Quality and quantity in New Zealand English vowel contrasts

Paul Warren

Victoria University of Wellington, New Zealand
paul.warren@vuw.ac.nz

Acoustic analysis of word-list and sentence data from a database of spoken New Zealand English is used to address the suggestion that a number of vowel contrasts in this variety are moving towards a quantity-based distinction. Data from 73 speakers across three age groups and two sexes are analysed, for a total of almost 6,000 vowel tokens. The move to a quantity distinction is confirmed for the STRUT–START contrast. In the case of DRESS and FLEECE, younger and female speakers show greater overlap of the two vowels but also stronger diphthongisation of FLEECE. The data also confirm the fronting of FOOT away from THOUGHT and towards GOOSE, but with no evident adjustment in durational contrasts or in the diphthongisation of either long vowel. KIT is becoming increasingly centralised, especially in the speech of young women, and is as a consequence becoming more distinct from NURSE. The findings are discussed in the context of the system of vowel contrasts in this variety.

1 Introduction

This paper presents acoustic analyses of four sets of vowels in contemporary New Zealand English (NZE): DRESS–FLEECE, KIT–NURSE, STRUT–START, and FOOT–GOOSE–THOUGHT.[1] Using an apparent-time approach, with samples from three age groups (as well as female and male speakers in each group), it explores claims that there have been changes in the overlap between vowels in each set in F1-F2 formant space, i.e. in vowel quality, and investigates associated changes in other characteristics that might distinguish these vowels, namely in their relative durations and in the degree of diphthongisation of the traditionally long vowels. The remainder of this introduction presents a description of the relevant parts of the NZE vowel system and a discussion of the parameters along which the vowels in question might be distinguished. Section 2 introduces the methodology and materials used in the current analyses, the results of which are presented in Section 3. The concluding discussion in Section 4 considers again the parameters distinguishing the vowels in the four sets analysed, in light of the results of the current study.

The NZE monophthongal vowel system is illustrated in Figure 1 (from Bauer et al. 2007). A number of developments in the pronunciation of New Zealand English (NZE) indicate that the phonemic contrast between the members of a small set of pairs in this monophthongal

---

[1] The lexical set labels introduced by Wells (1982) will be used to refer to the vowels of interest.
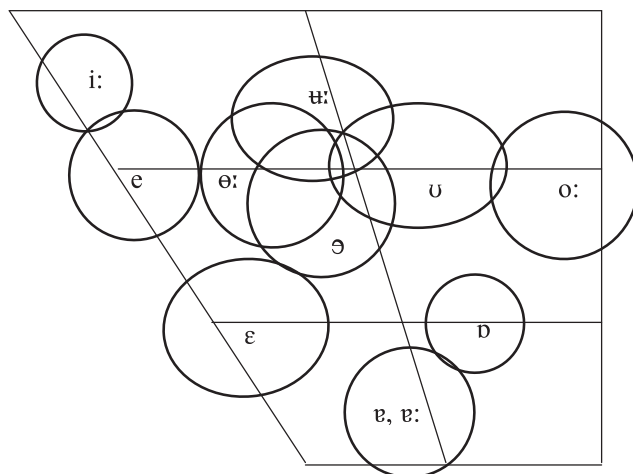
Figure 1 Approximate positions of New Zealand English short and long monophthongs (from Bauer et al. 2007, reproduced with permission. Note that the symbols are not centroids, but labels positioned to indicate clearly the different distributions.)

vowel system may be shifting towards vowel duration, and away from quality differences defined in terms of F1-F2 space. This has been suggested for the pairs DRESS–FLEECE, STRUT–START, FOOT–THOUGHT, and KIT–NURSE (Easton & Bauer 2000: 113; Bauer & Warren 2004: 588–589). While Figure 1 indicates complete overlap of STRUT and START (transcribed as /ɐ/ and /ɐː/ respectively, using the NZE transcription system suggested by Bauer & Warren 2004), DRESS (/e/) and FLEECE (/iː/) are shown to have minimal overlap, while FOOT (/ʊ/) overlaps partially with two long vowels, THOUGHT (/oː/) and GOOSE (/ʉː/). The somewhat crowded centre of the vowel space in NZE includes the substantial overlap of KIT (/ə/; KIT is famously centralised in NZE) with NURSE (/ɵː/), as well as the overlap of the latter with GOOSE. Previous work relating to each of these sets will now be presented.

In their description of NZE phonology, Bauer & Warren (2004: 588) state that '[i]f vowels are to be paired in terms of length/tension, then in New Zealand English the DRESS vowel should be paired with the FLEECE vowel, as being the closest long vowel in terms of quality'. Note that these authors make no distinction between tension and length in their treatment of any of the vowel sets under discussion. The focus in the current paper is on the relationship between quality distinctions and durational distinctions of the vowel pairs under consideration. However, I will return in the Discussion section to a consideration of the role of tense vs. lax in the description of the NZE vowel system.

It has been noted for some time (e.g. Maclagan 1982) that instead of the long/short pair of FLEECE and KIT that is found in British English Received Pronunciation, NZE appears to have a long/short pair in FLEECE and DRESS. Given this observation and the statement above, the representation of these vowels in Figure 1 would seem very conservative. In fact, Maclagan & Hay (2007) show that there is total overlap of DRESS and FLEECE in vowel space for many younger speakers, and it is even the case for some speakers that their DRESS vowel is closer (higher) and/or fronter than their FLEECE vowel. Whether or not the lack of a quality contrast is being compensated for by a durational contrast is somewhat uncertain. On the one hand, Maclagan & Hay (2004) found generally higher overall FLEECE:DRESS duration ratios for younger than for older NZE speakers, suggesting a possible increasing reliance on duration to distinguish the vowels. However, more detailed analysis by Maclagan & Hay (2007) indicates that the two vowels might not in fact be reliably distinguished by length, since there is considerable overlap between the durations of DRESS and FLEECE; in particular, many speakers in their sample have FLEECE vowels preceding voiceless coda consonants that

are not any longer than their DRESS vowels in the same contexts and are shorter than their DRESS vowels before voiced coda consonants. They also observe that for many speakers the distinction between DRESS and FLEECE now also involves the diphthongisation of FLEECE, with a noticeable on-glide, and that this is particularly likely when FLEECE is followed by a voiceless consonant, i.e. in the context where FLEECE is shortest and therefore most confusable with DRESS. Similar comments were made earlier by Easton & Bauer (2000: 100), who summarised that 'length and diphthongisation are becoming the major features distinguishing these two vowels for some speakers'. These two properties are reflected in the two transcription symbols suggested for FLEECE by Bauer & Warren (2004), namely [iː] and [ɪi].

For our second vowel pair, KIT and NURSE, Bauer & Warren (2004: 588) observe that for some NZE speakers *bid* and *bird* are distinguished only by length, and are a quantity pair for such speakers.[2] A high degree of overlap of these two vowels in the F1-F2 vowel space is also evident in the analysis presented for NZE (but not for Australian English) by Watson, Harrington & Evans (1998). Bauer & Warren also note that lip-rounding for NURSE is not very marked in NZE, but that the two vowels may nevertheless be distinguished on the basis of some rounding-like property. They observe that one of the speakers analysed for their study showed similar acoustic data for the NURSE vowel in the words *herd*, from a /hVd/ word list, and *word*, from an impromptu remark, but had more evident rounding (as shown in video recordings) in the more formal context of the word-list recordings. The authors comment that 'there would appear to be some other compensatory articulatory configuration that results in the rounded quality in the absence of rounded lip shape' in many productions of NURSE (Bauer & Warren 2004: 583).

The overlap of STRUT and START has been noted for some time, both in NZE and also in Australian English (in the latter as early as Bernard 1967). STRUT and START are similar in quality to one another in a number of other varieties, such as Fijian English (Tent & Mugler 2004), Singapore English (Wee 2004) and Malaysian English (Baskaran 2004). However, they tend to be back vowels in these varieties and are often not distinguished by length, while in both Australian and NZ English they are fronted and there is a reliable length difference between the two. The overlap in vowel quality is reflected also in the transcription symbols suggested for both of these varieties. As indicated above, Bauer & Warren (2004) in their description of NZE phonology recommend /ɐ/ and /ɐː/. The same symbols were recommended for both NZE and Australian English (AusE) by Watson et al. (1998). Easton & Bauer (2000) argue that STRUT and START are backer in AusE than in NZE, which supports Clark's (1989) suggestion of /ɐ/ and /ɐː/ as appropriate transcription symbols for these vowels in AusE. For NZE, Bauer & Warren (2004: 588) conclude that '[i]f vowels are to be paired in terms of length/tension, then in New Zealand English the STRUT vowel should be paired with the START vowel, with which it is virtually identical in terms of formant structure, resulting in a distinction primarily of length between *cut* and *cart*'. There appears to be little evidence in the literature that either STRUT or START are developing a diphthongal quality. In AusE this pair has been identified as the 'clearest example of a minimal length contrast', since for most other vowels 'length is correlated with a clear tendency to diphthongisation' (Durie & Hajek 1995: 231).

If we turn now to FOOT/THOUGHT/GOOSE, Easton & Bauer (2000: 113) observe that FOOT and THOUGHT overlap 'to a considerable extent' in the vowel space, while Bauer & Warren (2004: 589) make a slightly stronger claim that any pairing of FOOT on the basis of length should be with THOUGHT 'with which it is sometimes virtually identical in terms of formant structure, so that *put* and *port* may differ only in vowel length'. However, younger speakers of NZE have a more fronted FOOT than older speakers, so that while FOOT is closer in vowel space to THOUGHT for older speakers, this is not the case for younger speakers, for whom there is suggestion of overlap in acoustic space between FOOT and GOOSE (Kennedy 2004, Warren

---

[2] It should be noted that NZE is largely non-rhotic.

2004). Further, NZE has both conservative and innovative versions of FOOT – the conservative variant is back (though centralised) and slightly unrounded (Bauer & Warren 2004: 589 suggest [ʊ]), while the innovative form is central and unrounded (for which they suggest [ɨ]). This innovative form is found especially frequently in certain lexical items, such as 'good', as is reflected in the frequent spelling <gidday> for the greeting 'good day'. It has been claimed that FOOT and THOUGHT are less distinct durationally for younger than older speakers, though THOUGHT is still twice as long as FOOT, even for those speakers with the lowest THOUGHT:FOOT ratio (Kennedy 2004). Both long vowels in this set show evidence of diphthongisation, with an on-glide for GOOSE and an off-glide for THOUGHT. Bauer et al. (2007: 101) claim that THOUGHT is 'often diphthongised in lengthening environments'. Kennedy's (2004) small-scale study of the relationship between FOOT, GOOSE and THOUGHT in NZE reports that of these three vowels only THOUGHT showed any sign of diphthongisation, reinforcing her conclusion that for younger speakers there is a greater likelihood that FOOT overlaps with GOOSE than with THOUGHT. Easton & Bauer (2000) had however earlier reported that their speakers show a trend towards diphthongisation for GOOSE, supporting their claim that if there is a quantity pairing then it involves FOOT with THOUGHT rather than with GOOSE.

