Spectral change and duration as cues in Australian English listeners' front vowel categorization

Daniel Williams, Paola Escudero, and Adamantios Gafos

Citation: The Journal of the Acoustical Society of America **144**, EL215 (2018); doi: 10.1121/1.5055019

View online: https://doi.org/10.1121/1.5055019

View Table of Contents: https://asa.scitation.org/toc/jas/144/3

Published by the Acoustical Society of America

Spectral change and duration as cues in Australian English listeners' front vowel categorization

Daniel Williams,[1,a)] Paola Escudero,[2,b)] and Adamantios Gafos[1]

[1]*Linguistics Department, University of Potsdam, Haus 14, Karl-Liebknecht-Straße 24-25,*
14476 Potsdam, Germany
[2]*MARCS Institute, Western Sydney University, Building 1, Bullecourt Avenue, Milperra,*
New South Wales 2214, Australia
daniel.williams@uni-potsdam.de, paola.escudero@westernsydney.edu.au,
gafos@uni-potsdam.de

1904

Abstract: Australian English /iː/, /ɪ/, and /ɪə/ exhibit almost identical average first ($F1$) and second ($F2$) formant frequencies and differ in duration and vowel inherent spectral change (VISC). The cues of duration, $F1 \times F2$ trajectory direction (TD) and trajectory length (TL) were assessed in listeners' categorization of /iː/ and /ɪə/ compared to /ɪ/. Duration was important for distinguishing both /iː/ and /ɪə/ from /ɪ/. TD and TL were important for categorizing /iː/ versus /ɪ/, whereas only TL was important for /ɪə/ versus /ɪ/. Finally, listeners' use of duration and VISC was not mutually affected for either vowel compared to /ɪ/.

[BHS]
Date Received: May 21, 2018 **Date Accepted:** September 4, 2018

1. Introduction

Time-varying spectral information is very influential in vowel perception (Strange, 1989). By examining the perception of North American English (AE) co-articulated syllables, Strange and colleagues have repeatedly shown that identification accuracy remains high when the "steady-state" targets of vowels have been set to silence, leaving only their spectrally dynamic onsets and offsets (for a review, see Strange and Jenkins, 2013). Vowels may also display vowel inherent spectral change (VISC), patterns of spectral change characteristic of vowels themselves (Nearey and Assmann, 1986; Nearey, 2013), and this too can be perceptually relevant. For instance, Hillenbrand and Nearey (1999) found that AE vowels are more intelligible when their formant trajectories are spectrally dynamic like those in naturally produced AE vowels compared to when formant trajectories are spectrally flat. At the same time, the perceptual importance of VISC may not be uniform and its effects are generally greatest for AE vowels which are themselves produced with relatively large magnitudes of VISC (e.g., Hillenbrand and Nearey, 1999).

Despite growing evidence for the relevance of dynamic spectral information in vowel perception, it is an unresolved issue as to how it is best described (for a discussion, see Morrison and Nearey, 2007). It may be sufficient to refer to differences in formant frequencies between vowel onset and offset. On the other hand, it may be more appropriate to consider the formant frequencies themselves without referring to how they change. A further unresolved issue is the role of vowel duration in perceived vowel identity. In an identification task involving naturally produced AE vowels whose durations had been manipulated to be longer or shorter, Hillenbrand *et al.* (2000) found that some—but ultimately few—vowels were frequently misidentified. The authors hypothesize that listeners generally gave little weight to duration, as most AE vowels can be contrasted with their neighbors by spectral characteristics alone.

The neighboring Australian English (AusE) front vowels /iː/, /ɪ/, and /ɪə/ (as in the English words "bead," "bid," and "beard") provide a striking example of how both VISC and duration cues may potentially be exploited by listeners in vowel perception because the three vowels are spectrally very close in terms of mean or midpoint formant frequencies. Elvin *et al.* (2016) collected a corpus of vowels in the Western Sydney variety of AusE and examined their durations as well as their first ($F1$), second ($F2$), and third ($F3$) formant trajectories. Table 1 shows average values for /iː/, /ɪ/, and /ɪə/ as produced by male speakers. Table 1 also shows average $F1 \times F2$ *trajectory*

[a)]Author to whom correspondence should be addressed.
[b)]Also at: ARC Centre of Excellence for the Dynamics of Language, Western Sydney University, Locked Bag 1797, Penrith, New South Wales 2751, Australia.

