphthongs in a unified manner as vowels which have some pattern of spectral change (including the possibility of no spectral change).

VISC has been found to play an important role in speech perception: Listeners' vowel identifications change when they are presented with stimuli that have typical formant trajectories versus flat formant trajectories versus reversed formant trajectories (Nearey & Assmann, 1986; Nearey, 1995; Hillenbrand & Nearey, 1999; Assmann & Katz, 2000, 2005). For example, when formant trajectories are reversed, /e/ stimuli may be identified as /ɪ/, and /ɪ/ stimuli as /e/ (Nearey & Assmann, 1986). Listeners also give higher goodness ratings to synthetic versions of nominal monophthongs that include VISC (Nearey, 1995). In addition, when pattern recognition models are provided with information about formant trajectories in nominal monophthongs and phonetic diphthongs, as compared to formant measurements from a single point, higher correct classification rates are obtained, and there is higher correlation with listeners' perception patterns (see below).

Three basic hypotheses have been advanced as to the aspects of VISC which are perceptually relevant for vowel identification. This paper reviews previously published work with the primary aim of determining which of the three hypotheses is most likely to be correct. It also considers whether a more complex models of VISC outperform models based on the basic hypotheses.

2. Three Basic Hypotheses

Three basic hypotheses have been advanced as to the aspects of VISC which are relevant for the perception of vowel identity (Gottfried, Miller, & Meyer, 1993; Nearey & Assmann, 1986; Pols, 1977). All three hypotheses agree that the initial formant frequencies are perceptually relevant to vowel identification (for supporting evidence see Gay, 1970; Bladon, 1985; Nábělek Czyzewski, & Crowley, 1993; Nearey, 1995), but disagree on what additional cues are relevant in VISC perception:²

- The *offset* hypothesis states that the relevant perceptual cues are the formant values at the end of the vowel. This may be expressed as the relative change in formant values from onset to offset, i.e., $[\Delta F1, \Delta F2]$ or $\Delta \mathbf{F}$
- The *slope* hypothesis states that the relevant perceptual cues are the velocities of formant change, i.e., whether the change in the frequency of each formant is positive or negative and the rate of change in time. This may be expressed as $\Delta \mathbf{F}/\Delta t$
- The *direction* hypothesis states that the only relevant factor is the direction of formant movement in an F1–F2 (or similar) space. This may be expressed as $\angle \Delta \mathbf{F}$ or $\Delta \mathbf{F}/\|\Delta \mathbf{F}\|$

Under the direction hypothesis, the rate of change over time and the final formant values achieved are irrelevant. Under the slope hypothesis, the direction and speed of formant change are relevant but the final formant values achieved are irrelevant. Under the offset hypothesis, the speed of formant change is irrelevant but the final formant values achieved (and by implication the direction) are relevant.

The difference between the hypotheses can be illustrated using the stylised formant vectors which appear in Figure 2:

- Vectors A, B, C, and D have the same direction in the F1–F2 plane, and under the direction hypotheses they should therefore all be perceived as the same category.

² The terminology used here is based on Gottfried, Miller, & Meyer's (1993) "onset + offset", "onset + slope", and "onset + direction". Nearey & Assmann's (1986) "dual-target", "target-plus-slope", and "target-plus-direction" represent the same hypotheses. Gottfried, Miller, & Meyer only tested F2 slope for their onset + slope hypothesis, but both F1 and F2 slopes were tested in studies conducted by Nearey and colleagues. In contrast to Lehiste & Peterson's (1961) use of the term *target*, Nearey & Assmann's (1986) term *dual-target* does not imply that there must be steady states at the beginning and end of the diphthongs.

- Vectors A, B, and D have the same rate-of-change in formant values and under the slope hypothesis should therefore all be perceived as the same category. The rate-of-change of vector C is half that of vectors A, B, and D, and under the slope hypothesis it should therefore be perceived as a different category.
- Vectors A, C, and D have the same end-point in the F1–F2 plane, and under the offset hypothesis they should therefore all be perceived as the same category. Vector B is twice as long in the F1–F2 plane as vectors A, C, and D, and under the offset hypothesis it should therefore be perceived as a different category.