In summary, then, while there is consensus that NZE STRUT and START overlap in the vowel space and are the most likely pair to be distinguished only on the basis of length (Warren 2006), the case for the other contrasts in the sets of vowels under consideration is less clear. DRESS and FLEECE show more extreme overlap for younger speakers, so that the depiction in Figure 1 is more appropriate for older and more conservative speakers. However, there is evidence that the two vowels are becoming distinguished for younger speakers by increasing diphthongisation of FLEECE (Maclagan & Hay 2007). KIT and NURSE have been claimed to overlap in F1-F2 space, and yet are possibly distinguished for some speakers by a rounded quality to NURSE that is not found with KIT (Bauer & Warren 2004). FOOT and THOUGHT, which at one stage were looking likely to be a quantity pair, appear to be more distinct in quality for younger speakers than for older speakers, with a possible move towards FOOT being paired with GOOSE (Kennedy 2004, Warren 2004). But again, diphthongisation may have a role to play here, with both THOUGHT and GOOSE showing increasing evidence of diphthongal realisations (with off-glides and on-glides respectively) that would render them more distinct from the monophthongal short FOOT.

While there may be historical reasons for the treatment of NZE vowels as involving an opposition based on length or tension (reflecting the origins of the variety and the early settlement of New Zealand predominantly from southern England), recent changes in the NZE vowel system, and the observations above relating to these changes, suggest that it might be timely to revisit the nature of the vowel contrasts in this variety. With this in mind, the current paper presents acoustic analyses of tokens from the New Zealand Spoken English Database (NZSED, see Warren 2002), considering in particular contrasts in both vowel quality and vowel duration for the four vowel sets described above. Vowel quality here includes both the position of vowels in the F1-F2 vowel space, and – where appropriate – the nature and extent of any diphthongisation. As is clear from the descriptions above, earlier analyses of these vowel sets exist and these have helped to frame the current analysis. Importantly, however, in the analyses presented below all of the vowel tokens are produced by the same groups of speakers, so that the subsequent discussion can consider the vowels as part of a single vowel system, and can consider the relevance of the treatment of vowel contrasts in this variety in terms of either length or tension.

2 Method

To confirm the patterns noted above for the four sets of potentially quantity-distinguished vowels, new analyses were carried out on data collected over the period 1999–2002 as part of the New Zealand Spoken English Database (NZSED; Warren 2002). The NZSED materials

include recordings from three age groups (as well as from both females and males) and can therefore also be employed in an apparent-time study investigation of the extent to which NZE is moving towards quantity-based contrasts in the sets under consideration. NZSED was designed as a replication for NZE of the Australian National Database of Spoken Language (ANDOSL, see Millar et al. 1990) and consists of data from a number of speech tasks, including word lists where the traditionally monophthongal vowels are realised as the V in a /hVd/ sequence, and a set of 200 phonetically-rich sentences based on those developed for the SCRIBE corpus (see Vonwiller et al. 1995).

2.1 Speakers

The speakers were from two urban centres in New Zealand (Wellington and Hamilton), three age groups (older = 46–60 years, mid-age = 31–45 years, and younger = 18–30 years), and two sexes. The goal in setting up the database was to have six speakers in each of the 12 cells defined by region, age and sex. Due to circumstances to do with recruiting participants for the dialogue tasks for the NZSED corpus, i.e. for reasons not relevant to this paper, the Hamilton mid-aged female group happened to be made up of seven speakers. As there seemed to be no valid basis for choosing one of the speakers for exclusion, this group therefore contains seven speakers, and there is subsequently a total of 73 speakers in the current analysis. Regional differences are not strongly marked in New Zealand English (Gordon & Maclagan 2004), and are not explored in the current analysis.

2.2 Materials

Two sets of materials were included in the analysis. Each set involved the nine vowels discussed above: DRESS, FLEECE, KIT, NURSE, STRUT, START, FOOT, GOOSE, and THOUGHT. The first set (word-list data) consisted of relevant members of the /hVd/ word list (i.e. the above vowels between the consonants /h/ and /d/ in single CVC words), and included one token of each of the vowels of interest, for each speaker. The other set (sentence data) consisted of vowels from words in sentence readings recorded by the same speakers. Eight words were selected for each vowel, all appearing in accented contexts and with the vowel in a syllable carrying primary lexical stress. The token selection resulted in a range of phonetic contexts, i.e. the preceding and following consonants could not be held constant for the different vowels. However, with each set of vowels under comparison, an attempt was made to minimise differences in terms of the immediate phonetic context and syllable structure, as well as the syllable count of the word containing the vowel. Inevitably there was some variation in how speakers phrased the sentences, which affected rhythmic structures and foot length, which has also been shown to affect duration in NZE (Warren 1999). Such differences need to be borne in mind when interpreting the results reported below for vowel duration in the sentence contexts. The identical contexts of the word-list recordings therefore provide a check on durational differences between the vowels in each set. The full list of sentence contexts for the vowels analysed in this study is given in the appendix.

Phoneme-level labelling of all speech files had previously been carried out using the web-based implementation of *MAUS* (Kisler et al. 2016), selecting New Zealand English as the analysis language. Some hand correction of the automatic alignment was necessary. All tokens (from the word list and sentence list) were accessed from the database using *emuR* (Winkelmann et al. 2016).

2.3 Measuring formants and their distributions

Formant trajectories were computed using the *forest* function in the *wrassp* library in R. The *wrassp* library provides functionality for a range of speech signal analysis types, including *for*mant *est*imation (hence *forest*). *Forest* uses linear prediction and the Split-Levinson-Algorithm to derive raw resonance frequencies and bandwidth values. The resonances are

then classified using the Pisarenko frequencies (for more details, see Winkelmann 2016b). Gender-specific parameters were used for female and male speakers (i.e. an analysis window size and nominal F1 value of 12.5 ms and 560 Hz for women and 20 ms and 500 Hz for men). The resulting trajectories for the 5256 sentence tokens and 657 word-list tokens were visually inspected using the *EMU-webApp* (Winkelmann 2016a), and hand-corrected where necessary.

Many analyses of vowels (e.g. Watson & Harrington 1999) have measured formant values at points where F2 reaches a peak (for high and mid front vowels) or a trough (for high and mid back vowels) or where F1 reaches a peak (for open vowels). This approach depends on a reliable classification of vowels as front/back, open/close, etc., which is problematic for the vowel system of NZE, which has been undergoing change over recent generations, so that the 'same' vowel might for example be a back vowel for some speakers but a front vowel for others. An earlier comparison of different methods for measuring vowel formants, including values at single points determined by peaks or troughs in F2 or F1, values at the vowel midpoint, and values averaged over the entire vowel concluded that 'the differences between the various methods used are, in most respects, marginal and all methods used essentially give the same outcome. When studying vowel targets, the method that is most convenient can be used' (van Son & Pols 1990: 1692). The approach taken in the current study was to average the formant values over the central 40% of the vowel's duration. Taking an average, rather than the formant values at a specific point, reduces the impact of local fluctuations in formant values. Taking that average over the central 40% rather than over the entire vowel reduces the impact of in- and out-transitions. Values were calculated for the first three formants, though the majority of the analyses below will present data for F1 and F2 only. F3 values were analysed in the consideration of NURSE-rounding in the comparison of this vowel with KIT. In addition, to measure possible diphthongisation, values for early targets (at 25% of the vowel) and late targets (at 75% of the vowel) were taken for on-glides and off-glides respectively. These targets were considered far enough into the vowel to be uninfluenced by coarticulation from the preceding or following consonants.

To reduce the impact of physiological differences between speakers on comparisons of the vowel sets under consideration, the vowel formant data were normalised. In particular, comparisons of changes over apparent time in the male and female speakers in the dataset are problematic if the impact on formants of anatomical differences in vocal tract dimensions is not taken into account. Following Flynn & Foulkes (2011), a vowel-extrinsic, formant-intrinsic, speaker-intrinsic method was chosen as the most appropriate method for sociophonetic research. The method chosen was that developed by Watt and Fabricius (Watt & Fabricius 2002, Fabricius, Watt & Johnson 2009), and implemented in the *vowels* package (Kendall & Thomas 2015) in R. The method produces estimated centroid values S(Fn), where n = 1 or 2, based on each speaker's vowel space, which is then used to calculate normalised F1 and F2 values (shown as F1/S(F1) and F2/S(F2) in plots below).

The analysis below includes reports of Euclidean distances, either between the midpoints of two vowels or between target points within a vowel as a measure of diphthongisation. These distance measures were calculated for each speaker in the following ways. For vowel comparisons, the Euclidean distance for word-list tokens was based on midpoint formant values for the relevant /hVd/ tokens for each speaker. For the sentence data, average normalised F1 and F2 midpoint values were first calculated across the eight tokens of each vowel for each speaker, and a Euclidean distance computed from these averages. As a measure of diphthongisation, Euclidean distances were calculated between normalised F1 and F2 values for each vowel token's midpoint and the values for early or late targets, for on-glides and off-glides respectively. These Euclidean distances based on normalised F1 and F2 are reported below as distances in S units, using the Watt & Fabricius (2002) label for the estimated vowel centroid. The duration of each vowel was calculated based on transition points resulting from the hand-corrected automatic alignment of the database materials described above.

When the distributions of two vowels are compared using Euclidean distances, basing the statistical analysis on the distance between the mean normalised F1 and F2 values for each speaker reduces the overall distributional cloud for the vowel tokens, but nevertheless allows for an indicative statistical analysis of differences in the formant distributions for our speaker groups. However, Euclidean distance is a measure of distance, and not of overlap. As has been pointed out (Johnson 2015), measures comparing vowel distributions often fail to take into account both distance and overlap, and this can be problematic. For example, one pair of distributions might have the same distance between their centroids as another pair of distributions, but the overlap between the distributions in each pair may be considerably different. In such a case, a simple distance measure will not show that the distributions in each pair differ in their overlap (and confusability). Following Johnson (2015), I report below, for each vowel comparison, values of Bhattacharyya's affinity (Bhattacharyya 1943), computed in R using the *kerneloverlap* function in the *adehabitatHR* library (Calenge 2006). Bhattacharyya's affinity is a statistical measure of the affinity between two populations and is frequently used in spatial ecology to determine overlap in the habitats of two species. It is derived from the densities of two distributions A and B at each (x, y) coordinate (for details of the calculation of Battacharyya's affinity, and a comparison of this measure with others for assessing overlap, see Fieberg & Kochanny 2005). As pointed out by Johnson (2015), Bhattacharyya's affinity is superior to the Pillai trace score (as used by Hay, Warren & Drager 2006). In particular, if two distributions have the same centroids they will be given the same Pillai score even if the distributions differ (e.g. one distribution may be nested within the other, or they may be elliptical with different angles). This is not the case with Bhattacharyya's affinity. However, as an overlap measure, Bhattacharyya's affinity needs to be computed across two reasonably large distributions (30–50 observations, Seaman et al. 1999), which means that it would not be appropriate to calculate it separately for each of our speakers, since each has only eight tokens of each vowel in the sentence materials analysed for this study. As a result, a full statistical analysis of Bhattacharyya's affinity values is not possible. Instead, for each vowel comparison, a simple report of values will be given, derived from the sentence tokens (where there are more examples of each vowel than for the word-list tokens) broken down by each subgroup of speakers (defined by sex and age), as an indicator of the extent of overlap between the two vowels. Possible values for Bhattacharyya's affinity range from zero (no overlap) to one (total overlap).

2.4 Statistical tests

Euclidean distances (between vowels or between vowel midpoints and on- or off-glides) and durations served as dependent variables in linear mixed effects models constructed using the *lme4* library in *R* (Bates et al. 2015). Fixed effects varied depending on the analysis and will be presented with each result, but they included context (sentence vs. word lists), vowel (e.g. FLEECE vs. DRESS), speaker age and speaker sex. The random effects were speakers and, for most analyses, utterances. Utterances could not be included as a random effect when the dependent variable required averaging over utterances, e.g. in analyses that compared Euclidean distances between the centre points of two vowels in the sentence contexts, since these distances were based on the average formant values for each vowel for each speaker across the sentences, as explained above.