Table 1. Average acoustic information for /iː/, /ɪ/, and /ɪə/ as produced by male AusE speakers reported in Elvin *et al.* (2016). TL values were calculated by summing the 29 Euclidean distances between *F*1 and *F*2 frequencies sampled from 30 time points across the middle 60% of each vowel and TD assignment was based on average *F*1 × *F*2 values shown in Fig. 1.

Vowel	Duration (ms)	*F*1 (ERB)	*F*2 (ERB)	*F*3 (ERB)	TL (ERB)	TD
/iː/	168	8.25	21.73	23.93	1.60	Diverging
/ɪ/	101	8.57	21.15	23.52	0.52	Converging
/ɪə/	206	8.43	21.52	23.70	1.09	Converging
Mean	158	8.42	21.40	23.72	1.07	—

lengths (TLs), which are the Euclidean distances between vowel onset and offset in a *F*1 × *F*2 vowel space (cf. Fox and Jacewicz, 2009), i.e., the amount of *F*1 × *F*2 frequency change between the beginning and end portions of vowel segments. Additionally, Table 1 shows the three vowels' *trajectory directions* (TDs), that is, the course of *F*1 × *F*2 frequency change. Here "converging" refers to *F*1 and *F*2 frequencies moving closer together in vowel offset, while "diverging" denotes *F*1 and *F*2 frequencies moving apart in vowel offset (Morrison and Nearey, 2007). Figure 1 illustrates the three vowels' TDs and TLs in a *F*1 × *F*2 vowel space and, as can be seen, /iː/ and /ɪə/ exhibit roughly similar TLs, but their TDs proceed in opposite directions, /iː/'s is Diverging and /ɪə/'s is Converging.

To examine which acoustic parameters best separate /iː/, /ɪ/, and /ɪə/, Elvin *et al.* (2016) conducted discriminant analyses and found that using duration alone classified 70.2% of tokens correctly, with a notably high accuracy of 93.3% for /ɪ/. Formant means classified only 52.9% of tokens correctly, while values for a measure of VISC [the first discrete cosine transform coefficient which corresponds to the magnitude and direction (sign) of formant slope] classified 77.5% of tokens correctly, notably higher at 89.6% for /iː/. When duration and the VISC values were entered simultaneously, 93.1% of all three vowels' tokens were classified correctly, suggesting that the combination of these two acoustic parameters is more effective at separating /iː/, /ɪ/, and /ɪə/ than either formant means or duration values alone.

Given that AE listeners exploit both VISC and sometimes duration cues to distinguish neighboring vowels (e.g., Nearey and Assmann, 1986; Hillenbrand *et al.*, 2000), it is reasonable to expect AusE listeners will make use of these cues too. Examining the relative importance of individual cues as well as their combinations should provide insights into how VISC and duration are both used for identifying vowels and for distinguishing between them. In line with Elvin *et al.* (2016) and previous findings for AE, we predict that AusE listeners will be sensitive to durational differences and that this will be important for distinguishing /ɪ/ from /iː/ and /ɪə/. We also expect VISC cues will be perceptually most prominent for /ɪə/ and /iː/, as their TDs proceed in opposite directions and their TLs are much larger compared to /ɪ/. Finally, we compare the *relation* between *F*1 × *F*2 values at vowel onset and offset by means of both TLs and TDs. Thus, the present study investigates the relative use of the concurrent cues of duration and VISC and how to characterize VISC for perceiving spectrally neighboring vowels.

2. Methods

2.1 Participants

Twenty native monolingual speakers of AusE (12 female, 8 male) were recruited from Western Sydney University's subject pool who, at the time of testing, were young adults under the age of 30 (age range: 17 yrs 1 month—27 yrs 9 months; median: 20 yrs 10 months).

