Production and perception of length contrast in lateral-final rimes

Tunde Szalay, Titia Benders, Felicity Cox, Michael Proctor

Department of Linguistics, Macquarie University, Sydney, Australia
ARC Centre of Excellence in Cognition and its Disorders
{tunde.szalay, titia.benders, felicity.cox, michael.proctor}@mq.edu.au

Abstract

Words containing /l/-final rimes challenge listeners as coda /l/ reduces certain vowel contrasts. Lateral-final rimes therefore allow us to gauge the link between individuals' word recognition and production. We tested whether participants producing a larger durational contrast between word pairs containing the rimes /iːl-ɪl, ʉːl-ʊl, æɔl-æl , əʉl-ɔl/ were better at recognising minimal pairs contrasting the aforementioned rimes. 46 Australian English speakers produced 24 /l/-final minimal pairs and identified the same minimal pairs spoken by two speakers. Participants producing a longer durational contrast took longer to respond and were only more accurate when the stimuli contained a bigger durational contrast.

Index Terms: durational vowel contrast, production, perception, lateral-final rimes, Australian English

1. Introduction

A growing body of experimental evidence shows that individuals' speech production and perception are linked [1–4]. Listeners who robustly produce a contrast are better able to perceive the same contrast than listeners with less robust contrast production [1–4]. For instance, listeners who more accurately differentiate voiced from voiceless stops also produce longer voice onset time [1]. Listeners who are better at discriminating the /s-ʃ/ contrast in perception maintain a more consistent tongue-tip contrast in production [2]. Listeners who are better at discriminating /ɑ-ʌ, u-ʊ/ produce greater spectral differentiation between members within these vowel pairs [3].

In perception, the phonological contrast between vowels is cued by several acoustic cues, i.e. formant values of vowel targets [5], vowel inherent formant change [6], and duration [5]. English, including Australian English, listeners rely more on spectral than durational contrast and use durational contrast only when spectral contrast is diminished or unavailable [5, 7]. That is, spectrally similar vowels are more likely to be confused [8] even when they differ in length [9].

Spectral contrast is weighted more heavily than durational contrast at an individual level [10]. Listeners from a speech community where spectral contrast is maintained between the vowels in PULL-POOL-POLE in the pre-/l/ context cannot discriminate these vowels in the speech of another speech community, where only durational contrast is maintained [10]. However, speakers who reduce spectral difference but maintain durational difference in production can utilize durational cues in perception even when spectral cues are not available [10]. This indicates that listeners rely on the same cues in perception which they produce and perceive as phonologically contrastive. In contrast, in the production and perception of voiced and voiceless consonants, listeners were found to weight voice onset time and $f0$ differently [11].

The aforementioned studies tested contrast perception on continua of manipulated stimuli, therefore little is known about if and how contrast production is associated with listeners' ability to cope with variation in unmanipulated speech. To further our understanding of the production-perception link, this study examined if and how contrast production is associated with word recognition and processing in Australian English (AusE) lateral-final rimes.

The AusE vowel inventory contrasts 18 stressed vowels, using both spectral and durational contrasts [14]. Some vowel pairs differentiated by duration exhibit smaller spectral differences (e.g. /ɐː-ɐ, iː-ɪ/, in *cart-cut*, *beat-bit*), others exhibit bigger spectral contrast (e.g. /ʉː-ʊ/, in *kook-cook*) [14]. There are diphthong-monophthong pairs in which the first or the second target of the diphthong coincides with a monophthong [14]. These inherent spectral similarities increase vowel confusion [9]. Coda /l/ further reduces the spectral contrast between /iː-ɪ, ʉː-ʊ, æɔ-æ, əʉ-ɔ/ (e.g. *feel-fill, fool-full, howl-Hal, dole-doll*); however contrastive duration may be maintained [15]. It is not clear if listeners can use durational differences in /l/-final rimes.