In each analysis, simple and interaction effects were assessed by model comparison, using the *mixed* function in *afex* (Singmann, Bolker & Westfall 2015). Significance is assessed in *mixed* using the chi-square distribution of a likelihood ratio test that compares models with and without the effect (or interaction) in question, i.e. they assess whether exclusion of that effect would result in a significantly weaker account for the variation present in the model. Non-significant higher-order interactions were removed, and the resulting best-fit model for each analysis provides the results reported below.

3 Results

Figures 2 and 3 show ellipse plots in normalised F1 and F2 space for the nine vowels in the
/hVd/ word lists and sentence contexts respectively, based on midpoint formant data from the
12 speakers in each cell defined by age and sex, except for mid-aged females, where there
are 13 (because of the extra female recorded in the Hamilton sample). Comparison of the
female and male data for each age group and context shows that the normalisation procedure
outlined above was effective in reducing sex-based differences in vowel space. Comparison
of the ellipse sizes in Figure 3 with those in Figure 2 shows clearly the greater variation in
the vowels' realisations in the sentence contexts than in the /hVd/ context. The following
subsections explore each of the four vowel contrasts in turn, and present statistical analyses
of important aspects of each contrast.

3.1 DRESS–FLEECE

The ellipse plots in Figures 2 and 3 show considerable overlap in normalised F1-F2 space for
DRESS and FLEECE in both types of context. There is also evidence of change-in-progress, i.e.
the DRESS-raising documented by Maclagan & Hay (2004, 2007). That is, the distribution for
DRESS moves up in the vowel space to overlap increasingly with that for FLEECE as we move
from older to mid-age to younger. In addition, the overlap appears to be more advanced for
women in each age group than for men, suggesting that women are leading the change, as
noted also by Maclagan & Hay.

For each speaker, Euclidean distances were calculated between the (average) midpoint
values of their FLEECE and DRESS vowels in normalised F1-F2 space, derived as explained
above. If a speaker's token for DRESS was higher (had a lower F1) than their token for FLEECE,
then the Euclidean distance was expressed as a negative value.[3] Out of the total sample of 73
speakers, this was true of 9 speakers (7 female and 2 male; 7 younger and 2 mid-age) in the
/hVd/ set and of 17 (16 female and one male; 10 younger and 7 mid-age) in the sentence set.

Linear mixed effects models were constructed with Euclidean distance as the dependent
variable, with context (words vs. sentences), sex (female vs. male) and age (older, mid-age,
younger) as fixed effects, and with speaker as a random effect. All interactions of fixed effects
were explored. The best-fit model showed significant simple effects of context ($\chi^2(1) =
33.71$, $p < .0001$), sex ($\chi^2(1) = 33.70$, $p < .0001$) and age ($\chi^2(2) = 21.31$, $p < .0001$). There
were no interaction effects. Unsurprisingly, the distances for word-list data were significantly
greater than those for sentence data (means of 0.188 and 0.103 S units respectively). This is
because articulation is typically more careful in /hVd/ contexts. The effects for sex and age
are shown in Figure 4. The smaller average Euclidean distance for females (0.076 S units)
than for males (0.217 S units) reflects the differences evident in Figures 2 and 3. For age,
the distance between the vowels is smaller for younger speaker groups (0.211 vs. 0.145 vs.
0.079 S units for older, mid-age and younger respectively). The Euclidean distance value for
each age group was significantly different from those of the others. That is, inspection of
the initial model, in which 'older' was the intercept value for age, showed that both mid-age
and younger groups differed from the older group ($t = -2.35$ and -4.84 respectively). When
the analysis was re-run with 'mid-age' as the intercept value, the younger group was also
significantly different from the mid-age group ($t = -2.54$).

Together with the ellipse plots in Figures 2 and 3, this analysis of Euclidean distance by
age between DRESS and FLEECE indicates that these two vowels are becoming increasingly
similar in terms of vowel quality. This is supported in the Bhattacharyya's affinity values for
this pair of vowels in the sentence data, shown in Table 1. In particular, the values are higher
for mid-age and younger speakers than for older speakers and for females than for males,

---

[3] Note that Maclagan & Hay (2007) did not make this adjustment, as they were primarily interested in the
relative closeness of the two vowels to one another (Maclagan, personal communication).
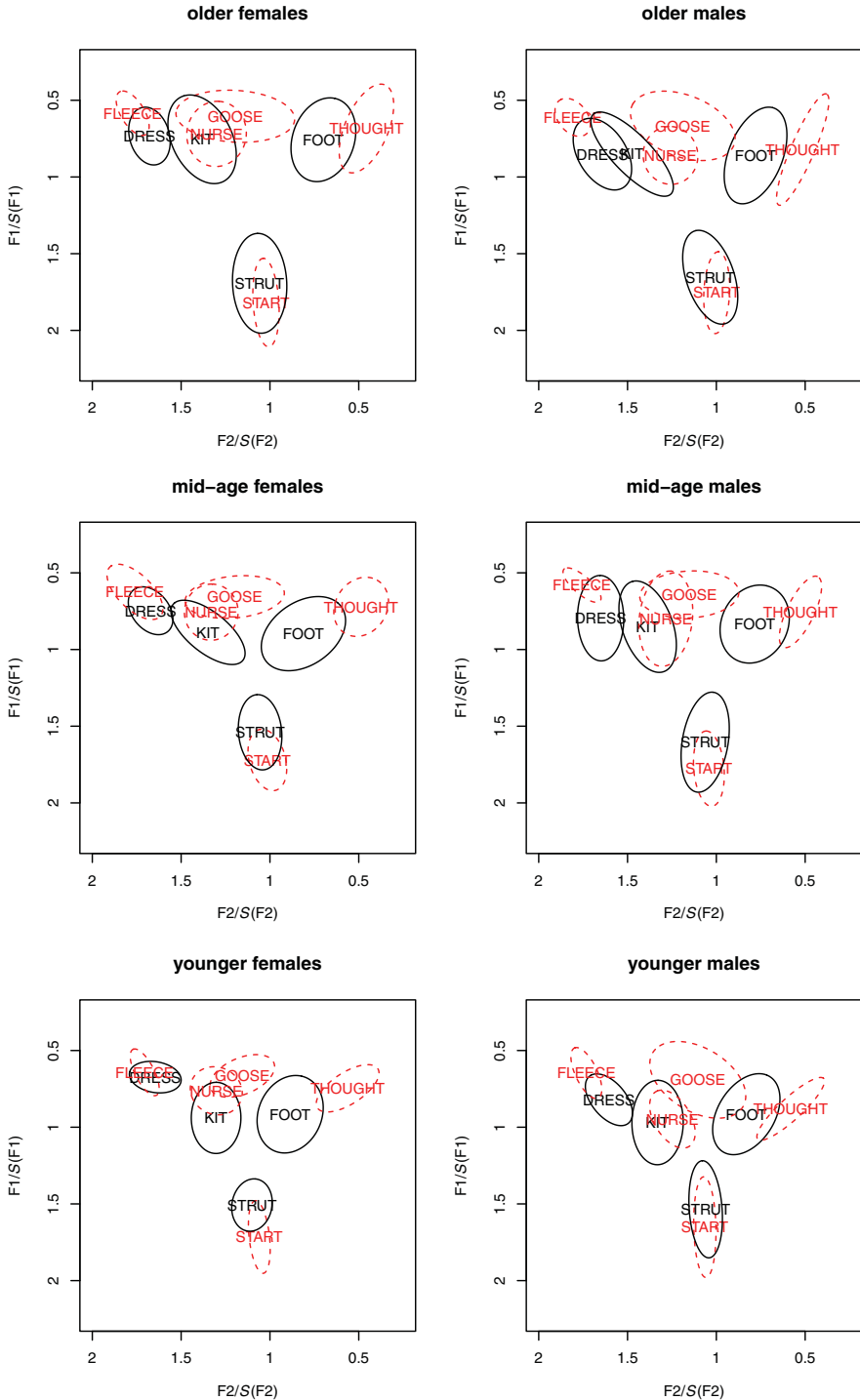
Figure 2 (Colour online) Normalised F1-F2 plots for vowel midpoints for the nine vowels of interest, from /hVd/ contexts, one token per speaker. Left hand panels show female speakers and right hand panels show male speakers. The top row is for older speakers, the middle row for mid-age speakers and the bottom row for younger speakers. Long vowels are shown in red with dotted ellipses.
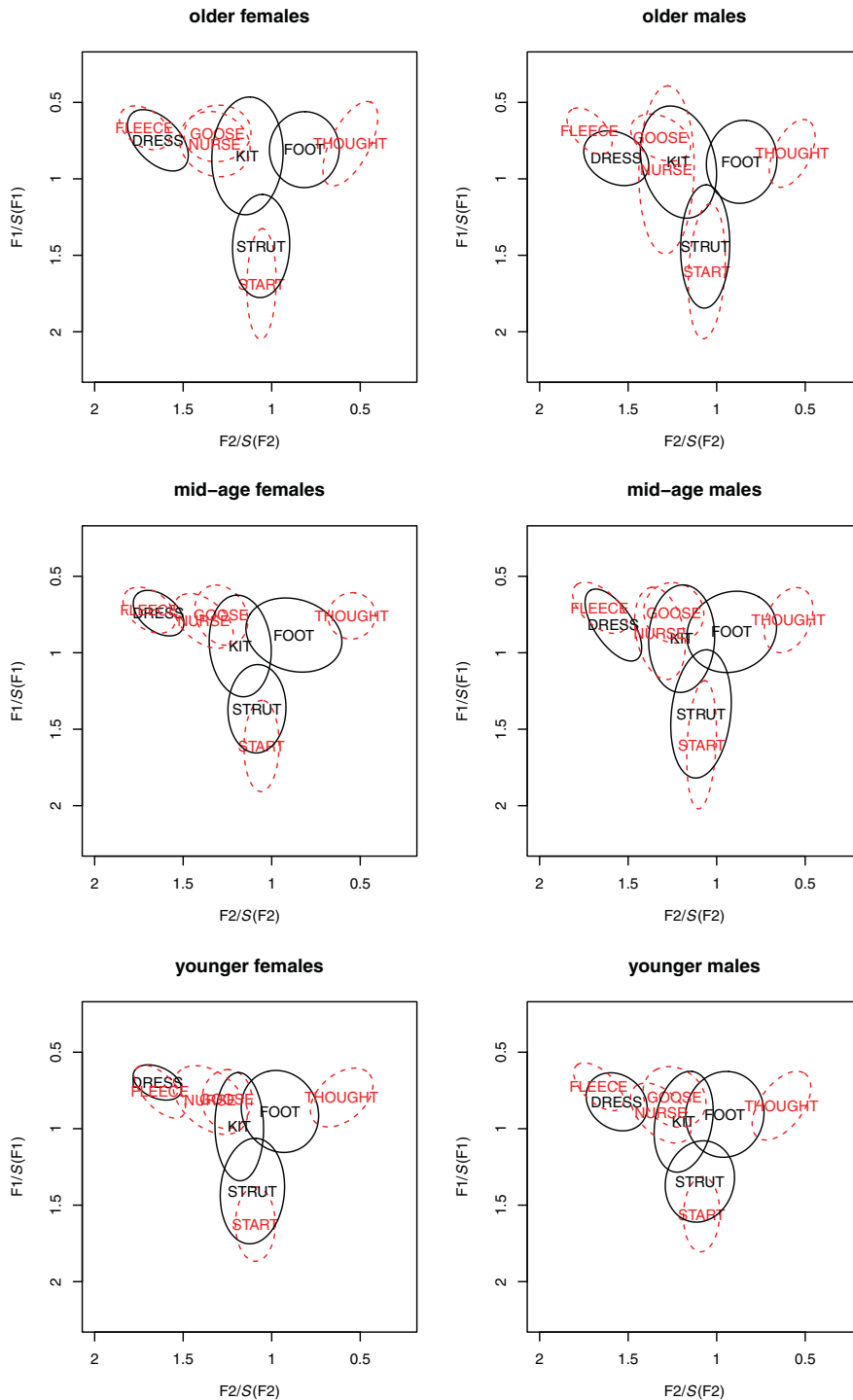
Figure 3 (Colour online) Normalised F1-F2 plots for vowel midpoints for the nine vowels of interest from sentence contexts, for eight accented tokens per vowel for each speaker. Left hand panels show female speakers and right hand panels show male speakers. The top row is for older speakers, the middle row for mid-age speakers and the bottom row for younger speakers. Long vowels are shown in red with dotted ellipses.

