Disambiguation of Australian English vowels

Tunde Szalay, Titia Benders, Felicity Cox, Michael Proctor

Department of Linguistics, Macquarie University, Sydney, Australia

{tunde.szalay, titia.benders, felicity.cox, michael.proctor}@mq.edu.au

Abstract

Discriminability of Australian English vowel pairs was examined using a lexical decision task. 15 native listeners categorized 10,800 Australian English words with /hVd/ structure in a binary forced choice task. 16 target words differing only in the nuclear vowel were presented as auditory stimuli, paired with each of the remaining 15 words as competitors. Accuracy and reaction time were measured. Results show that target vowel and the phonetic similarity of target vowel and competitor vowel affect lexical decision.

Index Terms: perception, lexical decision, Australian English, vowel discriminability

1. Introduction

A large body of research has examined word recognition, and the factors that are involved. Lexical frequency [1], familiarity [2], morpho-syntactic context [3, 4], and neighbourhood effects – both semantic [5] and phonological [6, 7, 8] – have been shown to influence the speed and accuracy with which an auditory lexical stimulus is classified. Many other factors are also involved in word recognition in non-auditory domains [9, 10].

Lexical access has been shown to be sensitive to within-category gradient variation in phonetic factors including VOT [11, etc.]. Listeners make different decisions according to the amount of information that is available [12, 13, 14], and these studies showed how and when phonemes are identified and disambiguated from each other.

While much is known about mechanisms of lexical disambiguation in American English, fewer studies have examined the details of these processes in Australian English (AusE). Most of this work has focused on perception at the level of the word and the segment [15]. AusE presents special challenges to the listener because it uses a large vowel inventory containing 18 stressed vowels (plus schwa) and incorporating phonemic vowel length contrast for certain spectrally similar pairs. For example, pairs of vowels such as /ɐ-ɐː, e-eː/, and /ɪ-iː/ have similar spectral quality but contrast in length. AusE also has monophthong-diphthong pairs such as /æ-æɪ, ɑe-ɔ/ and /æ-æɔ/ in which the monophthong is related to the first element of the diphthong, and /ʉː-əʉ, ɔ-æɔ/, in which the monophthong is related to the second element of the diphthong [16].

Such a vowel space makes intrinsic vowel similarities and vowel discriminability important.To better understand how listeners of AusE discriminate vowels in a rich and temporally-differentiated vowel space, we used a perception experiment to examine confusability. The aim of this experiment was to discover which vowels and vowel pairs are intrinsically hard to disambiguate.

2. Method

Vowel disambiguation was examined using a binary forced-choice lexical decision task [18]. Participants were presented with an auditory stimulus, and asked to identify the word by selecting one of two candidates presented orthographically on a computer screen. Participants were instructed to select the word they heard as quickly as they could.

2.1. Participants

15 native speakers of Australian English (14 female; ages 19 to 47; mean 22.5 years) took part in the experiment. All were undergraduate students of linguistics at Macquarie University who received class credit for participation. Fourteen participants were born in Australia and one immigrated before the age of two. 46% were monolinguals; other languages spoken by participants were Danish, Italian, Japanese, Korean Spanish, Teochew, and Vietnamese. One participant was left-handed.

2.2. Stimuli

The stimulus set consisted of recordings of single word utterances of the form /hVd/. 9 words and 7 non-words contrasted all the stressed vowels of Australian English other than /ɪə/ and /eː/. The informant was a 21 y.o. monolingual female university student, born in Sydney to Australian-born parents. All stimuli were recorded at 44.1 KHz, amplitude-normalized, truncated to common temporal landmarks, and digitized as 16 bit WAV files.

Stimuli were also presented orthographically to elicit participants' responses to audition. All words were presented in the form <hVd> to maintain orthographic regularity across items [9] and to avoid multi-morphemic representations [3, 4]: /hɜːd/, /hæɔd/, /hʉːd/ were presented as *herd*, *howd*, *hude* (avoiding *heard*, *how'd*, *who'd*). Non-words were represented with transparent regular spellings and piloted with native speakers of AusE.