2.2 Stimuli

The stimuli were isolated vowel segments created using the Klatt synthesizer (Klatt and Klatt, 1990) in Praat (Boersma and Weenink, 2016). The acoustic values of the stimuli were based on those of male speakers' vowels in the corpus of Western Sydney AusE (Elvin *et al.*, 2016). The vowel stimuli all shared the same midpoint formant values, which are the mean frequencies in Table 1, and varied in:

- *Duration*: four logarithmic steps; one unit was $e^{0.32}$ ms, yielding values of 94, 129, 177, and 244 ms;

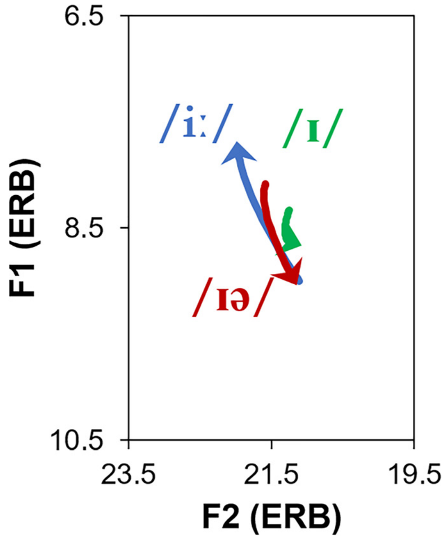

Fig. 1. (Color online) Average $F1 \times F2$ trajectories for male AusE speakers from the corpus of Elvin *et al.* (2016).

- *TD* (*Direction of change between onset and offset in F1 × F2 space*): Diverging, Zero or Converging;
- *TL* [Euclidean distance between onset and offset in F1 × F2 space using the Equivalent Rectangular Bandwidth (ERB) measure]: six steps which were 0.00, 0.30, 0.90, 1.50, and 2.10 ERB with an additional 3.90 ERB "exaggerated" step.

This procedure yielded 44 different vowels, as schematized in the left panel of Fig. 2; the right panel shows the ranges of $F1 \times F2$ values in the vowels' onsets and offsets. $F3$ was kept constant at the mean in Table 1, while $F4$ was constant at 25.55 ERB (Praat's suggested value for a male voice).

The 44 vowels in the present experiment served in another task for which different fundamental frequencies ($f0$s) were required. Each vowel was therefore synthesized with three male $f0$s (3.25, 3.80, and 4.90 ERB);[1] the final stimulus set thus consisted of 132 physically different stimuli (44 vowels × 3 $f0$s).

2.3 Procedure

On each trial of a multiple-alternative forced-choice task, participants heard one of the stimuli over headphones and selected one of three responses displayed on a computer screen—*bead* (= /iː/), *bid* (= /ɪ/), or *beard* (= /ɪə/)—by pressing a corresponding keyboard key. After a practice round of 15 trials, the task started. The 40 vowels containing spectral change (10 spectrally dynamic vowels × 4 durations) were presented on six trials, while the four spectrally static vowels (1 spectrally static vowel × 4 vowel durations) were presented on 12 trials. This resulted in 288 trials. Trial order was randomized across participants and breaks were given after every 48 trials.

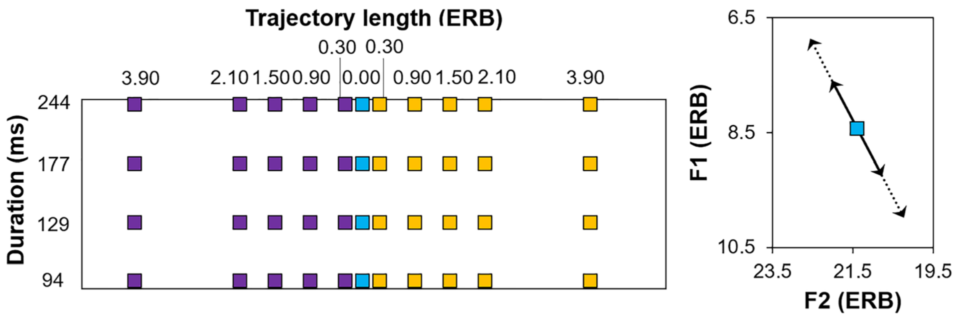

Fig. 2. (Color online) *Left panel:* The 44 vowel stimuli along the TL and Duration dimensions with the TD dimension shown by different colors (Diverging = purple, Zero = blue, Converging = orange). *Right panel:* $F1 \times F2$ vowel space of the stimuli. The square shows the midpoint shared by all stimuli. Most (36/44) vowel stimuli exhibit TLs < 2.10 ERB with onsets and offsets sampled between the arrowheads at the ends of the solid line. The remaining eight "exaggerated" stimuli exhibit onsets and offsets between the arrowheads at the ends of the dotted line.