Table 1 Bhattacharyya's affinity values for DRESS and FLEECE in the sentence tokens
(possible values range from zero (no overlap) to one (total overlap)).

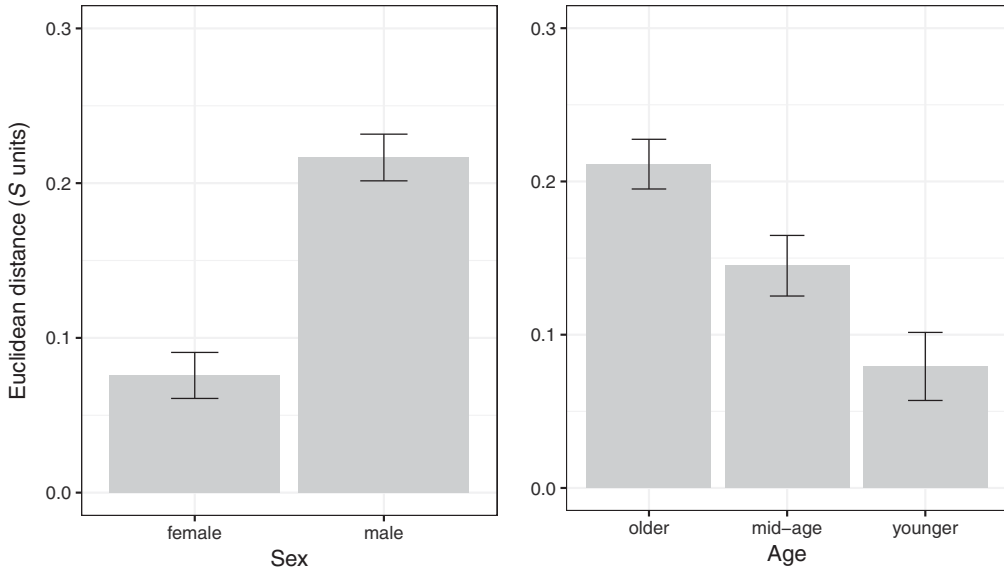| Group	Female	Male
Older	0.862	0.529
Mid-age	0.926	0.839
Younger	0.903	0.738

Figure 4 Euclidean distance between DRESS and FLEECE by speaker sex and age (means and standard errors).

mirroring the results for Euclidean distance in Figure 4. The overlap values are lower for the younger speakers than for the mid-age speakers because some of the DRESS tokens for the younger speakers are higher than their FLEECE tokens, i.e. there is greater separation, but in the opposite direction to that for the older speakers.

As noted in the Introduction, two additional means of distinguishing the vowels are through their duration and through the diphthongisation of the FLEECE vowel (Easton & Bauer 2000; Maclagan & Hay 2004, 2007). Linear mixed effects models for duration as the dependent variable were run with the fixed effects of context (sentences vs. words), vowel (DRESS vs. FLEECE), sex, and age, and with speaker and utterance as random effects. The best-fit model returned significant simple effects of vowel ($\chi^2(1) = 257.63$, $p < .0001$) and age ($\chi^2(2) = 10.20$, $p < .01$) and significant two-way interactions of context with age ($\chi^2(2) = 6.46$, $p < .05$) and with sex ($\chi^2(1) = 6.13$, $p < .05$). FLEECE durations were significantly longer than DRESS durations (165 vs. 116 ms). The age effect reflects shorter vowel durations (averaged across vowels and contexts) for younger speakers (129 ms) than for mid-age (143 ms) and older speakers (149 ms), and is probably indicative of differences in the approach to the recording situation. The interaction of age with context arises because the difference between the contexts was greater for younger speakers (sentences: 120 ms, words: 201 ms) than for mid-age speakers (sentences: 136 ms, words: 202 ms) and older speakers (sentences: 141 ms, words: 213 ms). In particular, vowels in the sentence context are shorter as age drops. The context-by-sex interaction reflects a greater context effect for the females (sentences: 132 ms, words: 210 ms) than for the males (sentences: 133ms, words: 200 ms). The absence in this analysis specifically of interaction effects involving the identity of the vowel together
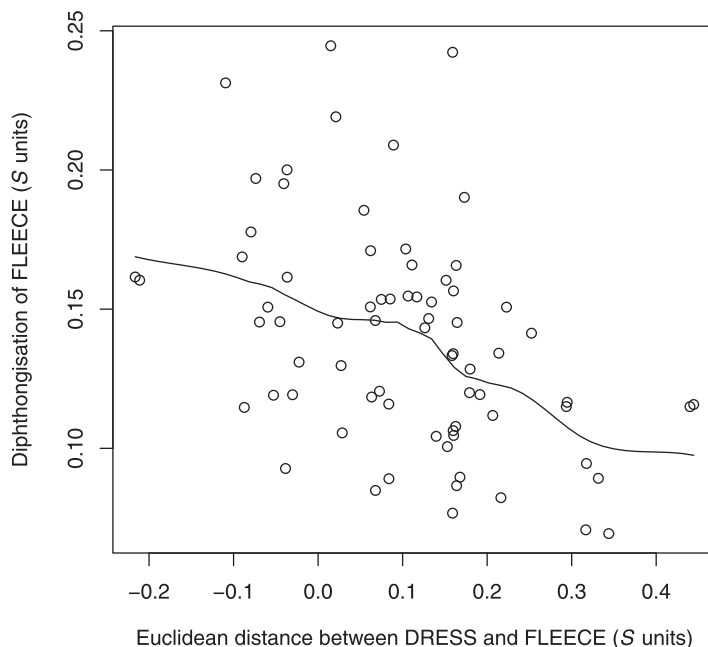
Figure 5 Euclidean distance between DRESS and FLEECE by average FLEECE diphthongisation. Points represent data in the sentence context for individual speakers and the line is a nonparametric smoother fit (loess).

with the other factors shows that the difference in the lengths of FLEECE and DRESS does not vary significantly by speaker sex or age.[4]

Following Maclagan & Hay (2007), the measure of diphthongisation of FLEECE is the Euclidean distance between early and midpoint targets, as defined above. The mean value of this measure across all speakers and contexts was 0.134 S units. Linear mixed effects models considered context, age and sex, together with their interactions, as fixed effects, with speakers and utterances as random effects. The only significant effect in the best-fit model was age ($\chi^2(2) = 9.60$, $p < .01$). The younger and mid-age speakers exhibited significantly more diphthongisation than the older speakers (0.137 and 0.148 S units vs. 0.117 S units; $t = 2.57$ and $t = 3.11$ respectively), but the younger and mid-age speakers did not differ from one another ($t = 1.15$, $p = .25$).

Maclagan & Hay (2007) observe that diphthongisation of FLEECE is more likely as other differences between DRESS and FLEECE reduce in significance (which would include the unreliable difference that they note between DRESS and FLEECE durations). They found a negative correlation between the extent of diphthongisation of FLEECE and the Euclidean distance of FLEECE from DRESS, which is significant when the analysis is restricted to the younger speakers in their sample (20–30 years; they had two age groups, the other speakers being in the range 45–60 years). An analysis of average Euclidean DRESS–FLEECE distances and average FLEECE diphthongisation from the sentence data in the current study (see Figure 5)

---

[4] It should be noted that because the environments for the two vowels are not fully controlled, vowel lengths could be affected by the accentual and rhythmic context in which they are found, and not just by vowel identity. However, the important considerations here are whether such vowel duration differences as exist are affected by speaker age and sex. Clearly they are not. In addition, in the more controlled (though also more limited) context of the /hVd/ word list, the only factor to influence vowel duration is the contrast between FLEECE and DRESS, i.e. the durational contrast appears quite stable across age groups and sex.
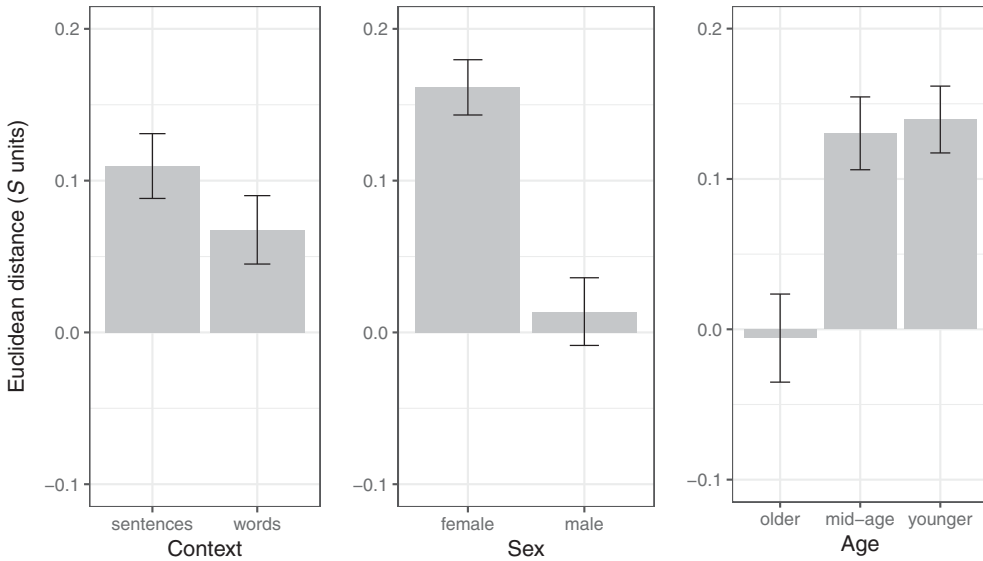
Figure 6 Euclidean distance between KIT and NURSE by context, sex, and age (means and standard errors).

also shows a significant negative correlation, but in this case over all three age groups (Spearman's rho = –0.47, $p < .001$). When the direction of the DRESS–FLEECE Euclidean distance is ignored (i.e. absolute values are used, as was done by Maclagan & Hay, rather than using negative values for cases where DRESS is higher than FLEECE), then the strength of the correlation is lowered, but is still significant (Spearman's rho = –0.37, $p < .01$).

In summary, the current apparent-time analysis confirms Maclagan & Hay's findings of an increasing overlap in vowel space of FLEECE and DRESS, accompanied by increasing diphthongisation of FLEECE.

3.2 KIT–NURSE

The ellipse plots in Figures 2 and 3 above show that the overlap of KIT and NURSE in F1-F2 space varies by speaker group. The trend, though, is for increasing centralisation of KIT over subsequent generations, and this appears to be more advanced for the female speakers (for earlier evidence of this change see also Bauer 1979, 1986; Bell 1997). Accordingly, for the analyses below, the Euclidean distance between KIT and NURSE was assigned a negative value if KIT was closer than NURSE, and as positive value if KIT was more open than NURSE.