2.3. Procedure

Participants were seated in a semi-enclosed booth wearing Sennheiser HD 380 pro headphones, facing a computer monitor. Visual stimuli were presented using E-Prime [19] software on an Asus laptop (60 Hz screen refresh). Participants responded by pressing one of two labelled buttons on the button box response tool to log response accuracy and RT.

The procedure consisted of three phases: a familiarisation phase, a practice test, and an experimental phase. During the familiarisation phase, a single word was presented orthographically for 4000 ms and the same word was presented auditorily

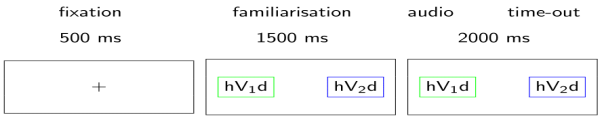

Figure 1: *Schematic structure of a trial in the test phase. Listeners selected the target word by pressing the corresponding button on the button box.*

at the midpoint of the visual presentation interval (at 2000 ms).

During the practice test, a non-word target and a non-word competitor were presented on the screen for 3000 ms. Audio presentation of the target word commenced at 1000 ms, allowing participants 2000 ms from the beginning of the auditory stimulus to make their choice. If they selected the target word correctly, they received the feedback "correct". If they gave an incorrect answer or did not provide an answer within 2000 ms, they received the feedback "incorrect" or "too slow" and the trial was repeated. The aim of the practice test was to strengthen the link between the spelling and pronunciation of non-words before the experimental phase.

During the experimental phase, participants first saw a fixation cross for 500 ms. The cross was followed by two words presented orthographically for 1500 ms, allowing participants time to read the words. 1500 ms after the two candidate words appeared on the screen, audio presentation of the target word began. Participants were allowed 2000 ms from the beginning of the auditory stimulus to identify the word by pressing a button on the same side as the corresponding orthographic representation. Participants were instructed to respond as quickly as possible. Figure 1 shows the schematic representation of trials in the experimental phase. If participants responded, the experiment moved on to the next trial without a feedback. If they did not respond within 2000 ms, they received the feedback "too slow" and the experiment only moved on to the next trial when the participant indicated that he/she was ready. The experimental phase was repeated three times, giving $3 \times 240 = 720$ trials.

2.4. Data analysis

We collected accuracy and reaction time (RT) data from the experimental phase. Accuracy and RT data were examined to determine whether any outliers should be removed. Participants' mean accuracy was 98.78% (range: 93%–100%), therefore no participant was excluded on the basis of accuracy. The mean RT of each participant fell within 2 standard deviations of the grand mean RT across all participants (644.8 ms), therefore no participant was excluded based on RT. Nine observations (three for each of the first three participants) were excluded, due to an error in stimulus presentation.

Raw data of all participants were transformed. Firstly, the percentage of inaccurate responses for each target word was calculated to determine which word was least accurately identified. Secondly, the percentage of inaccurate responses to each target and competitor pair was calculated to see which word-pair yielded the most inaccurate responses.

RT data was refined in two steps. The first step was to remove the incorrect responses or responses with an RT shorter than the onset of word-initial /h/+210 ms, as it takes approximately 210 ms to respond to stimulus [20]. Based on these criteria, 245 responses were excluded. Responses with too long RT were not trimmed [21] - the experiment had an inbuilt cutoff point at 2000 ms, because participants received a time-out message on screen after 2000 ms. Secondly, RT data was adjusted to two landmarks associated with the stimulus sound to accommodate the intrinsic length differences of the different vowels. The landmarks are shown in Figure 2. The first was the onset of the vowel (T1 on Figure 2), as marked by the end of the friction of /h/. The second was the offset of the vowel, (T2 on Figure 2) as marked by the closure of the /d/. Two RTs were calculated for each trial relative to the landmark vowel onset and the landmark vowel offset: RT from vowel onset is the time from the beginning of the vowel to the response and RT from vowel offset is the time from the end of the vowel to the response.