A complication for the direction hypothesis is the issue of whether some minimum magnitude of formant change is needed. Tokens of a vowel category with negligible VISC may have random fluctuations in the direction of formant movement that are not perceptually pertinent (Nearey & Assmann, 1986). Perception of a vowel as a diphthong, as opposed to a monophthong, may also require some minimum duration for the glide portion of the vowel (see Nábělek, Czyzewski, & Crowley, 1994). Note, however, that the requirements of minimum thresholds for magnitude and duration of movement also apply to the offset and slope hypotheses. Kewley-Port & Goodman (2005) reported that, under near optimal conditions, the mean perceptual threshold for the magnitude of second-formant movement in front vowels ranged from 16 to 51 Hz for rising F2, and 24 to 66 Hz for falling F2 (initial F2 values were 2068, 2272, and 2525 Hz). These values were at least a factor of four smaller than the magnitude of F2 movement produced in natural US English /i, ɪ, e, ε, æ/ vowels, leading Kewley-Port & Goodman to conclude that the formant movement in these vowels would be detectable by listeners.

3. Studies Comparing Models Parameterised Using Dynamic Spectral Properties Versus Models Parameterised Using Static Spectral Properties

Several studies have determined that a model which includes some parameterisation of VISC outperforms a model which is based on static spectral properties. Such models lend support to one or other of the VISC hypotheses without necessarily considering alternative VISC hypotheses.

Some studies have found evidence in support of the slope hypothesis. Gay (1970) claimed that slope was the primary cue for distinguishing between different diphthongs, e.g., /ɔɪ/–/aɪ/; however, his synthetic stimuli confounded either offset and slope or duration and slope, and his set of experiments

did not allow full separation of the effects of slope from its covariants.³ Assmann, Nearey, & Hogan (1982) applied pattern recognition models to measurements of formant values of Canadian English nominal monophthongs and phonetic diphthongs. They found that when they included formant slope parameters in addition to midpoint formant parameters they obtained higher correct classification and, more importantly, higher correlation with listeners' response patterns.

Other studies have found evidence in support of the offset hypothesis. Hillenbrand et al., (1995), Andruski & Nearey (1992), Hillenbrand & Nearey (1999), and Hillenbrand, Clark, & Nearey (2001) obtained higher correct classification or higher correlation with listeners' response patterns when they used two-point (onset + offset) versus one-point (mean or midpoint) parameterisations of Canadian and US English nominal monophthongs and phonetic diphthongs (see also Adank, van Hout, & Smits, 2004, for Dutch vowels). Andruski & Nearey (1992) conducted experiments using silent-centre natural /bVb/ stimuli (short portions extracted from the beginning and end of natural productions), silent-centre natural isolated vowel stimuli, and synthetic /bVb/ stimuli in which the vowel portion was a linear interpolation from initial to final target values.⁴ Since similar perceptual results were obtained for all three stimulus types, they argued that the perceptually relevant cues were those shared by all three, i.e., the onset and offset values (this is also a possible interpretation of the results of Strange, Jenkins, & Johnson, 1983).⁵ Using a different methodology with US English true

³ The interpretation of Gay's (1970) results is hindered by contradictions between the description of his stimuli and the discussion of the results. Discussion and graphical results suggest that, in his Experiment II, F2 offset did not covary with duration so as to maintain a fixed slope, rather F2 offset stepped up at a slower rate than duration, e.g., for /ɔ/-/ɔɪ/ stimuli with an F2 onset of 840 Hz, the first three duration steps of 100, 110, and 120 ms all had an F2 offset of 1320 Hz, and thus progressively shallower slopes of 4.80, 4.36, and 4.00 Hz/ms; the next two duration steps of 130 and 140 ms both had an F2 offset of 1440 Hz, and thus slopes of 4.62 and 4.29; etc..

⁴ Although the classical description of a diphthong includes an initial steady state, a glide, and a final steady state (Lehiste & Peterson, 1961), there is usually no second steady state (see Holbrook & Fairbanks, 1962), the first steady state may disappear at fast speaking rates (Gay, 1968), and diphthongs can be synthesised using only a glide (Gay, 1970).