Linear mixed effects regression models were built, with average Euclidean distance as the dependent variable, with context, sex and age as fixed effects, and with speakers as random effects. The best-fit model included significant simple effects of context ($\chi^2(1) = 3.75$, $p = .05$), sex ($\chi^2(1) = 20.23$, $p < .0001$) and age ($\chi^2(2) = 16.87$, $p < .001$). These effects are shown in Figure 6.

The vowels are more distinct in the sentence contexts than in the word-list context. This is primarily because the KIT vowel has its more conservative and closer realisation in the word lists. Likewise, Euclidean distances are significantly smaller for the males and for older speakers, i.e. for groups associated with more conservative speech patterns.

Table 2 presents the Bhattacharyya's affinity measures for these two vowels in the sentence data, broken down by age and sex. The values in the table confirm the greater overlap for males than for females, as well as for older speakers.

Table 2 Bhattacharyya's affinity values for KIT and NURSE in the sentence tokens (possible values range from zero (no overlap) to one (total overlap)).

Group	Female	Male
Older	0.710	0.913
Mid-age	0.525	0.806
Younger	0.605	0.724

The data reported above, based on normalised F1 and F2 values, show that KIT and NURSE are more distinct for women and for younger speakers. However, it was noted in the Introduction that NZE NURSE typically has a rounded quality (sometimes in the absence of visible lip-rounding). Because of rounding effects, F2 might be lower in NURSE than would be expected for a given tongue position, meaning that F2 is an underestimate of the degree of fronting, and that NURSE is actually fronter than would appear from Figures 2 and 3 (see also Watson et al. 1998: 194). If there is this rounded quality to NURSE, then we would also expect lower F3 values than would be found for KIT. F3 values for items from the sentence contexts (normalised to Bark values, because the Watt & Fabricius (2002) normalisation procedure outlined above only normalises F1 and F2) were entered into a mixed effects regression model, with vowel (KIT vs. NURSE), sex and age as fixed effects, and speakers and utterances as random effects. The best-fit model showed simple effects of sex ($\chi^2(1) = 99.49$, $p < .0001$), age ($\chi^2(2) = 7.85$, $p < .05$) and vowel ($\chi^2(1) = 3.95$, $p = .05$). Despite the Bark normalisation, females had higher F3 than males (15.16 vs. 14.23). Older speakers had lower F3 than mid-age or younger speakers (14.56 vs. 14.77 and 14.77 respectively). Importantly, F3 values are significantly lower for NURSE (14.55 Bark vs. 14.85 Bark), and there were no interactions of vowel with either sex or age. This absence of interactions shows that the KIT–NURSE difference in F3 is constant across speaker groups. It is probably safe to assume F2 values are also lower because of rounding and that NURSE is therefore fronter than the F2 analysis seems to suggest. The qualities of KIT and NURSE may be more different than indicated by either the ellipse plots or the Bhattacharyya's affinity values.

Durations of KIT and NURSE were compared in mixed effects models that included vowel, context, sex and age as fixed effects and speaker and utterance as random effects. The best-fit model returned significant simple effects of vowel ($\chi^2(1) = 24.41$, $p < .0001$), context ($\chi^2(1) = 24.46$, $p < .0001$) and sex ($\chi^2(1) = 12.11$, $p < .0005$), as well as interactions of vowel with each of context ($\chi^2(1) = 11.16$, $p < .001$), sex ($\chi^2(1) = 30.76$, $p < .0001$) and age ($\chi^2(2) = 43.86$, $p < .0001$). There were no other significant effects. The interactions are plotted in Figure 7. Durations are longer in the word-list context and the difference between KIT and NURSE is also greater in that context, reflecting the more deliberate style of word-list reading. Females have a greater difference between NURSE and KIT than males, while the difference diminishes with decreasing age. The pattern for sex does not reflect any compensatory relationship between the durational contrast and Euclidean distance, since for both measures the females show a greater difference than the males. The age pattern, however, suggests that as Euclidean distance gets larger, so the durational contrast reduces. That is, the younger speakers, for whom the vowels are more distinct in vowel space as KIT centralises, appear to rely less on the durational contrast between the two vowels. To examine this relationship, a correlation test was carried out between the ratios of average NURSE:KIT durations and the Euclidean distance for each speaker in each context. While this test returned a negative correlation (Spearman's rho = –0.15), this was not significant ($p = .07$).

While previous studies of NZE have not noted any diphthongisation of the NURSE vowel, an off-glide has been noted for rural white Southern US English (Thomas 2004), for older non-rhotic as well as rhotic speakers. As a measure of NURSE diphthongisation, Euclidean distances were calculated from the normalised values at the midpoint of the vowel to the later target as described earlier. The mean value of this measure across all speakers and contexts
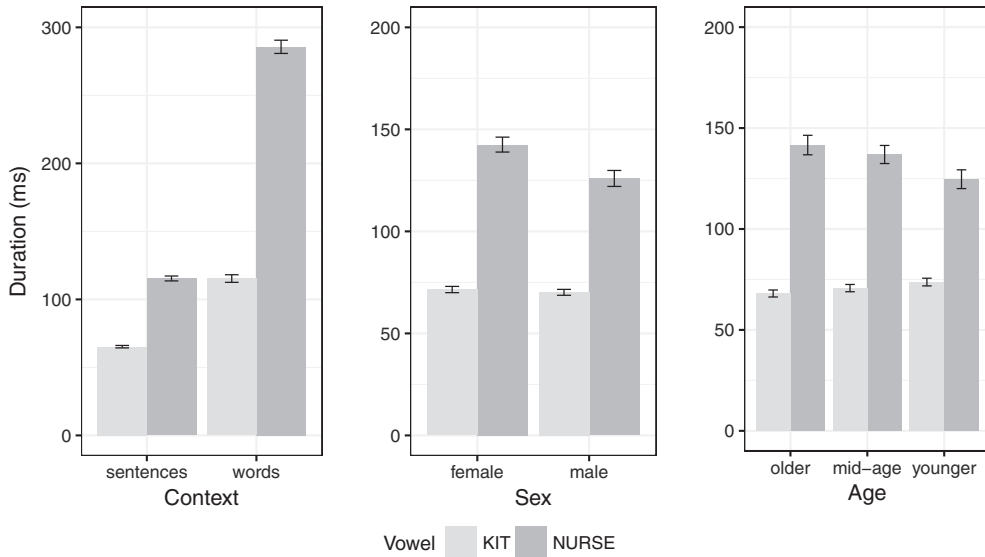
Figure 7 Average durations (in milliseconds) of KIT and NURSE by context, speaker sex and age (means and standard errors).

was 0.061 *S* units, which is considerably lower than the mean on-glide value for FLEECE, reported above, of 0.134 *S* units. Linear mixed effects modelling with context, speaker sex and age as fixed effects and speakers and utterances as random effects returned a best-fit model with the single significant effect of context ($\chi^2(2) = 5.85$, $p < .05$). Diphthongisation was greater in the sentence contexts than in the word-list context, presumably reflecting more careful speech in the word list. Importantly, diphthongisation shows no evidence of change across speaker age or sex.

In summary, while KIT and NURSE show considerable overlap in F1-F2 space for all speaker groups and in both contexts, the two vowels are still distinguished by rounding of NURSE, as well as by vowel duration. Because of NURSE-rounding, it would seem that there is little likelihood that this pair of vowels will become distinguished solely on the basis of quantity.

3.3 STRUT–START

The ellipse plots in Figures 2 and 3 show considerable overlap for the STRUT and START vowels for all age groups and for both female and male speakers. Linear mixed effects regression models for Euclidean distances, with context, speaker sex and age as fixed effects, and speakers as random effects, produced a best-fit model with context as the sole significant effect ($\chi^2(2) = 45.10$, $p < .0001$). STRUT and START were further apart in the sentence contexts than in the word-list contexts. Comparison of Figures 2 and 3 indicates that this is because of greater centralisation of STRUT in the sentence contexts, i.e. undershoot that affects the short vowel more than the long vowel. Indeed, comparisons of Euclidean distances between each speaker's word-list tokens and the average of their sentence tokens show a greater context effect for STRUT (mean distance: 0.203 *S* units) than for START (mean distance: 0.148 *S* units), which was statistically significant (paired-samples *t*-test: $t(72) = 3.24$, $p < .005$). It remains an open question as to whether the perceptual system compensates for undershoot in connected speech contexts to the extent that these vowels overlap perceptually more than they do acoustically (Verbrugge et al. 1976, Verbrugge & Shankweiler 1977).

Table 3 Bhattacharyya's affinity values for STRUT and START in the sentence tokens (possible values range from zero (no overlap) to one (total overlap)).

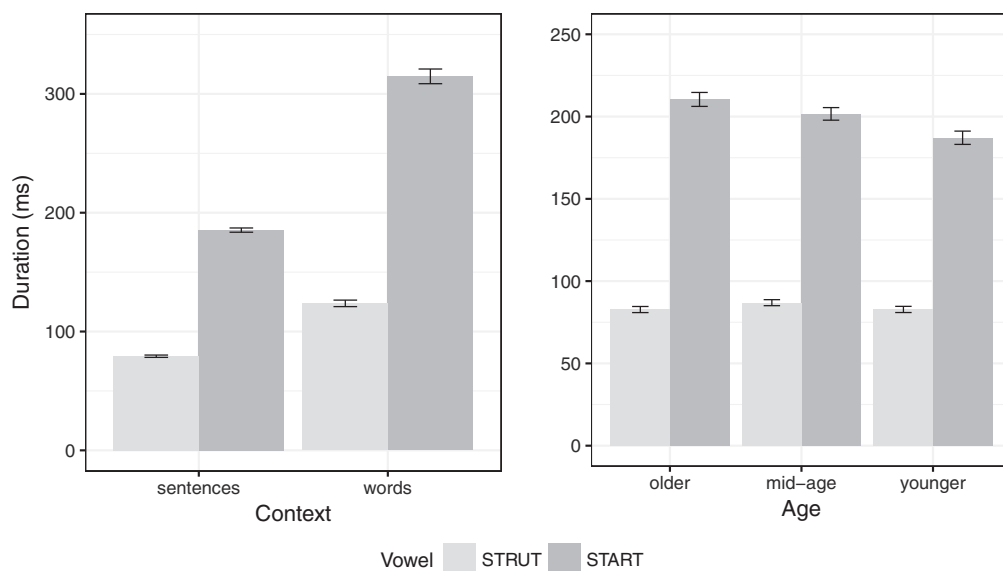| Group	Female	Male
Older	0.808	0.914
Mid-age	0.722	0.847
Younger	0.702	0.696

Figure 8 Average durations (in milliseconds) of STRUT and START by context and age (means and standard errors).

Importantly, the analysis of STRUT–START Euclidean distances showed effects neither of speaker sex nor of age, nor any interactions involving these two factors. However, the Bhattacharyya's affinity values for these vowels in the sentence data (Table 3) exhibit tendencies towards greater overlap for male than for female speakers and for older than younger speakers, despite the absence of any significant effects involving sex or age in the modelling of Euclidean distance reported above. Scrutiny of Figure 3 suggests that this is because of greater centralisation of STRUT for younger speakers and for females.