This study examined perception of duration contrast in CVl minimal pairs contrasting /iː-ɪ, ʉː-ʊ, æɔ-æ, əʉ-ɔ/ in the speech of two Source Speakers, one of whom maintains a more robust duration contrast than the other. The association between participants' production of the same duration contrast and their perception was tested in three hypotheses:

1. if AusE listeners rely on durational cues in /l/-final rimes, increased duration contrast in the stimuli would aid word recognition for all listeners regardless of their contrast production
2. if production and perception are linked, listeners producing a consistent length contrast would have an overall advantage in recognising /l/-final words that differ in the duration of the rime in the speech of both Source Speakers
3. if listeners rely more on cues that they themselves produce, then listeners who produce a more robust duration contrast would only perform better when the Source Speaker does so too.

2. Method[1]

2.1. Participants

Forty-six female [mean age = 21.5, range = 18 – 40] native speakers of AusE participated in the study. All participants were born in Australia to Australian-born parents. None of the participants reported any reading, hearing, or speaking disorders. Participants received course credit or $15 for participation.

[1] Data was collected as a part of a broader project.

2.2. Materials

The stimuli consisted of 32 CVC targets and 38 (C)V(C) fillers. 4 vowel pairs (/iː-ɪ, ʉː-ʊ, æɔ-æ, əʉ-ɔ/) were embedded in two sets of /l/-final and two sets of /d/-final minimal pairs to create 32 target words (See Table 1 for the /l/-final words). Here we analyse only production and perception data of /l/-final words.

Table 1: *Target words ending in /l/*

Vowel pair			
/iː-ɪ/	/ʉː-ʊ/	/æɔ-æ/	/əʉ-ɔ/
feel-fill, heel-hill	*fool-full, pool-pull*	*howl-Hal, vowel-Val*	*mole-moll, coal-Col*

To create the stimuli for the perception experiment, targets and fillers were read by two female native speakers (Source Speakers) of AusE upon orthographic random presentation on a computer monitor. Source Speaker 1 was 25, and Source Speaker 2 was 57 years old at the time of the recording. All stimuli were recorded with an AKG C535EB Condenser Microphone onto an iMac using Presonus Studio Live 16.2.4 AT Mixer in a sound treated studio. Stimuli were recorded at 44.1 KHz, amplitude-normalized, truncated to have 1 s silence before and after the word, and digitized as 16 bit WAV files.

Long:short rime duration ratios were calculated for the vowel-pairs /iː-ɪ, ʉː-ʊ, æɔ-æ, əʉ-ɔ/ from the experimental stimuli produced by the two Source Speakers (Table 2). Source Speaker 2 maintained a bigger long:short ratio, therefore maintained a bigger duration contrast for all vowel pairs except /æɔ-æ/.

Table 2: *Long:short rime duration ratios in the stimuli*

Informant	Vowel pair			
	/iː-ɪ/	/ʉː-u/	/æɔ-æ/	/əʉ-ɔ/
Source Speaker 1	1.27	1.3	1.23	1.23
Source Speaker 2	1.47	1.45	1.23	1.42

2.3. Procedure

The experiment consisted of a production task followed by a perception task, carried out in a one hour long session in a sound treated studio at Macquarie University, Sydney NSW. Participants were tested individually with the experimenter present.

Firstly, participants read orthographically presented words aloud. Words were pseudo-randomised, presented one by one three times in three blocks and recorded with an AKG C535EB Condenser Microphone onto an iMac using Presonus Studio Live 16.2.4 AT Mixer. The production task helped participants familiarise with the stimuli for the perception task.

Next, participants carried out the perception task, consisting of a practice phase and a test phase. In the practice phase, 10 single words were individually presented auditorily. Participants were asked to type the word that they heard quickly and accurately and received immediate feedback on what the correct responses were. In the test phase, participants were presented with individual words auditorily and were asked to type the words as they perceived them as quickly and accurately as possible. First, participants heard the words spoken by Source Speaker 2, repeated twice in two blocks, and then by Source Speaker 1, repeated twice in two blocks; blocks were separated by 30 s long forced break. Items within a block were pseudo-randomised so that no /l/-final words followed each other. Stimuli were presented with Expyriment [16] on an Asus X550JX laptop. Audio stimulus was presented via Sennheiser 380 Pro headphones at participants' preferred listening level. Participants' responses accuracy and response time (RT) of the first keypress were measured. After the word recognition task, participants were asked to fill out a self-evaluation questionnaire.