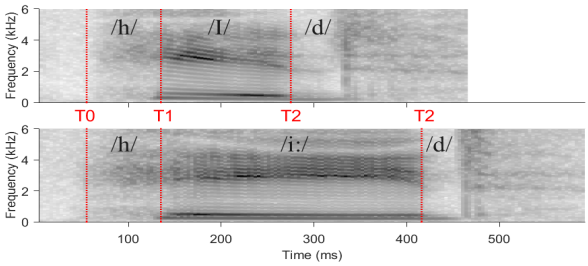

Figure 2: *Acoustic landmarks for vowel onset and offset exemplifed by the stimulus words* heed *and* hid*. T0 marks the onset of the stimulus, T1 marks the onset of the vowel and T2 marks the offset of the vowel. RT was measured from vowel onset (T1) and vowel offset (T2).*

The distribution of RT data followed Gamma distribution at both landmarks. Therefore, general-linear model with the family Gamma (GLM) was used to test RT. Gamma distribution is unable to handle the negative RT at vowel offset resulting from participants reacting before the end of the vowel. Therefore the constant 320 was added to RT from vowel offset prior to using GLM, as the lowest RT from vowel offset was -317 ms. All data analysis was conducted in R [22].

3. Results

3.1. Confusion result

The percentage of inaccurate responses per target vowel show which targets are the hardest to identify and which are the most confused vowel pairs. Table 1 shows the targets hardest to identify were /ɑɛ/ (4.1% of inaccurate responses), /æɔ/ (3.6%), /ɐ/ (3.4%), and /əʊ/ (3.1%). Table 1 also shows these targets also have a clear competitor, whereas /ɜː/ (the least confused vowel) was not confused with the same vowel more than once. There are also target vowels which were confused equally often with more than one competitor; in these cases all competitors were given in Table 1 and the % of errors shows the tie. However, confusion data has its limitations. Firstly, listeners are at ceiling in a binary choice task due to the high number of easy comparisons. Secondly, inaccurate answers may be a result of mispressing the buttons rather than confusing the two vowels, as 14 out of the 15 participants reported noticing they had made a mistake after pressing the answer button.

3.2. RT results

Preliminary analysis was conducted to examine the effect of target and competitor vowel on RT at vowel onset and offset. A null model without a factor and two full models with either target or competitor vowel as a factor were constructed. Target vowel had a significant effect on reaction time ($p \geq 0.0001$ at

onset and offset, df=15 at both landmarks). Competitor vowel did not have an effect ($p = 0.997$ at onset and $p = 0.996$ at offset, df=15 at both landmarks). The interaction of the two factors was not significant.

To examine how target vowels affect RT, and if the 16 phonemes can be grouped according to their phonetic features, target vowels were assigned binary features following the system of [23] and using the values appropriate for AusE [17]. The features are ±high, ±low, ±front, ±back and ±long. Diphthongs were classified according to their first element, because the first element is likely to be responsible for the confusion [16, 24]. The values were coded as 0 and 1 in R. The classification of vowels is shown in Table 1.

To examine the effect of target vowel on RT relative to vowel offset and offset, a null model without a factor and five full models with each feature as a factor were constructed. At vowel onset, the features +front and +long had a significant effect, as long targets have significantly longer RT, and front targets have significantly shorter RT. At vowel offset, +low and +long had a significant effect, as long and low targets have significantly shorter RT. These results are shown in Figure 3. Testing the effects of the features of competitor vowels did not return significant results.