Linear mixed effects modelling of START and STRUT durations returned a best-fit model with vowel ($\chi^2(1) = 42.39$, $p < .0001$), context ($\chi^2(1) = 25.58$, $p < .0001$) and age ($\chi^2(2) = 10.29$, $p < .01$) as simple effects, along with interactions of vowel with context ($\chi^2(1) = 9.63$, $p < .01$) and with age ($\chi^2(2) = 52.97$, $p < .0001$). Figure 8 shows these interactions. The interaction of vowel with context is similar to that observed for KIT and NURSE (Figure 7), i.e. the durational difference was considerably larger in the word-list context. The simple effect of context reflects the longer overall durations in the word-list context, and that for vowel reflects the longer durations for START than for STRUT. The interaction of vowel with age is due to shorter durations of the START vowel for younger speakers, reflecting perhaps a less formal style of reading, which has a more marked impact on the longer vowel. It is clear though that the durational difference between STRUT and START is maintained across all age groups.

Table 4 Bhattacharyya's affinity values for FOOT and THOUGHT in the sentence tokens (possible values range from zero (no overlap) to one (total overlap)).

Group	Female	Male
Older	0.465	0.357
Mid-age	0.274	0.353
Younger	0.364	0.374

What is interesting is the finding that any apparent-time change in the durational difference between the two vowels is in the opposite direction to what would be predicted if a quantity contrast was replacing a quality contrast. However, note that there was no age effect in the analysis of Euclidean distance, suggesting (as the ellipse plots in Figures 2 and 3 also indicate) that the quality overlap is established already for our older speakers, and that the patterns of durational difference are incidental to the vowel contrast. In addition, a test of the relationship between START:STRUT duration ratios and START–STRUT Euclidean distance showed that these were not correlated (Spearman's rho = 0.030, $p = .72$).

Inspection of the average formant tracks for each vowel in each context, and by speaker sex and age, showed no evidence of diphthongisation. The F1 trajectories were in all cases bowed, with higher values in the central portion of the vowel, but this is to be expected given the need for the tongue to move from a fairly close position for most consonant contexts to the open position for the vowel.

To summarise, STRUT and START show a high degree of overlap in F1-F2 space, a consistently significant durational difference, any variation in which does not appear related to quality differences, and no evidence of diphthongisation. This appears to be a clear quantity pair in NZE.

3.4 FOOT–GOOSE–THOUGHT

In this section, comparisons are made of FOOT first with THOUGHT and then with GOOSE. In the Introduction it was reported that for some speakers FOOT and THOUGHT might constitute a quantity pair (Bauer & Warren 2004: 589). While this may be the case for individual speakers, the ellipse plots in Figures 2 and 3 indicate that it is not generally true across the speaker groups analysed here. Linear mixed effects modelling of FOOT–THOUGHT Euclidean distances with context, sex and age as fixed effects and speakers as a random effect returned a best-fit model in which context ($\chi^2(1) = 93.95, p < .0001$) and age ($\chi^2(2) = 6.56, p < .05$) were significant simple effects, and age interacted significantly with sex ($\chi^2(2) = 7.47, p < .05$). The effect of context was that the word-list recordings had a greater Euclidean distance between FOOT and THOUGHT than the sentence recordings. The interaction effect of age and sex, shown in Figure 9, is that while mid-age and younger female speakers have a greater Euclidean distance between these two vowels than older females, there is no age effect for the male speakers. The women in the sample are clearly also carrying the simple age effect returned in the statistical analysis. The values for Bhattacharyya's affinity (Table 4) show a similar pattern to that found for Euclidean distances.

Euclidean distances between FOOT and GOOSE were tested in modelling with the same structure as that reported for FOOT and THOUGHT. The best-fit model showed significant simple effects of context ($\chi^2(1) = 7.55, p < .01$) and age ($\chi^2(2) = 23.73, p < .0001$). The FOOT–GOOSE distance was greater for word-list recordings than for sentence recordings. The same was reported above for FOOT–THOUGHT, and these two effects reflect a general pattern of a more spread vowel space for the word-list context (see Figures 2 and 3). The age effect results from a closing of the distance between FOOT and GOOSE as we move from older (0.499 *S* units) to mid-age (0.429) to younger speakers (0.345) (all comparisons significant with

Table 5 Bhattacharyya's affinity values for FOOT and GOOSE in the sentence tokens (possible values range from zero (no overlap) to one (total overlap)).

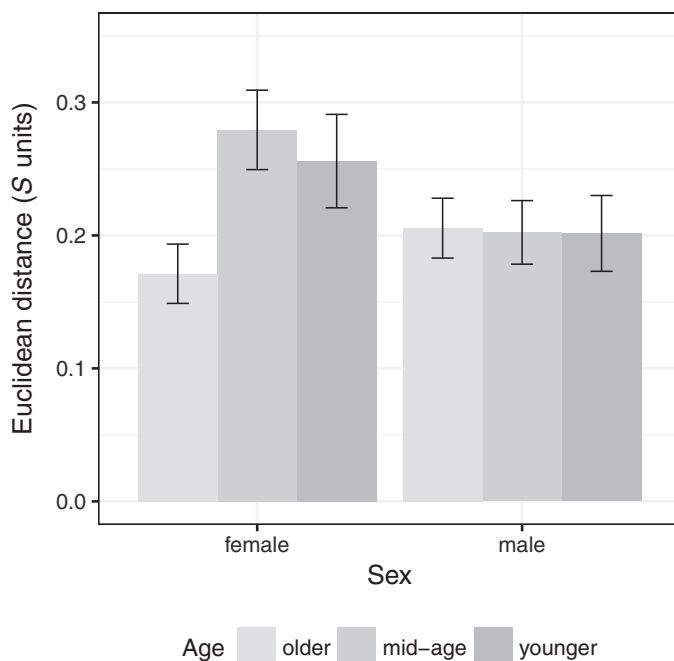| Group	Female	Male
Older	0.086	0.071
Mid-age	0.270	0.303
Younger	0.344	0.449

Figure 9 Euclidean distance between FOOT and THOUGHT by speaker sex and age (means and standard errors).

|t| > 2 and $p < .05$). This is reflected also in the pattern of Bhattacharyya's affinity for this pair in the sentence data (Table 5), which shows greater FOOT–GOOSE overlap with decreasing age. These findings are compatible with the increasing fronting of FOOT reported elsewhere (Kennedy 2004, Warren 2004).

These F1-F2 data for FOOT, THOUGHT and GOOSE show little evidence of overlap of FOOT with either of the longer vowels, while supporting the claim of increasing fronting of FOOT towards GOOSE. This claim is also supported by a Bayesian classification of the FOOT vowel. Using normalised F1 and F2 values, Gaussian models for THOUGHT and GOOSE were built for each speaker group (i.e. age by sex grouping), using the *train* function in *emuR* (emuR 2016), and each FOOT vowel token was then classified (using *emuR's classify* function) as THOUGHT or GOOSE using the model for the relevant speaker group. Logistic mixed effects regression of the classification data, with age and sex as predictors and speaker as a random effect, showed a significant simple effect of age ($\chi^2(2) = 8.04$, $p < .05$) but no other significant effects. The age effect reflects an increasing shift in categorisation of FOOT as GOOSE as we move from older to mid-age to younger speakers.

Linear mixed effects regression modelling of vowel durations, with vowel (FOOT, GOOSE, THOUGHT), context, speaker sex and age as fixed effects and speakers and utterances as random effects returned a best-fit model in which there were significant simple effects for
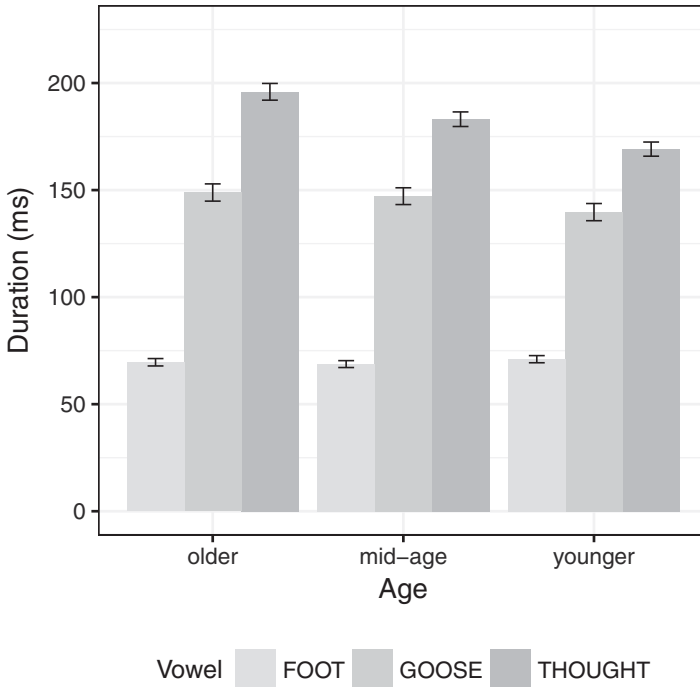
Figure 10 Durations of FOOT, GOOSE and THOUGHT by age (means and standard errors).

vowel ($\chi^2(2) = 156.05$, $p < .0001$) and context ($\chi^2(1) = 29.58$, $p < .0001$), as well as interactions of context with vowel ($\chi^2(2) = 17.69$, $p < .0001$), with sex ($\chi^2(1) = 17.00$, $p < .0001$), and with age ($\chi^2(4) = 73.91$, $p < .0001$), and of vowel with age ($\chi^2(2) = 7.85$, $p < .05$). The overall simple effect of context was that durations were longer in the word-list context than in the sentence context. The interaction of context with vowel arose because the context effect was stronger for the longer vowels (GOOSE and THOUGHT). The interaction of context with sex resulted from a stronger effect of context for the female speakers, and the interaction of context with age arose because the younger speakers showed a stronger effect (largely through greater vowel length reduction in the sentence context) than the mid-age and older speakers. The remaining interaction, of vowel with age, is illustrated in Figure 10. While there is some reduction in the length of both GOOSE and THOUGHT with decreasing age, the differences between FOOT and both of these vowels are consistently present, and there is no clear support for Kennedy's claim that FOOT–THOUGHT duration differences are less evident for younger speakers (Kennedy 2004).

It has been noted (Easton & Bauer 2000, Bauer et al. 2007) that both of the longer vowels in this set are prone to diphthongisation, with an off-glide for THOUGHT and an on-glide for GOOSE. The average Euclidean distance from the midpoint of THOUGHT to the later measurement point (0.75 duration) for that vowel is 0.137 S units, while that from the early measurement point (0.25 duration) to the midpoint of GOOSE is 0.110 S units. Recall that the on-glides for FLEECE averaged 0.134 S units.

Linear mixed effects modelling of THOUGHT off-glides and GOOSE on-glides as dependent variables included context, speaker sex and age as fixed effects and speakers and utterances as random effects. The best-fit model for GOOSE on-glides returned no significant effects at all, that is, GOOSE on-glides appear stable across the contexts and speaker groups examined here. However, the best-fit model for THOUGHT off-glides returned a significant simple effect for age ($\chi^2(2) = 6.11$, $p < .05$), but no other effects. The diphthongisation of THOUGHT increases

as age decreases (older: 0.124; mid-age: 0.134; younger: 0.153; the difference between older and younger was significant with $t = -2.44$, $p < .05$, but mid-age differed significantly from neither of the other groups). Interestingly, average THOUGHT diphthongisation is positively correlated with average Euclidean distance between FOOT and THOUGHT (Spearman's rho $= 0.187$, $p < .05$). That is, diphthongisation of THOUGHT is clearly not a result of any increasing overlap in vowel space with FOOT, since it increases as THOUGHT and FOOT become more separate.