⁵ Fox (1989) found that, when presented with very short extracts from consonant transitions, listeners extrapolated the trajectories of vowels from the dynamic information available in the consonant transitions, rather than using the absolute formant values immediately before and after silent centres. This is not necessarily inconsistent with the onset + offset hypothesis if one assumes that listeners extrapolated the trajectories to include the characteristic onset and offset values / targets values of the vowel. The initial and final portions in Andruski & Nearey's (1992) silent centre stimuli were relatively long and may therefore have actually reached the target values.

diphthongs, phonetic diphthongs, and nominal monophthongs, Fox (1983) also obtained results consistent with the offset hypothesis. In a multidimensional scaling experiment, Fox extracted four perceptual dimensions: the first dimension was most highly correlated with F2 formant values measured at the end of the vowels, and the third dimension with F2 formant values measured at the beginning of the vowels.

Other studies have found evidence in support of an onset + midpoint + offset hypothesis; however, this has not been proven to be superior to the onset + offset hypothesis. Huang (1992) for non-back US English nominal monophthongs and /e/, and Harrington & Cassidy (1994) for Australian English diphthongs and nominal monophthongs, found that pattern classifiers based on three-point models (e.g., measurements taken at 25%, 50%, and 75% of vowel duration) outperformed one-point models (e.g., measurements taken at 50% of vowel duration). The authors did not claim that the three-point model was the absolute correct parameterisation, only that more than a one-point model was necessary.⁶ Hillenbrand et al. (1995) compared one-point, two point, and three-point parameterisations of US English nominal monophthongs. Substantially higher correct classification rates were obtained for two-point models compared to one-point models, but three-point models offered little or no improvement over two-point models.

4. Studies Comparing Two of the Hypotheses

Much of the work on theories of VISC has presented arguments against one of the hypotheses in an attempt to falsify it and leave one of the other hypotheses as the alternative.

Contra the offset hypothesis and in support of the slope hypothesis, Gay (1968) found

⁶ In a small-scale study Neel (2004) investigated the perception of synthetic 1, 2, 3, 5, and 11 point versions of US English phonetic diphthongs and nominal monophthongs. Each stimulus was based on the formant tracks from a single /dVd/ production from one of two speakers (problems with the study may be related to idiosyncrasies in the small number of productions). Two-point stimuli based on formant measurements at 10% and 90% of duration were poorly identified, typically at rates substantially worse than one-point stimuli based on formant measurements at 50% of duration. A possible reason for this is that the 10% and 90% points may actually have been in the consonant transitions and were therefore not representative of the vowels' characteristic onset and offset values: Identification rates were generally high for five-point stimuli based on formant measurements at 10%, 30%, 50%, 70% and 90% of duration, which would be expected since these stimuli included some approximation of initial consonant transition, vowel onset to vowel offset transition (via a midpoint value), and final consonant transition.

substantial speaking-rate dependent differences in final formant values and more consistency in slope (see also Borzone de Marique, 1979, for slope consistency in Spanish, and Pols, 1977, for direction consistency in Dutch); however, it could be argued that listeners are able to compensate for target undershoot, and that substantial variability in target may be unproblematic if there are only a few widely separated targets, and thus little chance of confusion between them (Bladon, 1985).

Contra the slope hypothesis and in support of the offset hypothesis, Bond (1978, 1982) found that changing the duration of the glide between onset and offset had little effect on vowel identification, and in some cases even deleting the glide completely had no effect (for glide deletion see also Wise, 1965; Bladon, 1985; Nearey & Assmann, 1986; Andruski & Nearey, 1992). In Kewley-Port & Goodman's (2005) study in the perceptual threshold for F2 movement perception, stimuli included long and short synthetic vowels, where a long vowel had the same slope as one of the shorter vowels but the same offset value as a different shorter vowel. They found no significant effect for duration, i.e., no difference in ΔF threshold values for long and short stimuli with the same end points but different slopes, leading them to conclude that their results supported the offset hypothesis over the slope hypothesis.