3. Data analysis

3.1. Production data

Recordings were segmented automatically [17]; rime durations were extracted automatically [18]. Rime duration is a measure combining vowel and coda /l/ length. Duration values 1.5 times above or below the interquartile range for a given vowel were hand-checked and corrected for measurement errors.

Mean rime duration was calculated by participant and vowel. The ratio of long:short vowels for each vowel pair and for each participant was calculated; increased ratio indicates an increased duration contrast.

3.2. Perception data

Responses to 46 (participants) x 64 (/l/-final tokens) = 2944 trials were collected. Responses were rated for accuracy. Responses were classified as Intended Answer, Phonetic Respelling, Typo, Minimal Pair Error, and Other Error. Responses were classified as Intended Answer if spelled as the target. Unambiguous but nonstandard phonetic spellings (e.g. *cole* for *coal*) were classified as Phonetic Respellings. Single letter deletions, additions, letter transpositions, and substitutions within one key distance of the target letter were classified as Typos [19], unless the result was an English lexical item. Confusion of members of minimal pairs (e.g. *fool* for *full*) was classified as Minimal Pair Error. Any other error (e.g. *cool* for *pool*, *howled* for *howl*) were classified as Other Error. For the purposes of the analysis of accuracy, Intended Answers, Phonetic Respellings and Typos were accepted as Correct; Minimal Pair Errors and Other Errors were classified as Incorrect.

RT of the first keypress was collected. RT within 210 ms [20] or above 5000 ms [21] of stimulus onset were excluded from analysis. Individual RT exceeding or less than mean+±2*sd for each participant were excluded from analysis [22]. 5.1% of responses were excluded according to these criteria, leaving 2,794 tokens for analysis.

4. Results

4.1. Individual variation in production and perception

Participants produced /l/-final rimes with a mean long:short ratio of 1.34 and a range of 0.99-1.38.[2] Participants consistently produced a decreasing durational contrast from /iː-ɪ/ to /ʉː-ʊ/ to /æɔ-æ/ to /əʉ-ɔ/. In the perception data, participants were consistent across the vowel pairs.

4.2. Production-perception link

To measure the association between accuracy, RT, and duration ratio, we constructed two Generalised Linear Mixed-effect models [23] with the dependent variables Accuracy and RT. The independent variables were Participant Duration Ratio (long:short, scaled), Vowel Pair (contrast coded and each compared against the grand mean), Source Speaker (contrast coded), and Lexical Frequency (from [24], log-normalised); Participant and Block were random intercepts. All two-way interactions

[2]Mean long:short vowel ratio was 1.64 in /d/-final rimes, as in [14].

between Duration Ratio, Vowel Pair, and Source Speaker were included in the model, but three-way interactions were not; Lexical Frequency did not interact with the other independent variables. Effects on accuracy were tested using the binomial family and effects on RT with the gaussian family with log-normal link, as raw RT followed a log-normal distribution.

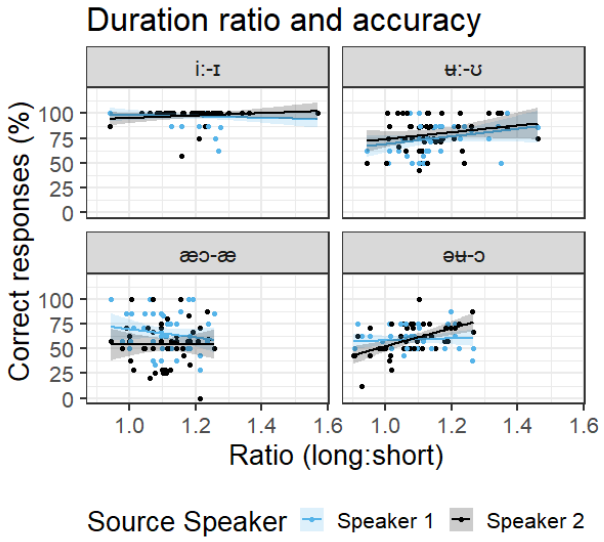

Figure 1: *Correlation of participants' duration ratio (x-axis) and recognition accuracy (y-axis) by Source Speaker (blue: Speaker 1, black: Speaker 2) and Vowel Pair (panels). Top: /iː-ɪ/ and /ʉː-ʊ/ contrast. Bottom: /æɔ-æ/ and /əʉ-ɔ/ contrast.*