4 Discussion

The acoustic analysis above of four sets of vowel comparisons in word-list and sentence materials sampled from NZSED has investigated the possibility that a number of vowel contrasts in NZE are moving towards a quantity distinction. Claims for such a development were earlier made (e.g. Bauer & Warren 2004, and other authors discussed in the Introduction above) for DRESS and FLEECE, KIT and NURSE, STRUT and START, and FOOT and THOUGHT. One of the advantages afforded by the use of the NZSED recordings is that we can consider all of these vowel contrasts for the same set of speakers, including comparisons of female and male speakers and with age-grading showing possible changes in apparent time.

The Euclidean distance measures (normalised) for the vowel comparisons showed two significant effects for speaker sex. Females had a smaller distance between DRESS and FLEECE than males, but a greater distance between KIT and NURSE. Although these results appear to point in contradictory directions, they both reflect more advanced innovation in the female speech. That is, the DRESS–FLEECE difference results from the more advanced raising of DRESS into FLEECE-space for females, and the KIT–NURSE difference arises because the centralisation of KIT, a widely reported feature of NZE, is more advanced for females. The same explanation can be put forward for the contrasting trends in age effects on Euclidean distances – DRESS and FLEECE are significantly closer for younger than for mid-age and for mid-age than for older speakers (Figure 4), while KIT and NURSE are further apart for younger and mid-age than for older speakers (Figure 6). That is, younger speakers are more advanced in terms of DRESS raising and KIT centralisation than older speakers.

In neither of these vowel contrasts is there strong evidence that a difference in Euclidean distances is compensated for by a difference in durational contrasts. Importantly, the analysis of DRESS and FLEECE revealed that the durational patterning of the vowels did not change significantly in response to any other factors (as shown by the absence of interactions of vowel identity with the other factors in the analysis of duration). While there was a suggestion that NURSE:KIT duration ratios decrease as Euclidean distance increases when we move from older to mid-age and then to younger speakers, these measures were not significantly correlated. What is more, such apparent-time changes in the NURSE:KIT duration ratio would indicate a move away from a quantity contrast, with younger speakers having smaller durational differences, rather than towards one. Other aspects of the analyses did suggest some alternative compensatory measures. In particular, the analysis of DRESS and FLEECE returned a significant correlation of the overlap of these vowels in vowel space with the diphthongisation of FLEECE.

As noted in the Introduction, the long-short distinction between members of NZE vowel pairs has often been conflated with the tense-lax distinction (e.g. by Bauer & Warren 2004). This would appear to place the NZE vowel system closer to a British English RP system, where vowel distinctions have traditionally been indicated in terms of both quality/peripherality and length (e.g. transcriptions that distinguish between /iː/ and /ɪ/), than to a General American system, where vowel length is not generally considered to be distinctive and is therefore not marked in most transcription systems (thus /i/ and /ɪ/). In phonological theory, the usefulness of the tense-lax distinction in the description of English has been the subject of much debate (see summaries in e.g. Durand 2005, Kwon 2011). While the distinction may remain useful for descriptions of phonotactic behaviour (e.g. to distinguish vowels that may occur in open syllables from those that may not), it is not clear that the correlation of tense-lax with long-

short is helpful in the description of the NZE vowel system. For instance, Labov (1994: 285), in his description of the short front vowel shift in NZE, argues that the short front vowels in NZE should in fact be treated as tense. These NZE short front vowels have undergone a chain shift in which TRAP and DRESS have raised, and KIT has retracted (Bauer 1979, 1986; Bell 1997; Trudgill, Gordon & Lewis 1998; Watson & Harrington 1999; Gordon et al. 2004; Langstrof 2006). As Labov (1994: 138) notes, this shift provides an exception to two of his general principles governing chain-shifts in vowels, since in such shifts short (i.e. lax) front vowels are expected to fall rather than rise (Principle II), and back vowels are expected to move to the front rather than vice versa (Principle III). If the NZE short front vowels are accordingly regarded as tense, then the tense-lax distinction for front vowels in NZE collapses.

As Maclagan & Hay point out (2007: 20), a treatment of the NZE short front vowels as tense may also help to account for the linkage between the movement of DRESS and the realisation of FLEECE, since this treatment would place the two in the same subsystem of vowels. From their analysis of two age groups (20–30 years and 45–60 years), Maclagan & Hay (2007: 12) concluded that FLEECE is diphthongising in response to the raising of DRESS, by moving out of the monophthongal system via some process such as Labov's Upper Exit Principle (Labov 1994: 281). Note though that while the analysis of NZED data presented above revealed significant differences between each age group in the overlap of DRESS and FLEECE, the younger and mid-age groups did not differ from one another in the extent of FLEECE diphthongisation, yet both had greater FLEECE diphthongisation than the older group. If the diphthongisation of FLEECE is equally strong for the younger and mid-age groups, but the overlap of DRESS and FLEECE distinguishes between these two groups, then this suggests that the diphthongisation of FLEECE might be occurring in advance of the overlap of the two vowels. So the current data suggest that FLEECE might in fact be making way for DRESS-raising, rather than responding to it. However, regardless of whether FLEECE-diphthongisation is a precursor to or consequence of DRESS-raising, both Maclagan & Hay (2007) and the current analysis place the two vowels in the same subsystem.

A similar breakdown of the traditional distinction of subsets on the basis of tension or length is suggested by Gordon et al. (2004: 206–207) in a tentative proposal that sees movement of START as the catalyst, historically, for the raising of the short front vowels in NZE. The suggestion is that as START fronted, TRAP raised, leading to a chain reaction that also involved DRESS and KIT. Given the current overlap of START and STRUT, and the relatively open and front position of STRUT in data that Gordon et al. present for speakers born between 1870 and 1889 (their Figure 6.5), it is possible that STRUT has also been involved in this chain shift. START and STRUT present a pair for which overlap in vowel quality has long been noted in NZE, and the only pair for which a revision of the transcription system (Bauer et al. 2007) suggested the same basic symbols, with the addition of the length mark for START. In the data above it is the clearest case where quantity is the sole point of distinction, and for which there is also no indication of diphthongisation for either vowel. Interestingly, the analysis above suggests that the quantity distinction may be reducing. However, since START is still at least twice the length of STRUT for even the youngest speakers, it would be premature to predict a complete merger of these two vowels. In addition, this reduction in the quantity distinction is not found in the word-list data but only in the sentence recordings, where there are lexical and contextual cues to the identity of the words containing these vowels. Other research has been exploring the consequences of the vowel changes discussed in this paper for lexical access (Warren 2006).

Finally, the comparison of FOOT with THOUGHT and GOOSE supports earlier observations that FOOT is fronting and becoming more similar to GOOSE than to THOUGHT. Thus the Euclidean distance between FOOT and THOUGHT increases and that between FOOT and GOOSE decreases as we move from older to younger speakers. In neither case does there appear to have been any compensatory adjustment of vowel durations or diphthongisation. FOOT remains consistently shorter than either THOUGHT or GOOSE. The fact that diphthongisation of THOUGHT is positively correlated with FOOT–THOUGHT distance indicates that the relationship

between these two measures is not compensatory. At the same time, the pattern for GOOSE does not indicate that it is diphthongising any more strongly as FOOT approaches its vowel space.

In summary, although a shift to a length-based contrast has been suggested for multiple vowel oppositions in NZE, this is only clearly supported in the current analysis for the STRUT–START pair. For the other sets under consideration, there has either been a further move of one of the vowels out of the space of the other (i.e. a continuing process of centralisation sees KIT moving away from the territory of NURSE, and fronting has moved FOOT away from THOUGHT), or diphthongisation appears to be moving the longer vowel out of the monophthongal subset (as is most clearly the case with FLEECE, although diphthongal realisations of THOUGHT and GOOSE are also prevalent).

Acknowledgments

Appendix. Tokens taken from sentence recordings

The word containing the target vowel is in bold and the target vowel itself is underlined. Note that the NZ English of the samples is non-rhotic.

DRESS

S001: The price range is smaller than any of us **exp<u>e</u>cted**
S023: I'll hedge my **b<u>e</u>ts** and take no risks
S038: You ought to brush your teeth before you go to **b<u>e</u>d**
S084: The coach swerved to the left in an attempt to avoid the blind **ped<u>e</u>strian**
S092: Compared with the previous **gu<u>e</u>sts** she appeared coy and demure
S169: Cath opened the larder door to fetch the iceberg lettuce and a chunk of **ch<u>e</u>ddar**
S189: I can't pretend to know the exact answer but I can make an educated **gu<u>e</u>ss**
S194: It's obvious that the student was amply rewarded for his **end<u>e</u>avours**

FLEECE

S038: You ought to brush your **t<u>ee</u>th** before you go to bed
S059: **P<u>e</u>te** couldn't bear to show us his scar so soon after the accident
S072: Everyone talks of the birds and the **b<u>ee</u>s** but they never mention wasps
S129: The yacht capsized in the choppy **s<u>ea</u>s** but all the crew swam to safety
S170: The babe was starting to **t<u>ee</u>the** but her favourite toy appeared to soothe her
S173: The topic of Jeff's **th<u>e</u>sis** is beginning to annoy me
S179: You always manage to **bes<u>ie</u>ge** me with other people's blasted problems
S187: I'll never know whether it was the alcohol or the lack of **sl<u>ee</u>p** that gave me blurred vision

KIT

S010: It is futile to offer any further **res<u>i</u>stance**
S023: I'll hedge my bets and take no **r<u>i</u>sks**
S037: The bath plug is **m<u>i</u>ssing** so you'll have to take a shower
S103: The plaintiff felt thwarted because the judge didn't revoke his **dec<u>i</u>sion**
S135: Some pop stars just mouth the words because they can't **s<u>i</u>ng**
S140: He clenched his **f<u>i</u>st** and hammered it against the wall with a resounding smack of authority
S187: I'll never know whether it was the alcohol or the lack of sleep that gave me blurred **v<u>i</u>sion**

S188: I'm intrigued that social values can change so **qu<u>i</u>ckly**

NURSE

S006: John could lend him the latest draft of his **w<u>or</u>k**
S017: It was important to be **p<u>er</u>fect** since there were no prompts
S053: There'll be big trouble if you dare to touch that **s<u>ur</u>face**
S056: Doctor Philips raised a number of moot points about the Professor's abridged article in the recent **j<u>our</u>nal**
S113: Mary and Elizabeth both aim to be company directors by the age of **th<u>ir</u>ty**
S125: Dad wanted us to leave the surfboard in the boat until our **ret<u>ur</u>n**
S162: The Cuthbert family live by the village **ch<u>ur</u>ch**
S197: The pretty manageress treated all her shop assistants like **d<u>ir</u>t**

STRUT

S011: They launched into battle with all the forces they could **m<u>u</u>ster**
S035: It's a shame that architects design for themselves and not for the general **p<u>u</u>blic**
S070: They were still frisked by **C<u>u</u>stoms** though they had nothing to declare
S103: The plaintiff felt thwarted because the **j<u>u</u>dge** didn't revoke his decision
S111: Those texts you mention were **disc<u>u</u>ssed** in depth by the brethren last Sunday
S133: I'm obliged to tell you that most women loathe their **h<u>u</u>sbands**
S142: You should eat less of that **f<u>u</u>dge** cake if it really is aggravating your sore tooth
S162: The **C<u>u</u>thbert** family live by the village church