Contra the direction hypothesis and in support of the offset hypothesis, Bladon (1985) found that phonetically trained listeners transcribed truncated diphthongs with pairs of symbols appropriate for monophthongs at the initial and final formant values of the stimuli. The second symbol varied with the final formant values and was not invariant with direction. Unfortunately, Bladon's (1985) choice of stimuli make the relevancy of the results questionable: He removed the latter portions of /ia, iɛ, ie/, all three have similar initial formant values and a similar direction, but different final targets. However, it is not clear that /ia, iɛ, ie/ really are phonemes, i.e., that they are perceived holistically as single units rather than as a sequence of two phonemes. Although there are clearly some similarities, findings based on a sequence of two vowels (or a glide plus vowel) may have little relevance for the perception of true diphthongs, and even less relevance for phonetic diphthongs and nominal monophthongs. Jacewicz, Fujimura, & Fox (2003) found that listeners' responses shifted from /a/ to /aɪ/ as F2 offset was increased and from /aɪ/ to /ɛɪ/ as F2 onset was increased. Although they argued that the initial and final formant values did not characterise the diphthong, their interpretation of the results amounts to a thresholded version of the offset hypothesis: Once formant movement in a synthetic vowel has achieved roughly half the formant movement found in natural /aɪ/ productions, listeners fairly reliably identify the synthetic vowel as the diphthong /aɪ/. The threshold was much greater than the just

noticeable difference for F2 movement (Kewley-Port & Goodman, 2005), hence my claim that this is a thresholded version of the offset hypothesis rather than a thresholded version of the direction hypothesis.

5. Studies Testing All Three Hypotheses

Nearey & Assmann (1986) tested the three VISC hypotheses using pattern recognisers trained on different parameterisations of Canadian English nominal monophthongs and phonetic diphthongs. Parameters were initial F1 and F2 values plus: final F1 and F2 values for the offset hypothesis ($\Delta\mathbf{F}$); change in F1 and F2 values over the duration of glide for the slope hypothesis ($\Delta\mathbf{F}/\Delta t$); and change in F1 and F2 values each over the magnitude of the total change ($\Delta\mathbf{F}/\|\Delta\mathbf{F}\|$) for the direction hypothesis (all formant values were transformed to log-hertz prior to making any other calculations). Correlations with listeners' responses were slightly higher for the offset and direction parameterisations than for the slope parameterisation, but in general all three parameterisations provided adequate characterisations of listeners' response patterns.

Gottfried, Miller, & Meyer (1993) compared the three hypotheses using pattern recognisers trained on different parameterisations of US English phonetic and true diphthongs. They used two sets of parameterisations: one was similar to that of Nearey & Assmann (1986) in that it used log-F1 and log-F2 measurements, but differed in that only the F2 slope was included,⁷ and that the direction was specified as an angle in degrees ($\angle\Delta\mathbf{F}$), with adjustments made to avoid discontinuities at $0^\circ / 360^\circ$. The second set of parameters transformed F1, F2, and F3 values into Miller's *auditory-perceptual space* (APS: Miller, 1989). Across speaking conditions (slow stressed, slow unstressed, fast stressed, and fast unstressed) the log-formant parameterisations had slightly higher correct classification rates for the offset and slope hypotheses than for the direction hypothesis, and in the APS parameterisations the offset hypothesis had higher correct classification rates than the slope and direction hypotheses. However, no one hypothesis was superior to the others in all contexts.

Morrison & Nearey (2007) tested the three hypotheses using a synthetic Canadian English

⁷ Assmann & Katz (2000) tested the perception of stimuli in which the F1 trajectory was flattened and F2 unchanged, and stimuli in which the F2 trajectory was flattened and F1 unchanged. Listeners' correct identification rates for the set of US English nominal monophthongs and phonetic diphthongs significantly decreased when either formant was flattened. Although some vowels were affected more by F1 flattening, some were affected more by F2 flattening. The results indicate that a VISC theory applicable across vowel categories should refer to formant movement in both F1 and F2.

/i/-/e/-/ɪ/ continuum which had a fixed onset but systematic variation in duration, offset values, and slope. For each set of final formant values, two slope values were created by synthesising one stimulus with a straight line trajectory in the log-hertz F1–F2 space, and synthesising another with an elbow (piecewise linear). The elbowed stimuli had no formant movement during the first quarter of the vowel and straight transition from onset to offset values during the last three quarters. Pairs of elbowed and straight stimuli therefore had the same offset values, but the elbowed member of each pair had a slope which was a third steeper than the straight member. Direction was fixed to a single diagonal in the F1–F2 space, but stimuli had several magnitudes of positive and negative movement along this diagonal, thus multiple stimuli had the same direction but different offset values. Formant movement along the diagonal was either in the rising-F1–falling-F2 (/ɪ/-like) direction, the opposite falling-F1–rising-F2 (/e/-like) direction, or had zero formant movement (/i/-like direction). Listeners identified the stimuli, and general additive models were fitted to the perception results. Adding offset parameters to models already containing slope or direction parameters resulted in a significant improvement of fit to listeners’ responses, but adding slope or direction parameters to a model already containing offset parameters did not. Compared to the direction and slope hypotheses, the offset hypothesis therefore gave a better account of human listeners’ VISC perception.