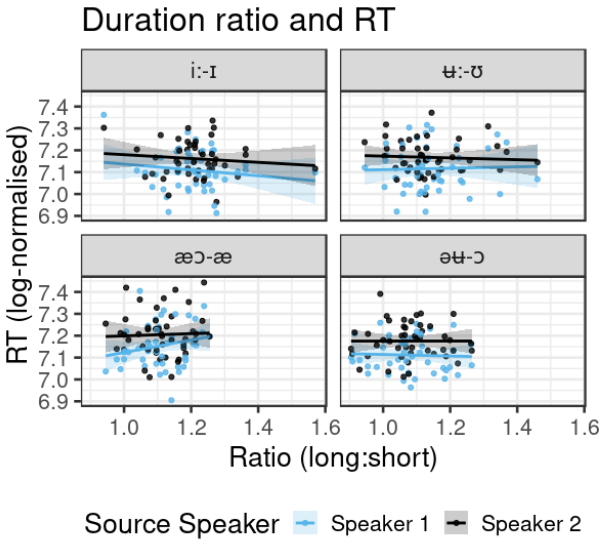

Figure 2: *Correlation of participants' duration ratio (x-axis) and perceptual RT (y-axis) by Source Speaker (blue: Speaker 1, black: Speaker 2) and Vowel Pair (panels). Top: /iː-ɪ/ and /ʉː-ʊ/ contrast. Bottom: /æɔ-æ/ and /əʉ-ɔ/ contrast.*

Participant Duration Ratio did not affect Accuracy significantly, but participants with larger Participant Duration Ratio had significantly slower RT (β=0.02, F(1, 4097)=9.53, p<0.001). Source Speaker did not affect Accuracy significantly, but participants responded more slowly to words produced by Source Speaker 2 (β=0.03, F(1, 4097)=0.0002, p=0.01). Participant Duration Ratio showed a significant positive interaction with Source Speaker 2 on accuracy (β=0.13, F(1, 5572)=9.74, p=0.002): participants with a larger long:short ratio recognised words more accurately when produced by Source Speaker 2, who produced larger duration contrast. Participant Duration Ratio and Source Speaker did not show significant interaction on RT.

Vowel Pair effects showed that /iː-ɪ/ was disambiguated more accurately (β=1.43, F(3, 5572)=105.95, p<0.001) and more quickly (β=-0.64, F(3, 4097)=99.11, p<0.001) than other Vowel Pairs. /ʉː-ʊ/ was disambiguated less accurately (β=-0.92, F(3, 5572)=105.92, p<0.001) but more quickly (β=-0.05, F(3, 4097)=99.11, p<0.001) than other Vowel Pairs. /ʉː-ʊ/ and Source Speaker 2 showed a negative interaction on RT (β=-0.02, F(3, 4097)=8.25, p<0.001): the RT difference between responses to Source Speaker 1 and 2 was smaller for /ʉː-ʊ/ than for other Vowel Pairs. /æɔ-æ/ was disambiguated less accurately (β=-0.18, F(3, 5572)=105.92, p=0.048) and more slowly (β=0.11, F(3, 4097)=99.11, p<0.001) than other pairs with 59% response accuracy and log-normalised 7.23 ms RT, in contrast with the overall response accuracy of 73% and log-normalised RT of 7.18 ms. Interactions between Participant Duration Contrast and Vowel Pair /æɔ-ɔ/ showed that participants with larger long:short ratio disambiguated /æɔ-ɔ/ less accurately (β=-0.22, F(3, 5572)=3.08, p=0.012) and more slowly (β=0.1, F(3, 4097)=1.69, p=0.04). Interaction between Source Speaker 2 and Vowel Pair /æɔ-æ/ showed that /æɔ-æ/ was disambiguated less accurately when produced by Source Speaker 2 (β=-0.25, F(3, 5572)=7.23, p=0.001).