Table 2. Counts and proportions (in parentheses) of vowel labels assigned to the vowel stimuli according to Duration, TD, and TL. Note that the count of responses was not the same across cue dimensions. For each level of Duration, responses summed to 1440. For TD, responses to Diverging and Converging stimuli each totaled 2400, while responses to Zero stimuli totaled 960. For each level of TL, responses added up to 960.

Duration	/iː/	/ɪ/	/ɪə/	TD	/iː/	/ɪ/	/ɪə/	TL	/iː/	/ɪ/	/ɪə/
94 ms	347	962	131	Diverging	1562	601	237	0.00 ERB	263	418	279
	(0.24)	(0.67)	(0.09)		(0.65)	(0.25)	(0.10)		(0.27)	(0.44)	(0.29)
129 ms	482	670	288	Zero	263	418	279	0.30 ERB	265	439	256
	(0.33)	(0.47)	(0.20)		(0.27)	(0.44)	(0.29)		(0.28)	(0.46)	(0.27)
177 ms	634	216	590	Converging	271	904	1225	0.90 ERB	364	345	251
	(0.44)	(0.15)	(0.41)		(0.11)	(0.38)	(0.51)		(0.38)	(0.36)	(0.26)
244 ms	633	75	732					1.50 ERB	382	287	291
	(0.44)	(0.05)	(0.51)						(0.40)	(0.30)	(0.30)
								2.10 ERB	404	255	301
									(0.42)	(0.27)	(0.31)
								3.90 ERB	418	179	363
									(0.44)	(0.19)	(0.38)

3. Results

Table 2 displays counts and proportions of how the vowel stimuli were categorized according to Duration, TD, and TL. As can be seen, listeners favored the /ɪ/ label for short vowels and the /iː/ and /ɪə/ labels for vowels of longer durations. With respect to TD, listeners chose /ɪ/ most often for Zero vowels and /iː/ and /ɪə/ with roughly equal frequencies for the remainder of responses. For Diverging vowels, listeners clearly preferred selecting /iː/, while for Converging vowels the /ɪ/ and /ɪə/ labels were chosen more often. For TL, the general trend is for vowels with less spectral change to be labeled as /ɪ/ and vowels with greater spectral change to be labeled as /iː/ or /ɪə/.

To determine the extent to which the cues of Duration, TD, and TL contributed to listeners' categorization of the vowel stimuli as /iː/, /ɪ/, or /ɪə/, we used the *MCMCglmm* package (Hadfield, 2010) in the program *R* (R Core Team, 2017) which fits generalized mixed-effects models utilizing Markov chain Monte Carlo sampling for Bayesian statistics. A mixed-effects multinomial logistic regression was fit to the data, which estimates the log-odds for several alternative outcomes occurring compared to one particular outcome (a reference category) based on a given set of fixed and random effects. We selected /ɪ/ to be the reference category because—in speech—it is the shortest of the three vowels, displays the least amount of spectral change, and its onset and offset lie between those of /iː/ and /ɪə/ (Fig. 1). The fixed factors were Duration, TD, and TL and their interactions. By-participant random intercepts and slopes were entered for the three fixed factors, their interactions and *f*0[2], as these were repeated across subjects (Barr *et al.*, 2013). Duration and *f*0 values were mean-centered around zero, while TL values were ordered from zero upwards (i.e., 0.00, 0.30, 0.90, 1.50, 2.10, and 3.90 ERB). The levels of TD were coded as −1, 0, and 1 to correspond to Diverging, Zero, and Converging, respectively. Subsequently, to ensure comparability across cues, Duration, TD, TL. and *f*0, all values were standardized by dividing them by their respective standard deviations. In this way, the model's intercept represents the estimated log-odds of selecting the reference category /ɪ/ when VISC cues are absent (i.e., both TD and TL are zero) and when Duration is at its mean. Thus, each fixed effect or interaction represents the estimated log-odds of selecting the alternative categories (/iː/ or /ɪə/) with each one-standard-deviation shift along the respective cue dimension(s). Last, model coefficients were sampled from the posterior distribution using four Markov chains conditioned on priors recommended in Hadfield (2017) appropriate for this kind of model.[3]

Figure 3 plots the posterior coefficients of the model. Reliable effects are those with 95% credible intervals which do not cross zero. As predictors were standardized, comparisons can be made regarding the relative importance of Duration, TD, and TL and their interactions for the log-odds of selecting /iː/ and /ɪə/ relative to the reference category /ɪ/. That is, coefficient values can be ranked from most important (largest) to least important (smallest) in selecting /iː/ or /ɪə/ over the reference category /ɪ/. If listeners' use of one cue is moderated by the value(s) of at least one other cue, interactions should be apparent. A lack of interactions, on the other hand, would suggest that a particular cue provides robust information to vowel identity without being dependent on the values of other cues.