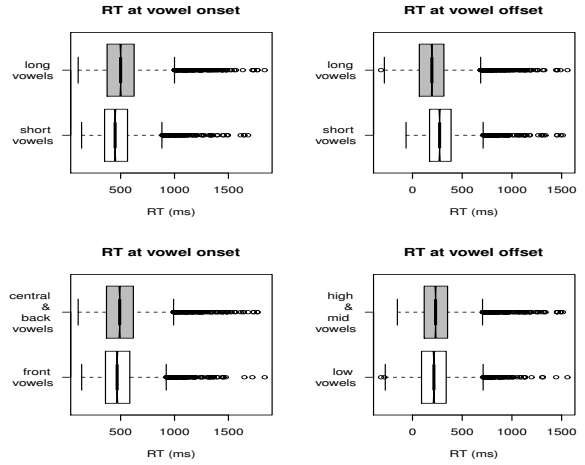

Figure 3: *RT measured from vowel onset (right panel) and from vowel offset (left panel) with the binary features ±long (upper panel), ±front (bottom left) and ±low (bottom right) as factors.*

To examine whether target vowel and competitor vowel interact five models were constructed for RT measured from vowel onset and from vowel offset. The dependent variable was RT and one feature of the target and the same feature of the competitor vowel were interacting factors. The goal was to determine if shared features of target and competitor vowel affect RT. All features except ±long interacted, showing that disambiguation is slower when target and competitor vowel share the features ±front, ±back, ±high, or ±low feature. Target and competitor length did not interact at either landmarks, showing that sharing the feature ±long does not affect disambiguation.

RT of target vowels was calculated across all competitors, and target vowels were sorted according to mean RT (from quickest to slowest). Table 2 shows that at vowel onset, short targets have short RT and long and diphthong targets have long RT. However, this is reversed at the vowel offset, when long and diphthong targets have short RT and short targets have long RT.

Table 1: *Classification of AusE vowels according to their binary features and the strongest competitor for each target according to percentage of incorrect responses and longest RT*

Target	Front	Back	High	Low	Long	Confused with	% of errors / strongest competitor	Overall accuracy (%)	Long RT
iː	+	–	+	–	+	ɪ	11	98.6	ɪ
ɪ	+	–	+	–	–	iː	11	97.8	ɐː
e	+	–	–	–	–	ɑe	4.4	98.6	iː
æ	+	–	–	+	–	æɔ	15	97.8	æɔ
ʉː	-	–	+	–	+	əʉ, ʊ	8	97.4	ʊ
ɜː	-	–	–	–	+	e, ɑe oː, ɐ, ʉː	2	99.1	ɐː
ɐː	-	–	–	+	+	æe	13	97.8	ɐ
ɐ	-	–	–	+	–	ɐː	20	96.6	ɐː
ʊ	-	+	+	–	–	oː, æɔ, ɔɪ	4	98.8	əʉ
oː	-	+	+	–	+	ɔɪ	11	98	ɔɪ
ɔ	-	+	–	–	–	ɔɪ	11	97.1	əʉ
æɪ	+	–	–	+	+	e	6	98.3	æɔ
ɑe	-	+	–	+	+	ɐː	22	95.9	ɐː
æɔ	+	–	–	+	+	æ	20	96.4	æ
əʉ	-	–	–	-	+	ɔɪ	11	96.9	ɔɪ
ɔɪ	-	+	+	–	+	oː, æɔ	8	97.7	æɔ

Next, two strongest competitors were selected: one with which the target was the most often confused, and a second with the longest pairwise RT. (The strongest competitor based on RT was the same at vowel onset and offset.) The strongest competitors are shown in Table 1. Table 1 also shows competitors usually share features with their targets.

Table 2: *Target vowels from shortest to longest mean RT at vowel onset and offset. Short monophthongs are in white cells, long monophthongs are in light grey cells and diphthongs are in grey cells.*

Time	Target vowel													
Onset	ɪ æ e	iː	ʊ	ɜː	ɔɪ	ɐ	æɪ æɔ	ɔ	oː	ɑe	ɐː	əʉ	ʉː	
Offset	ɐː ɜː oː	ɑe	iː	æɔ ɔɪ æɪ			æ e	ʉː	əʉ	ɐ ɪ ɔ ʊ				