6. Curve Fitting Parameterisations

Parameterisations based on formant measurements at two or three points in the vowel have been criticised as being crude measures incapable of capturing all the relevant details of inherently complex time-varying patterns (Clermont, 1993; Jenkins, Strange, & Miranda, 1994). A concrete problem is the choice of timepoints at which to measure formant values. Different studies have selected different points at which to measure the vowel onset and offset values, e.g., at the earliest and latest measurable values with amplitudes not less than 15dB below the vowel’s maximum amplitude (Nearey & Assmann, 1986), 40 ms after the initial consonant release and 40 ms before the final consonant closure (Andruski & Nearey, 1992), at 20% and 80% of the duration of the vowel (Hillenbrand & Nearey, 1999), and at 20% and 70% of the duration of the vowel (Hillenbrand, Clark, & Nearey, 2001). Gottfried, Miller, & Meyer (1993) measured at points immediately following and preceding the consonant transitions, which they determined on the basis of an algorithm which made use of the speed of formant movement. The choice of timepoints will clearly have an influence on the onset + offset parameterisation. It may also affect the slope parameterisation: If there is any steady-state portion

between the measurement points then the true slope will be underestimated (most studies using data based on acoustic measurements of productions have not attempted to divide vowels into steady-state and glide portions). If correct, the direction parameterisation is least likely to be affected. These philosophical and practical problems may be overcome by using more sophisticated curve-fitting parameterisations.

Zahorian & Jagharghi (1991, 1993) fitted *discrete cosine transforms* (DCTs) to the time-varying spectral properties of US English duration-equalised nominal monophthongs (although /e/ was excluded, /o/ was included).⁸ Spectral slices were parameterised as formant values and as cepstral coefficients. For both spectral-slice parameterisations, the highest correct classification rates and highest correlations with listeners' responses were obtained for models that included the first two DCT coefficients. The first DCT coefficient gives the mean value of a formant/cepstral coefficient over time, and the second coefficient is a measure of time-normalised slope of the formant/cepstral coefficient trajectory: a half period of a cosine is fitted to the values of the formant/cepstral coefficient measured from the beginning to the end of the vowel.⁹ When vowel durations are equalised, the value of a second DCT coefficient is therefore a symmetrically constrained measure of the direction and distance of the onset and offset of a formant/cepstrum measured relative to the mean value of that formant/cepstrum. This parameterisation is therefore similar to the onset + offset parameterisation, but based on a curve fitted to the whole trajectory rather than only two points. Models including DCT parameterised dynamic information outperformed one-point models in terms of correct-classification rates and correlation with human listeners' perception. Zahorian & Jagharghi (1993) reported also testing Legendre polynomial basis functions, and least-squares polynomial fitting, but obtained higher correct-classification rates for DCTs. The relative success of DCTs may be related to greater stability due to their edge constraints. No comparisons were made with models parameterised in terms of the basic offset, slope, or direction hypotheses.

Watson & Harrington (1999) fitted DCTs to formant trajectories from Australian English

⁸ Zahorian & Jagharghi reserved the term “discrete cosine transform” for what is normally referred to as a “cepstrum”, and used the term “discrete cosine series” for curves fitted to a time-ordered series of spectral frames.

⁹ The third DCT coefficient is a whole period of a cosine, the fourth one-and-a-half periods etc.. In the present paper ordinal numbering starts with the first coefficient (the mean or dc offset), but in some papers (e.g., Harrington, 2006) this coefficient is referred to as the zeroth coefficient, the half cosine is referred to as the first coefficient, etc..

vowels. They obtained higher correct-classification rates for models using the first and second DCT coefficients compared to models using only the first, but no significant additional improvement when the first three coefficients were used. Watson & Harrington made no quantitative comparisons with offset, slope, or direction parameterisations.