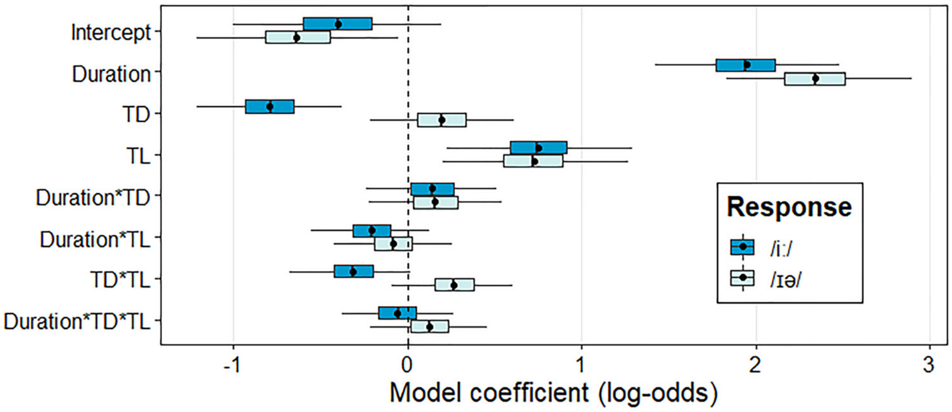

Fig. 3. (Color online) Model fixed-effect coefficients from the posterior distribution sampled by four Markov chains. The boxes represent the interquartile range, the vertical line within each box represents the posterior median, and the whiskers represent the 95% credible interval. The dots within each box represent the posterior mean.

Of all model coefficients, Duration was by far the most important cue for selecting both /ɪ/-/iː/ and /ɪ/-/ɪə/, and more important than VISC cues. Recall that this effect indicates the influence of vowel duration on the labeling of those vowel stimuli without spectral change, i.e., when TD and TL are zero. As the log-odds of selecting the alternative vowel over the reference category are positive for /ɪ/-/iː/ and for /ɪ/-/ɪə/, AusE listeners preferred both /iː/ and /ɪə/ to exhibit longer durations than /ɪ/, mirroring speech production. In fact, the effect is larger for /ɪ/-/ɪə/, indicating that vowel duration is more important for contrasting this vowel pair. This also reflects speech production because, of the three AusE front vowels, /ɪ/ and /ɪə/ display the shortest and longest durations, respectively (Table 1).

Turning to the two VISC cues, it is immediately clear that these were used differently depending on the vowel pair in question. First, while TD was expected to be equally important for /iː/ and /ɪə/, there is a significant effect only for /ɪ/-/iː/. As the TD coefficient is negative, listeners preferred the alternative category /iː/ to have a more Diverging $F1 \times F2$ trajectory than the reference category /ɪ/, which mirrors speech production (Fig. 1). On the other hand, TD does not appear to contribute to the categorization of /ɪə/ compared to /ɪ/, presumably because these two AusE vowels are both produced with Converging $F1 \times F2$ trajectories. Second, there are reliable effects of TL and the positive coefficients indicate that, as expected, listeners were more likely to select the /iː/ or /ɪə/ labels than the /ɪ/ label for vowels with greater spectral change. For /ɪ/-/iː/, the sizes of the effects of TL and TD are roughly similar, indicating both cues are of approximately equal importance.

There is some, albeit less reliable, evidence to support a TD × TL interaction involving /iː/, as the 95% credible intervals only just cross zero. In fact, the 95% credible intervals of the posterior distribution from one of the four Markov chains came close to but did not cross zero. The direction of the TD × TL interaction for /iː/ is as expected: the negative coefficient indicates listeners' preferences for selecting this option over the reference category /ɪ/ were stronger as TLs become larger and when TD is Diverging. However, the size of the TD × TL interaction for /iː/ is much smaller than the independent effects of either TD or TL, indicating that the interaction between the two VISC cues is much less important for selecting this vowel over the reference category /ɪ/.