Increased Lexical Frequency lead to increased accuracy (β=0.52, F(1, 5572)=256.3, p<0.001) and to increased RT (β=0.02, F(1, 4097)=63.23, p<0.001). Increase in RT with the increase in Lexical Frequency was probably due to the fact that there were more high frequency words among the targets with long acoustic duration.

4.3. Summary of findings

1. Contra to hypothesis 1, increased durational contrast in the speech of Source Speaker 2 did not assist word recognition, suggesting that not all listeners rely on durational cues.
2. Contra to hypothesis 2, participants who produced an increased durational contrast were not overall better at word recognition but they were overall slower.
3. In accordance with hypothesis 3, participants producing a larger duration contrast were more accurate on the contrast produced by Source Speaker 2, who, like them, maintained a larger durational contrast.

5. Discussion

Accuracy data showed that increased duration contrast in the stimuli aided word recognition only when participants also produced a more robust durational contrast. This indicates that perception is aided by cues that speakers themselves produce, but speaker-listeners without a robust durational cue production could not gain perceptual benefits. We found no evidence for overall better perception by participants with more robust duration contrast, contrary to [2, 3]. These discrepancies may be attributed to the differing methods, as we used an open-ended word recognition task, not contrast discrimination.

RT data showed that participants' increased rime duration contrast was associated with overall longer RT, indicating that these participants might consistently monitor for durational contrast. Durational contrast might take longer to process than spectral cues, as spectral cues may be available earlier in the

vowel, whereas the whole rime needs to be processed for the perception of durational cues [25, 26, c.f. 27]. The overall increase in RT with the increase in durational contrast in production indicates that speaker-listeners who rely on durational contrast in perception always monitor for it. However, the fact that these speaker-listeners are not overall more accurate indicates that they cannot always find durational contrast.

All participants responded more slowly to Source Speaker 2, despite Source Speaker 2 producing overall shorter target words than Source Speaker 1. The reason might lie in the potentially different spectral quality of the Source Speakers' vowels, in Source Speaker 2 always being presented first, or in the fact that Source Speaker 1 was closer in age to the participants.

Words contrasting the four vowel pairs were recognised differently and showed complex interactions with participants' production. Words contrasting /iː-ɪ/ were recognised more efficiently, potentially due to the F2 differences between /iː/ and /ɪ/ at vowel onset in the stimuli. Minimal pairs contrasting /æɔ-æ/ were poorly recognised, probably because neither of the Source Speakers produced a robust durational contrast for this vowel pair. All participants performed less accurately on Source Speaker 2's production of the /æɔ-æ/ contrast. Moreover, participants with a bigger durational contrast performed *worse* on the overall recognition of the /æɔ-æ/ contrast. That is, participants with bigger durational contrast did not perform better on Source Speaker 2, contrary to their performance with other vowel contrasts, as they may have been looking for a durational contrast that was not present. Patterns of minimal pair recognition contrasting /æɔ-æ/ are consistent with hypothesis 3, in which listeners' perception is aided by cues that they themselves produce.

These findings suggest that listeners can only benefit from durational cues in vowel perception when they themselves produce it. Similarly, in [10]'s study listeners who could not use durational contrast were members of a different speech community and maintained spectral contrasts (and presumably a non-phonological durational contrast as well), whereas participants in our study were members of a single speech community. These results do not allow us to determine the cues that listeners without a durational contrast use to identify /l/-final words. Future work will analyse listeners' spectral contrast production and link it to their perception of /l/-final minimal pairs.

6. Conclusion

Slower discrimination of /l/-final rimes by individuals who produce larger durational contrast implies that these speaker-listeners may monitor for durational contrast. This makes word identification slower, but only leads to increased accuracy when the speaker produces a sufficient durational contrast too. This implies that robust durational contrast production may come at a price and with limited benefits in word recognition.

7. Acknowledgements

We thank the Phonetics Lab at Macquarie University. This research was supported in part by iMQRTP 2015144, ARC DE150100318, and MQSIS 9201501719 grants.