Hillenbrand, Clark, & Nearey (2001) reported fitting polynomials and DCTs to formant trajectories from US English nominal monophthongs and phonetic diphthongs, and comparing the results to two-point parameterisations. They concluded that in terms of correct-classification rates the curve-fitting parameterisations were not superior to the simpler two-point parameterisation. Thus, there has been no proof that more sophisticated curve-fitting parameterisations are superior to the dual-target parameterisation with respect to the substantive issues of correct classification and correlation with listeners' responses.

7. Conclusion

The weight of evidence in the literature reviewed here indicates that the onset + offset hypothesis provides a better account of the perceptually relevant aspects of VISC with respect to vowel identity than either the slope or the direction hypotheses. In terms of correct-classification rates and correlation with listeners responses, more sophisticated curve-fitting parameterisations of VISC have not, as yet, been found to outperform simple two-point parameterisations of the onset + offset hypothesis. The most successful curve-fitting model, discrete cosine transform, is interpretable as a parameterisation of the onset + offset hypothesis. One should not, however, conclude that onset + offset models of VISC will necessarily capture all perceptually relevant information: For example, perception of vowel-plus-consonant diphones may require more detailed descriptions of VISC (Jacewicz, Fujimura, & Fox, 2003; Moreton, 2004), and more complex VISC models are more effective for speaker identification (McDougall, 2006; Morrison, 2008).

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada. Thanks to Terrance M. Nearey for comments on an earlier version of this paper.

References

- Adank, P. van Hout, R., & Smits R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America*, 116, 1729–1738. DOI: 10.1121/1.1779271.
- Adank, P., van Hout, R., and van de Velde, H. (2007). An acoustic description of the vowels of Northern and Southern Standard Dutch II: Regional varieties. *Journal of the Acoustical Society of America*, 121, 1130–1141. DOI: 10.1121/1.2409492.
- Andruski, J. E., & Nearey, T. M. (1992). On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables. *Journal of the Acoustical Society of America*, 91, 390–410. DOI: 10.1121/1.402781.
- Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America*, 71, 975–989. DOI: 10.1121/1.387579.
- Assmann, P. F. & Katz, W. F. (2000). Time-varying spectral change in the vowels of children and adults. *Journal of the Acoustical Society of America*, 108, 1856–1866. DOI: 10.1121/1.1289363.
- Assmann, P. F. & Katz, W. F. (2005). Synthesis fidelity and time-varying spectral change in vowels. *Journal of the Acoustical Society of America*, 117, 886–895. DOI: 10.1121/1.1852549.
- Bladon, A. (1985). Diphthongs: A case study of dynamic articulatory processing. *Speech Communication*, 4, 145–154. DOI: 10.1016/0167-6393(84)90040-2.
- Bond, Z. S. (1982). Experiments with synthetic diphthongs. *Journal of Phonetics*, 10, 259–264.
- Bond, Z. S. (1978). The effects of varying glide duration on diphthong identification. *Language & Speech*, 21, 253–278.
- Borzone de Manrique, A.M. (1979). Acoustic analysis of Spanish diphthongs. *Phonetica*, 36, 194–206.
- Clermont, F. (1993). Spectro-temporal description of diphthongs in F_1 – F_2 – F_3 space. *Speech Communication*, 13, 377–390.
- Fox, R. (1989). Dynamic information in identification and discrimination of vowels. *Phonetica*, 46, 97–116.
- Fox, R. (1983). Perceptual structure of monophthongs and diphthongs in English. *Language & Speech*, 26, 21–49.
- Fox, R. A., & McGory, J. T. (2007). Second language acquisition of a regional dialect of American