Finally, none of the interactions involving Duration and TD and/or TL are stable (Duration × TD, Duration × TL and Duration × TD × TL). Therefore, it can be concluded that the influence of vowel duration on vowel categorization does not clearly depend on the particular acoustic values of VISC cues.

4. Discussion

AusE listeners were clearly sensitive to both duration and the two VISC cues in front vowel categorization involving stimuli with identical midpoint and mean formant frequencies. Notably, the use of three cues was not uniform. As expected, vowel duration was most significant for selecting /iː/ or /ɪə/ relative to /ɪ/, reflecting that, in speech production, /ɪ/ is shorter than /iː/ and /ɪə/. Also in line with speech production, AusE listeners preferred /iː/ or /ɪə/ to show greater levels of spectral change than /ɪ/. Furthermore, a Diverging $F1 \times F2$ trajectory is evidently important for perceiving /iː/,

whereas the direction of a vowel's $F1 \times F2$ trajectory is not specifically used for categorizing /ɪə/ as different from /ɪ/, presumably because both exhibit Converging trajectories in speech.

Morrison and Nearey (2007) hypothesize that the use of VISC in vowel categorization is best explained by the formant frequency values in vowel onset and offset. The present study finds support for this hypothesis with regard to the *relation* between vowel onset and offset. AusE /iː/ and /ɪə/ display large TLs in speech production, i.e., their onsets and offsets exhibit markedly different formant frequencies, and this cue was, as expected, perceptually important. However, TD was not used in the same way for the two vowels. Since formant frequency differences between onset and offset proceed in the same direction for /ɪ/-/ɪə/, this particular onset-offset relation may not act as an effective cue for differentiating the two vowels, leaving the size of the onset-offset formant frequency difference, in addition to duration, as a more crucial cue. As formant frequency differences between the onsets and offsets for /ɪ/ and /iː/ proceed in opposite directions, this relation successfully distinguishes the two vowels just as well as its size. Thus, the ways in which onsets and offsets explain vowel categorization is dependent on the vowel and contrast in question.

Hillenbrand *et al.* (2000) hypothesize that the use of vowel duration is influenced by the degree to which a given vowel can be spectrally distinguished from its neighbors. In the case of AusE /iː/, /ɪ/, and /ɪə/, while duration is undoubtedly a critical cue, the use of duration versus VISC may function in a "complementary" fashion. Specifically, duration was of greater importance for /ɪ/-/ɪə/ compared to /ɪ/-/iː/. At the same time, only TL was important for /ɪ/-/ɪə/, while for /ɪ/-/iː/ TL and TD were of approximately equal importance. As /ɪ/-/ɪə/ are not distinguished from one another by TD, duration may therefore serve as a more informative cue than for /ɪ/-/iː/.

The classification analyses of Elvin *et al.* (2016) were partly motivated by reports of monophthongized variants of AusE /ɪə/ as [ɪː], possibly due to sociophonetic variation or sound change (e.g., Cox, 2006). The present study offers some insights into why this variation is more plausible than, say, monophthongized variants of AusE /iː/. A less spectrally dynamic but relatively long realization of /ɪə/ is unlikely to be detrimental to its perception as the intended vowel than a less spectrally dynamic realization of /iː/. This is because duration is used more strongly for /ɪə/ than for /iː/ and, concurrently, VISC cues, such as TD, are more important for /iː/.

The present findings also shed some light on AusE acquisition. It is likely that learning to exploit VISC cues is challenging for infants because, as the present results reveal, their use may involve attending to more than one dimension rather than a single dimension (e.g., Goudbeek *et al.*, 2008). Escudero *et al.* (2017) report that AusE 15-month-olds habituate more readily to AusE /ɪ/ than to AusE /iː/ when contained in a /dVt/ frame, leading to the successful detection of a switch from /dɪt/ to /diːt/, but not from /diːt/ to /dɪt/. As VISC cues are presumably less important for successfully recognizing /dɪt/ compared to /diːt/, an inability to attend to VISC cues therefore should not impede detecting a switch from /dɪt/ to /diːt/.