START

S002: They asked if I wanted to come along on the **b<u>ar</u>ge** trip
S054: Tom says that ancient **S<u>aa</u>bs** are far more stylish than British Leyland Triumphs
S068: He remembered he needed a **p<u>a</u>ssport** to get a visa stamp for France
S090: We were plunged into **d<u>ar</u>kness** as the clouds engulfed the moon
S101: We'll need to put the poodle back on the leash when we leave the **p<u>ar</u>k**
S151: He caught a glimpse of what looked like a badger in the **m<u>ar</u>sh**
S168: I feel too **leth<u>ar</u>gic** to wash up tonight
S177: The ivy wound its way grotesquely round the neck of the **g<u>ar</u>goyle**

FOOT

S021: Thank **g<u>oo</u>dness** it's Friday and time to go home
S031: It's difficult to choose between two such equally **g<u>oo</u>d** alternatives
S064: We'd be hard **p<u>u</u>shed** to catch the bus to Newcastle tonight
S102: The Olympic torch shines as a symbol of hope which has **p<u>u</u>shed** aside the barriers of race
S115: Cobwebs slowly find their way into every **n<u>oo</u>k** and cranny
S136: The toddler stamped his **f<u>oo</u>t** in annoyance and gouged a large hole in his parents' table out of spite
S184: The Websters are going to see "**P<u>u</u>ss** in Boots" just after Christmas
S186: I'm well aware that raw vegetables contain more vitamins than **c<u>oo</u>ked** ones

GOOSE

S028: The mud squelched loudly and he realised that his suede **b<u>oo</u>ts** were doomed
S079: The **f<u>oo</u>d** varies from place to place but the price remains fairly constant
S083: We really will need to defrost the fridge now that the strawberries have **<u>oo</u>zed** out all over the place
S142: You should eat less of that fudge cake if it really is aggravating your sore **t<u>oo</u>th**
S146: This is the eighth week that parsnip **s<u>ou</u>p** has been served at this hotel

S181: Cinderella waltzed through the night while poor Buttons **mooched** aimlessly about the house

S184: The Websters are going to see "Puss in **Boots**" just after Christmas

S198: I wish I lived a little bit nearer to the **supermarket**

THOUGHT

S019: We have proof that the regime wields sufficient power in the **North** to exploit the entire population

S034: When **forced** to make a choice, Sarah chose ping-pong as her favourite game

S089: The questionnaire about "King Lear" was **short** and to the point

S097: It seems as if Susan does all the **chores** for this household

S098: I want each person here to give **George** their report by Thursday afternoon

S102: The Olympic **torch** shines as a symbol of hope which has pushed aside the barriers of race

S155: The archbishop has resolved not to disclose the details of the bitter **divorce**

S175: Her wealth's certain to buy her not only a **Porsche** but also considerable popularity

References

Baskaran, Loga. 2004. Singapore English: Phonology. In Kortmann et al. (eds.), 1034–1046.

Bates, Douglas, Martin Maechler, Ben Bolker & Steven Walker. 2015. *lme4* (Version 1.1-8). http://CRAN.R-project.org/package=lme4 (retrieved 22 February 2016).

Bauer, Laurie. 1979. The second great vowel shift. *Journal of the International Phonetic Association* 9(2), 57–66.

Bauer, Laurie. 1986. Notes on New Zealand English phonetics and phonology. *English World-Wide* 7, 225–258.

Bauer, Laurie & Paul Warren. 2004. New Zealand English: Phonology. In Kortmann et al. (eds.), 580–602.

Bauer, Laurie, Paul Warren, Dianne Bardsley, Marianna Kennedy & George Major. 2007. Illustrations of the IPA: New Zealand English. *Journal of the International Phonetic Association* 37(1), 97–102.

Bell, Allan. 1997. The phonetics of fish and chips in New Zealand: Marking national and ethnic identities. *English World-Wide* 18, 243–270.

Bernard, J. R. 1967. Length and the identification of Australian English vowels. *AUMLA (Journal of the Australian Universities Language and Literature Association)* 27, 37–58.

Bhattacharyya, Anil Kumar. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* 35, 99–109.

Calenge, Clément. 2006. The package "adehabitat" for the R software: A tool for the analysis of space and habitat use by animals. *Ecological Modelling* 197, 516–519.

Clark, John. 1989. Some proposals for a revised phonetic transcription of Australian English. In Peter Collins & David Blair (eds.), *Australian English: The language of a new society*, 205–213. Brisbane: University of Queensland Press.

Durand, Jacques. 2005. Tense/lax, the vowel system of English and phonological theory. In Philip Carr, Jacques Durand & Colin Ewen (eds.), *Headhood, elements, specification and contrastivity*, 77–97. Amsterdam: John Benjamins.

Durie, Mark & John Hajek. 1995. Getting it right: More on an orthography for Australian English. *Australian Journal of Linguistics* 15(2), 227–240.

Easton, Anita & Laurie Bauer. 2000. An acoustic study of the vowels of New Zealand English. *Australian Journal of Linguistics* 20(2), 93–117.

Fabricius, Anne, Dominic Watt & Daniel Ezra Johnson. 2009. A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change* 21, 413–435.

Fieberg, John & Christopher O. Kochanny. 2005. Quantifying home-range overlap: The importance of the utilization distribution. *The Journal of Wildlife Management* 69(4), 1346–1359.

Flynn, Nicholas & Paul Foulkes. 2011. Comparing vowel formant normalization methods. In Wai-Sum Lee & Eric Zee (eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (ICPhS XVII), 683–686. Hong Kong: City University of Hong Kong.

Gordon, Elizabeth, Lyle Campbell, Jennifer Hay, Margaret Maclagan, Andrea Sudbury & Peter Trudgill. 2004. *New Zealand English: Its origins and evolution*. Cambridge: Cambridge University Press.

Gordon, Elizabeth & Margaret Maclagan. 2004. Regional and social differences in New Zealand: Phonology. In Kortmann et al. (eds.), 603–613.

Hay, Jen, Paul Warren & Katie Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34(4), 458–484.

Johnson, Daniel Ezra. 2015. Quantifying vowel overlap with Bhattacharyya's affinity. Presented at NWAV44. https://danielezrajohnson.shinyapps.io/nwav_44/ (retrieved 4 September 2017).

Kendall, Tyler & Erik R. Thomas. 2015. vowels (version 1.2-1). https://cran.r-project.org/web/packages/vowels/ (retrieved 2 September 2016).

Kennedy, Marianna. 2004. Vowel movement, vowel duration and regional variation: Another adventure in the vowel system of New Zealand English. BA Honours project, Victoria University of Wellington.

Kisler, Thomas, Uwe D. Reichel, Florian Schiel, Christoph Draxler, Bernhard Jackl & Nina Pörner. 2016. BAS Speech Science Web Services: An update of current developments. Presented at 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.

Kortmann, Bernd, Edgar W. Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.). 2004. *A handbook of varieties of English: A multimedia reference tool*, vol. 1. Berlin: Mouton de Gruyter.

Kwon, Young-Kook. 2011. Doing without [tense/lax] in English. *Linguistic Research* 28(3), 605–624.

Labov, William. 1994. *Principles of language change*, vol. 1: *Internal factors*. Oxford: Blackwell.

Langstrof, Christian. 2006. Acoustic evidence for a push-chain shift in the Intermediate Period of New Zealand English. *Language Variation and Change* 18, 141–164.

Maclagan, Margaret. 1982. An acoustic study of New Zealand vowels. *The New Zealand Speech Therapists Journal* 37(1), 20–26.

Maclagan, Margaret & Jen Hay. 2004. The rise and rise of New Zealand English DRESS. In Steve Cassidy, Felicity Cox, Robert Mannell & Sallyanne Palethorpe (eds.), *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, 183–188. Canberra: Australian Speech Science and Technology Association Inc.

Maclagan, Margaret & Jen Hay. 2007. Getting fed up with our feet: Contrast maintenance and the New Zealand English "short" front vowel shift. *Language Variation and Change* 19, 1–25.

Millar, J. Bruce, Philip Dermody, Jonathan Harrington & Julie Vonwiller. 1990. *A national database of spoken language: Concept, design, and implementation*. Presented at International Conference on Spoken Language Processing (ICSLP-90), Kobe, Japan.

Seaman, D. E., J. J. Millspaugh, Brian J. Kernohan, Gary C. Brundige, Kenneth J. Raedeke & Robert A. Gitzen. 1999. Effects of sample size on KERNEL home range estimates. *Journal of Wildlife Management* 63(2), 739–747.

Singmann, Henrik, Ben Bolker & Jake Westfall. 2015. afex: Analysis of Factorial Experiments (Version 0.13-145). https://cran.rproject.org/web/packages/afex/index.html (retrieved 9 December 2016).

Tent, Jan & France Mugler. 2004. Fiji English: Phonology. In Kortmann et al. (eds.), 750–779.

Thomas, Erik R. 2004. Rural Southern white accents. In Kortmann et al. (eds.), 300–324.

Trudgill, Peter, Elizabeth Gordon & G. Lewis. 1998. New dialect formation and southern hemisphere English: The New Zealand short front vowels. *Journal of Sociolinguistics* 2, 35–51.

van Son, R. J. J. H. & Louis C. W. Pols. 1990. Formant frequencies of Dutch vowels in a text, read at normal and fast rate. *The Journal of the Acoustical Society of America* 88(4), 1683–1693.

Verbrugge, R. R. & D. P. Shankweiler. 1977. Prosodic information for vowel identity. *Haskins Laboratories Status Report on Speech Research* SR-51/52, 27–35.

Verbrugge, R. R., W. Strange, D. P. Shankweiler & T. R. Edman. 1976. What information enables a listener to map a talker's vowel space? *The Journal of the Acoustical Society of America* 60, 198–212.

Vonwiller, Julie, Inge Rogers, Chris Cleirigh & Wendy Lewis. 1995. Speaker and material selection for the Australian national database of spoken language. *Journal of Quantitative Linguistics* 2(3), 177–211.

Warren, Paul. 1999. Timing properties of New Zealand English rhythm. In John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville & Ashlee C. Bailey (eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (ICPhS XIV), 1843–1848. San Francisco, CA: Regents of the University of California, Berkeley.

Warren, Paul. 2002. NZSED: Building and using a speech database for New Zealand English. *New Zealand English Journal* 16, 53–58.

Warren, Paul. 2004. On the front foot: The quality and quantity of ANZE vowels. PANZE workshop keynote presentation, The Tenth Australian International Conference on Speech Science & Technology (SST 2004), Sydney.

Warren, Paul. 2006. Oops, I've done a futt: Quality and quantity in a New Zealand vowel contrast. *Te Reo* 49, 125–143.

Watson, Catherine I. & Jonathan Harrington. 1999. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America* 106(1), 458–468.

Watson, Catherine I., Jonathan Harrington & Zoe Evans. 1998. An acoustic comparison between New Zealand and Australian English vowels. *Australian Journal of Linguistics* 18(2), 185–207.

Watt, Dominic & Anne Fabricius. 2002. Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1-F2 plane. *Leeds Working Papers in Linguistics and Phonetics* 9, 159–173.

Wee, Lionel. 2004. Singapore English: Phonology. In Kortmann et al. (eds.), 1017–1033.

Wells, John C. 1982. *Accents of English*. Cambridge: Cambridge University Press.

Winkelmann, Raphael. 2016a. EMU-webApp (version 0.0.34). http://ips-lmu.github.io/EMU-webApp/ (retrieved 2 September 2016).

Winkelmann, Raphael. 2016b. wrassp (version 0.1.4). https://github.com/IPS-LMU/wrassp (retrieved2 September 2016).

Winkelmann, Raphael, Klaus Jaensch, Steve Cassidy & Jonathan Harrington. 2016. emuR: The R library for the Emu Speech Database System (version 0.1.9). https://github.com/IPS-LMU/emuR (retrieved 2 September 2016).