- English by native Japanese speakers. In O.-S. Bohn, & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 117–134). Amsterdam: John Benjamins.
- Gay, T. (1968). Effects of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America*, *44*, 1570–1573. DOI: 10.1121/1.1911298.
- Gay, T. (1970). A perceptual study of American English diphthongs. *Language & Speech*, *13*, 65–88.
- Gottfried, M., Miller, J. D., & Meyer, D. J. (1993). Three approaches to the classification of American English diphthongs. *Journal of Phonetics*, *21*, 205–229.
- Hargus Ferguson, S., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, *112*, 259–271. DOI: 10.1121/1.1482078.
- Harrington, J. (2006) An acoustic analysis of ‘happy-tensing’ in the Queen’s Christmas broadcasts. *Journal of Phonetics* *34*, 439–457. DOI: 10.1016/j.wocn.2005.08.001.
- Harrington, J., & Cassidy, S. (1994). Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English. *Language & Speech*, *37*, 357–373.
- Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111. DOI: 10.1121/1.411872.
- Hillenbrand, J. M., Clark, M. J., & Nearey, T. N. (2001). Effect of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, *109*, 748–763. DOI: 10.1121/1.1337959.
- Hillenbrand, J. M., & Nearey, T. N. (1999). Identification of resynthesized /hVd/ syllables: Effects of formant contour. *Journal of the Acoustical Society of America*, *105*, 3509–3523. DOI: 10.1121/1.424676.
- Holbrook, A., & Fairbanks, G. (1962). Diphthong formants and their movements. *Journal of Speech and Hearing Research*, *5*, 38–58.
- Huang, C. B. (1992). “Modelling human vowel identification using aspects of format trajectory and context,” in *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Ohmsha, Tokyo / IOS, Amsterdam), pp. 43–61.
- Jacewicz, E., Fujimura, O., & R. A. Fox (2003). Dynamics in diphthong perception. In . J. Sole, D. Recasens, & J. Romero(Eds.), *Proceedings of the 15th International Congress of Phonetic*

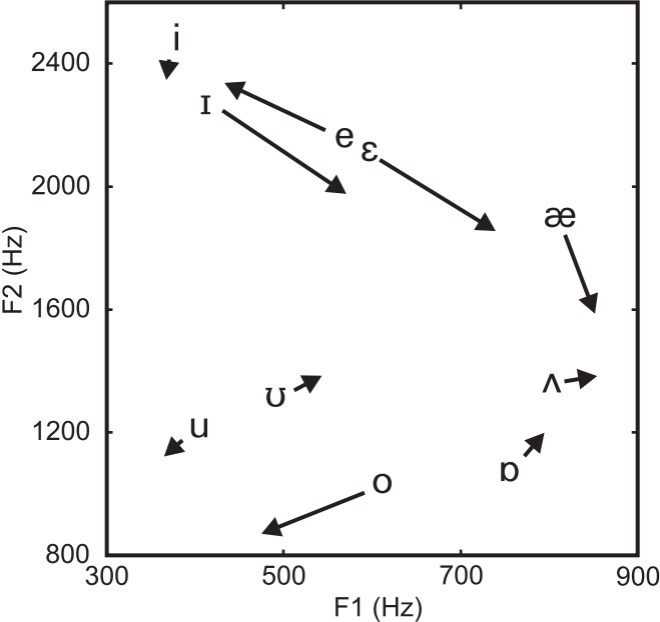
- Sciences, Barcelona* (pp. 993–996). Rundle Mall, SA, Australia: Causal Productions.
- Jenkins, J. J., Strange, W., & Miranda, S. (1994). Vowel identification in mixed-speaker silent-center syllables. *Journal of the Acoustical Society of America*, *95*, 1030–1043. DOI: 10.1121/1.410014.
- Kewley-Port, D., & Goodman, S. G. (2005). Thresholds for second formant transitions in front vowels. *Journal of the Acoustical Society of America*, *118*, 3252–3560. DOI: 10.1121/1.2074667.
- Lehiste, I., & Peterson, G. E. (1961). Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America*, *33*, 268–277. DOI: 10.1121/1.1908681.
- McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. *Speech, Language and the Law*, *13*, 89–126.
- Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, *85*, 2114–2134. DOI: 10.1121/1.397862.
- Moreton, E. (2004). Realization of the English postvocalic [voice] contrast in F1 and F2. *Journal of Phonetics*, *32*, 1–33. DOI: 10.1016/S0095-4470(03)00004-4.
- Morrison, G. S. (2008). Forensic speaker recognition using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. Manuscript submitted for publication.
- Morrison, G. S., & Nearey, T. M. (2007). Testing theories of vowel inherent spectral change. *Journal of the Acoustical Society of America*, *122*, EL15–EL22. DOI: 10.1121/1.2739111.
- Nábělek, A. K., Czyzewski, Z., & Crowley, H. (1994). Cues for perception of the diphthong /aɪ/ in either noise or reverberation. Part I. Duration of the transition. *Journal of the Acoustical Society of America*, *95*, 2681–2693. DOI: 10.1121/1.409837.
- Nábělek, A. K., Czyzewski, Z., & Crowley, H. (1993). Vowel boundaries for steady-state and linear formant trajectories. *Journal of the Acoustical Society of America*, *94*, 675–687. DOI: 10.1121/1.406885.
- Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*, 2088–2113. DOI: 10.1121/1.397861.
- Nearey, T. M. (1995). Evidence for the perceptual relevance of vowel-inherent spectral change for front vowels in Canadian English. In K. Elenius, & P. Branderud (Eds.), *Proceedings of the 13th Congress of Phonetic Sciences, Stockholm*, (pp. 678–681). Stockholm, Sweden: KTH.
- Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of vowel inherent spectral change in