To conclude, VISC and duration are crucial acoustic cues for explaining the perception of neighboring front vowels in AusE, and the present findings were largely predictable based on acoustic patterns found in naturally produced vowels (Elvin *et al.*, 2016). The current results on AusE listeners' categorization of /iː/, /ɪ/, and /ɪə/ add to the evidence that vowel perception involves, in addition to vowel duration, attending to time-varying spectral information rather than simply to static targets. Moreover, the relative importance and incorporation of VISC cues can vary across vowels and complements duration, highlighting that perceptual categorization involves attending to and integrating multiple cues in different ways.

Acknowledgments

This work was supported by an Endeavour Research Fellowship (No. 4699_2015), the Australian Research Council Centre of Excellence for the Dynamics of Language (Grant No. CE140100041), and an Advanced Grant from the European Research Council (Grant No. 249440).

References and links

[1] In vowel perception, higher *f*0s result in slightly poorer identification accuracy than lower and mid-range *f*0s (Ryalls and Liberman, 1982), but this effect is negligible in close vowels (Hirahara, 1988).

[2] *f*0 was included as a by-subject random effect as it was not a variable of interest.

[3] A normal distribution prior was used for the fixed effects and an inverse-Wishart distribution was used for variance and covariance. Priors for residual covariance were identical to those recommended by Hadfield

(2017, p. 97) for a multinomial logistic regression involving three outcome categories. For both fixed and random effects, the degree of belief parameter was set to the lowest bound to give the weakest possible priors, and full covariances were estimated for the random effects. The four Markov chains began sampling at different starting points. For each chain, the first 50 000 iterations were discarded and the next one million were thinned with an interval of 1000, leaving 1000 samples per chain. Convergence was determined by visual inspection of traces and posterior distributions and by negligible autocorrelation between samples. Also, the Gelman-Rubin diagnostic, which evaluates differences between chains, showed all four chains converged on the same posterior distribution (all <1.008).

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Mem. Lang. **69**, 255–278.

Boersma, P., and Weenink, D. (**2016**). Praat: Doing phonetics by computer [Computer program]. Version 6.0.13, http://www.praat.org/ (Last viewed February 6, 2016).

Cox, F. (**2006**). "/hVd/ vowels in the speech of some Australian teenagers," Aust. J. Linguist. **26**, 147–179.

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

Escudero, P., Mulak, K., Elvin, J., and Traynor, N. M. (**2017**). " 'Mummy, keep it steady': Phonetic variation shapes word learning at 15 and 17 months," Develop. Sci. #e1260.

Fox, R. A., and Jacewicz, E. (**2009**). "Cross-dialectal variation in formant dynamics in American English vowels," J. Acoust. Soc. Am. **126**, 2603–2618.

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

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

Hadfield, J. D. (**2017**). "MCMCglmm Course Notes," https://cran.r-project.org/web/packages/MCMCglmm/ (Last viewed November 4, 2017).

Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," J. Acoust. Soc. Am. **105**, 3509–3523.

Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (**2000**). "Some effects of duration on vowel recognition," J. Acoust. Soc. Am. **108**, 3013–3022.

Hirahara, T. (**1988**). "On the role of fundamental frequency in vowel perception," J. Acoust. Soc. Am. **84**, S156.

Morrison, G. S., and Nearey, T. M. (**2007**). "Testing theories of vowel inherent spectral change," J. Acoust. Soc. Am. **122**, EL15–EL22.

Nearey, T. M. (**2013**). "Vowel inherent spectral change in vowels in North American English," in *Vowel Inherent Spectral Change*, edited by G. S. Morrison and P. F. Assmann (Springer, Berlin), pp. 49–85.

Nearey, T. M., and Assmann, P. F. (**1986**). "Modeling the role of vowel inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**, 1297–1308.

R Core Team (**2017**). R: A Language and Environment for Statistical Computing. Version 3.3.3, retrieved March 6, 2017 from https://cran.ma.imperial.ac.uk/.

Ryalls, J. H., and Liberman, P. (**1982**). "Fundamental frequency and vowel perception," J. Acoust. Soc. Am. **72**, 1631–1634.

Strange, W. (**1989**). "Evolving theories of vowel perception," J. Acoust. Soc. Am. **85**, 2081–2087.

Strange, W., and Jenkins, J. J. (**2013**). "Dynamic specification of coarticulated vowels: Research chronology, theory and hypotheses," in *Vowel Inherent Spectral Change*, edited by G. S. Morrison and P. F. Assmann (Springer, Berlin), pp. 87–115.