4. Discussion

4.1. RT results at vowel onset and offset

Short vowels have shorter RT than long vowels when RT is measured from vowel onset, and long vowels have shorter RT when RT is measured from vowel offset. The fact that short vowels have shorter RT from onset is inherent to the task, because all information on a short vowel becomes available sooner for short vowels than for long vowels. That is, listeners have access to the whole vowel sooner when the vowel is short than when it is long. The fact that the relationship between vowel length and RT is reversed at the offset shows that listeners do not need to wait until the end of the long vowel to identify it. This is further supported by the finding that RT measured from the offset is negative in 1.26% of the responses for short targets, and RT is negative in 12.5% of the responses when the target is long. There are two possible interpretations for this result. The first is that short vowels are harder to identify, and the second is that there is a minimum exposure that is required for the identifica-

tion of the target vowel. The current data does not allow us to choose between these two explanations, and making this choice is beyond the scope of the present study.

Ordering target vowels from shortest to longest RT (see Table 2) points to disruptions in the general pattern. At the onset, where short targets have short RTs, long /iː/ patterns with the short vowels, and short /ɐ/ and /ɔ/ pattern with the long vowels. This may contribute to finding a significant effect of target frontness at vowel onset, as /ɔ/ is a short back vowel patterning with the long vowels, and /iː/ is a long front vowel patterning with the short vowels. Also, out of the 10 non-front vowels 7 are long, which can lead to frontness showing a significant effect. At vowel offset, low targets returned a significantly shorter RT than non-low targets; however, Table 2 shows that low vowels are spread evenly. Therefore the feature ±long seems to explain the differences in RT.

4.2. Target - strongest competitor pairs

The effect of binary features on vowel identification can be seen on the target-strongest competitor pairs. Table 3 shows that vowels confused with each other share one or more features. Also, the 5 idiosyncratic pairs share features between the target and the second element of the diphthong competitor. Vowel pairs that have the longest RT also share features.The targets /æ, æɪ, æɔ/ share both the +front and the +low features.

Table 3: *Target words (left) with their strongest competitors (right), as explained by shared features*

	Explained				Idiosyncratic
	Front	Back	High	Low	
Acc.	iː-ɪ, ɪ-iː æ-æɔ,æɪ-e, æɔ-æ	ʊ-oː,oɪ, ɔ-oɪ, oɪ-oː	iː-ɪ, ɪ-iː, ʉː-ʊ	æ-æɔ, ɐː-ɑe, ɐ-ɐː, ɑe-ɐː, æɔ-æ	e-ɑe, ʉː-əʊ, ʊ-æɔ, əʊ-ʉː, oɪ-æɔ
RT	iː-ɪ, e-iː, æ-æɔ,æɪ-æɔ, æɔ-æ	o-oɪ ʉː-ʊ, oː-oɪ	iː-ɪ, ʉː-ʊ oː-oɪ	æ-æɔ, ɐː-ɐ, ɐ-ɐː, æɪ-æɔ, ɑe-ɐː, æɔ-æ	ɪ-ɐː, ʊ-əʉ, ɔ-əʉ, oɪ-æɔ

The target-competitor pairs tend to differ in length: /iː-ɪ/, /ʉː-ʊ/, /ɐː-ɐ/, /æɔ-æ/. This confirms that AusE long-short vowel pairs are hard to disambiguate. Additionally, there are the monophthong-diphthong pairs, /æɔ-æ/, /oː-oɪ/, that share the first element and are perceptually similar. The limitation of choosing the strongest competitor for individual targets is however that the differences in RT between target-competitor pairs were small and may have been affected by lexical frequency [1].

5. Conclusions

Our results show that vowel disambiguation becomes harder when target vowel and competitor vowel share features. Disambiguation is the hardest when target and competitor only differ in length: long vowels are intrinsically hard to disambiguate from short vowels when the members of the vowel pairs have similar spectral qualities (including diphthongs that share their first element with a particular monophthong). We have identified five vowel pairs that have a high feature overlap (/iː-ɪ/, /ʉː-ʊ/, /ɐː-ɐ/, /æɔ - æ/, /oː-oɪ/) making them hard to disambiguate. Thus they are good candidates for more targeted research on the effects of phonemic similarity on vowel disambiguation.