8. References

[1] Newman, R., "Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report", J. Acoust. Soc. Am., 113(5):2850–2860, 2003.

[2] Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H. Zandipour, M., Marrone, N., Stockmann, E. and Guenther, F. H, "The distinctness of speakers' /s/-ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect", J. Speech Hear. Res., 47(6):1259–1269, 2004.

[3] Perkell, J. S., Guenther, F. H, Lane, H., Matthies, M. L., Stockmann, E., Tiede, M. and Zandipour, M., "Cross-subject correlations between measures of vowel production and perception", J. Acoust. Soc. Am., 116(4):2338–2344, 2004.

[4] Zellou, G., "Individual differences in the production of nasal coarticulation and perceptual compensation", J. of Phonetics 61:13–29, 2017.

[5] Bennett, D. C., "Spectral form and duration as cues in the recognition of English and German vowels", Language & Speech 11(2):65–85, 1968.

[6] Nearey, T. M. and Assmann, P. F., "Modelling the role of inherent spectral change in vowel identification", J. Acoust. Soc. Am., 80(5):1297–1308, 1986.

[7] Liu, S., "The effect of vowel duration on native Mandarin listeners' perception of Australian English vowel contrasts in voiced and voiceless coda contexts", M.Res. thesis, Dept. of Linguistics, Macquarie Univ., Sydney, NSW. 2016.

[8] Neel, A. T. "Vowel space characteristics and vowel identification accuracy", J. Speech Hear. Res., 51(3):574–585, 2008.

[9] Szalay, T., Benders, T., Cox, F. and Proctor, M., "Disambiguation of Australian English vowels", in C. Carignan and M. D. Tyler [Eds] Proc. of 16th Speech Sci. and Technol. Conf., 73-76, 2016.

[10] Wade, L., "The role of duration in the perception of vowel merger" J. of Laboratory Phonology 8(1), 2017.

[11] Shultz, A. A., Francis, A. L. and Llanos, F., "Differential cue weighting in perception and production of consonant voicing", J. Acoust. Soc. Am., 132(2):EL95–EL101, 2012.

[14] Cox, F., "The acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers", Australian J. of Linguistics 26(2):147–179, 2006.

[15] Palethorpe, S. and Cox, F., "Vowel modification in pre-lateral environments", Int. Seminar on Speech Prod., Sydney, 2003.

[16] Krause, F. and Lindemann, O., "Expyriment: A Python library for cognitive and neuroscientific experiments", Behaviour Res. Methods, 46(2):416–428, 2014.

[17] Schiel F., "Automatic phonetic transcription of non-prompted speech", in J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey [Eds], Proc. of the ICPhS, 607-610, 1999.

[18] Boersma, P. and Weenink, D., Praat: doing phonetics by computer v6.0.25, Online: http://fon.hum.uva.nl/praat/, accessed in 2017.

[19] Luce, P. A. and Pisoni, D. B., "Recognizing spoken words: The neighborhood activation model", Ear & Hearing 19(1):1-36, 1998.

[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 of Human Neuroscience 131(9):1-12 2015.

[21] Baayen, H. R. and Milin, P., "Analysing reaction times", Int. J. of Psychological Res. 3(2):12–28, 2010.

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

[23] Bates, D., Mächler, M., Bolker, B. and Walker, S., "Fitting linear mixed-effects models using lme4", J. of Statistical Software 67(1):1–48, 2015.

[24] Davies, M., "Corpus of global web-based English: 1.9 billion words from speakers in 20 countries", Online: https://corpus.byu.edu/glowbe/, accessed in 2017.

[25] McMurray, B., Clayards, M. A., Tanenhaus, M. K. and Aslin, R. N., "Tracking the time course of phonetic cue integration during spoken word recognition", Psychonomic Bulletin & Review, 15(6), 1064-1071 2008.

[26] Reinisch, E. and Sjerps, M. J., "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context", J. of Phonetics, 41(2), 101-116 2013.

[27] Tillman, G., Benders, T., Brown, S. D. and van Ravenzwaaij, D., "An evidence accumulation model of acoustic cue weighting in vowel perception.", J. of Phonetics, 61, 1-12 2017.