- vowel identification. *Journal of the Acoustical Society of America*, 80, 1297–1308. DOI: 10.1121/1.394433.
- Neel, A. T. (2004). Formant detail needed for vowel identification. *Acoustic Research Letters Online*, 5, 125–131. DOI: 10.1121/1.1764452.
- Pols, L. C. W. (1977). Spectral analysis and identification of Dutch vowels in monosyllabic words. PhD dissertation, University of Amsterdam. Amsterdam: Academische Pers B. V.
- Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, 74, 695–705. DOI: 10.1121/1.389855.
- Watson, C., & Harrington, J. (1999). Acoustic evidence of dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, 106, 458–468. DOI: 10.1121/1.427069.
- Wise, C. M. (1964). Acoustic structure of English diphthongs and semivowels *vis-a-vis* their phonetic symbolization. In E. Zwirner, & W. Bethge (Eds.), *Proceedings of the 5th International Congress on Phonetic Sciences, Münster* (pp. 589–593). Basel, Switzerland: S. Karger.
- Zahorian, S. A., & Jagharghi, A. J. (1991). Speaker normalisation of static and dynamic vowel spectral features. *Journal of the Acoustical Society of America*, 90, 67–75. DOI: 10.1121/1.402350.
- Zahorian, S., & Jagharghi, A. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, 94, 1966–1982. DOI: 10.1121/1.407520.

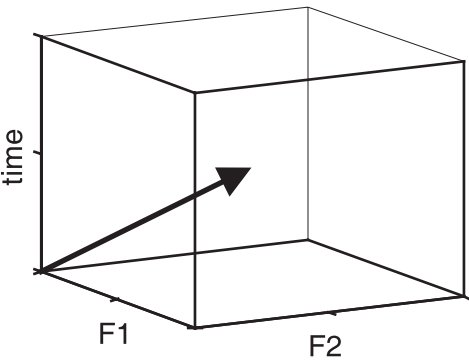
Figure captions:

Fig. 1. Diphthongisation of traditional monophthongs and phonetic diphthongs in Western-Canadian English. Adapted from Nearey & Assmann (1986). Arrow tails: Mean formant values measured at the beginning of the vowel. Arrow heads: Mean formant values measured at the end of the vowel.

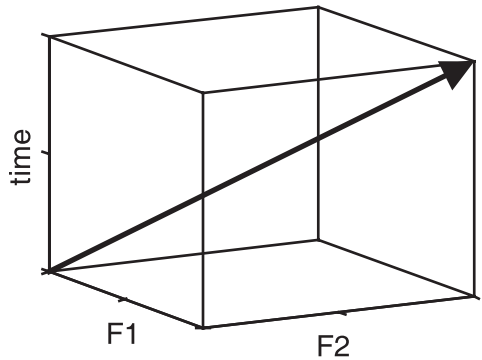
Fig. 2. Stylised formant vectors $[\Delta F1, \Delta F2, \Delta \text{time}]$ used to illustrate the differences between the three basic hypotheses, see text. Vector A: $[1, 1, 1]$. Vector B: $[2, 2, 2]$. Vector C: $[1, 1, 2]$. Vector D: piecewise $[0, 0, 1] + [1, 1, 1]$.



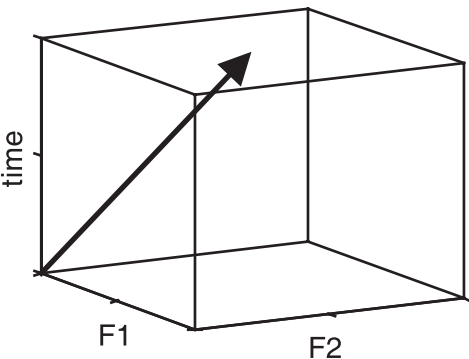
A



B



C



D