6. References

[1] Meunier, F. and Segui, J. "Frequency Effects in Auditory Word Recognition: The Case of Suffixed Words", Journal of Memory and Language 41:327–344, 1999.

[2] Connine, M. C., Mullennix, J., Shernoff E. and Yelen, J. "Word Familiarity and Frequency in Visual and Auditory Word Recognition", Journal of Experimental Psychology: Learning, Memory, and Cognition 16(6):1084-1096, 1990.

[3] Röder, B., Demuth, L., Streb, J. and Rösler, F. "Semantic and morpho-syntactic priming in auditory word recognition in congenitally blind adults", Language and Cognitive Processes, 18(1):1-20, 2003.

[4] Vannest, J., Newport, E. L., Newman, A. J. and Bavelier, D., "Interplay between morphology and frequency in lexical access: The case of the base frequency effect", Brain Research 1373:144 – 159, 2011.

[5] Buchanan, L., Westbury, C. and Burgess, C., "Characterizing semantic space: Neighborhood effects in word recognition", Psychonomic Bulletin and Review, 8(3):531-544, 2001.

[6] Goldinger, S. D., Luce, P. A. and Pisoni, D. B., "Priming lexical neighbors of spoken words: Effects of competition and inhibition", Journal of Memory and Language, 28:501–518, 1989.

[7] Luce, P. A., Pisoni, D. B. and Goldinger, S. D., "Similarity neighborhoods of spoken words', in G. T. M. Altmann [Ed], Cognitive models of speech processing: Psycholinguistic and computational perspectives, 122–147, MIT Press 1990.

[8] Cluff, M. S. and Luce, P. A., "Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation.", Journal of Experimental Psychology: Human Perception and Performance, 16:551–563 1990.

[9] Ziegler, J. C., Ferrand, L. and Montant, M., "Visual phonology: The effects of orthographic consistency on different auditory word recognition tasks", Memory and Cognition 32(5):732–741, 2004.

[10] Assmann, P. F., Nearey, M. T. and Hogan, T. J., "Vowel identification: Orthographic, perceptual, and acoustic aspects", The Journal of the Acoustic Society of America 71:975-989, 1982.

[11] McMurray, B., Tanenhaus, M. K., Aslin, R. N., and Spivey, M. J., "Probabilistic Constraint Satisfaction at the Lexical /Phonetic Interface: Evidence for Gradient Effects of Within-Category VOT on Lexical Access" Journal of Psycholinguistic Research 32:77-97, 2003.

[12] Grosjean, F., "Spoken word recognition processes and the gating paradigm", Perception and Psychophysics 28:267-283, 1980.

[13] McQueen, J. M. and Viebahn, M. C., "Tracking recognition of spoken words by tracking looks to printed words", The Quarterly Journal of Experimental Psychology 60:661-671, 2007.

[14] Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K., "Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models", Journal of Memory and Language 38:419439, 1998

[15] Taft, M., "Lexical access codes in visual and auditory word recognition", Language and Cognitive Processes, 1(4):297–308, 1986.

[16] Cox, F. "Vowel Change in Australian English", Phonetica 56:1–27, 1999.

[17] Cox, F., Australian English Pronunciation and transcription, Cambridge University Press, 2012.

[18] Cutler, A., Native Listening: Language Experience and the Recognition of Spoken Words, MIT Press, 2012.

[19] Psychology Software Tools, I. E-Prime 2.0 2012.

[20] Woods, D. L., Wyma, J. M., Yund, E. W., Herron, T. J. and Reed, B. "Factors influencing the latency of simple reaction time", Frontiers in Human Neuroscience 9(131):1-12, 2015.

[21] Ratcliff, R., "Methods for dealing with reaction time outliers", Psychological Bulletin 114(3):510-532, 1993.

[22] R version 3.2.3 2015.

[23] Chomsky, N. and Halle, M., The sound pattern of English, MIT Press, 1968.

[24] Nearey, T. M. and Assmann, P. F., "Modeling the role of inherent spectral change in vowel identification", The Journal of the Acoustic Society of America 80:1287-1308, 1986.