Cross-accent word recognition is affected by perceptual assimilation

Sarah M. Wright[1], Mark D. Lathouwers[2], Catherine T. Best[1,3,4], Michael D. Tyler[1,2]

[1]The MARCS Institute, Western Sydney University, Australia
[2]School of Social Sciences & Psychology, Western Sydney University, Australia
[3]School of Humanities & Communication Arts, Western Sydney University, Australia
[4]Haskins Laboratories, New Haven, Connecticut, USA
sarah.wright@westernsydney.edu.au

Abstract

A single-item shadowing task was conducted to determine how identification of London-accented words by Australian listeners is affected by perceptual assimilation. This was evaluated in conjunction with two other well-established effects on word recognition: word frequency and talker variability. The results replicate frequency and talker variability effects, support the hypothesis that talker and accent normalisation operate at different processing stages, and show that words with nativelike assimilation of all phonemes are identified more accurately than those with category goodness or category shifting assimilation. Results are evaluated in view of episodic theories of lexical access.

Index Terms: cross-accent speech perception, perceptual assimilation, shadowing, word frequency, talker variability

1. Introduction

Variation in speech results from a variety of factors including individual differences in speaker gender, vocal tract characteristics, and speaker origin [1]. Despite this, the average person usually quite easily understands words across their many varying forms without any conscious awareness of adjusting for talker or token variability.

Accented speech is commonly encountered, and may affect comprehension and speed of processing, at least initially. For example, although initially disturbed, comprehension of an unfamiliar accent returns to nativelike levels within one minute of listening [2]. Additionally, [2] suggests that comprehension and speed of processing may vary depending on how 'thick' the accent is perceived to be. That is, accents can be ranked on a perceptual scale as a function of their distance from the native accent with foreign accents being more accented compared to regional accents of the native language. This suggestion is supported by their finding of a 100-150 ms delay in word identification in a foreign accent compared to the native accent, whereas [3] found only a 30 ms delay in word identification by French participants listening to an unfamiliar French accent.

However, even anecdotally, there are likely to be situations where the perceived strength of an accent is not related to whether it is a foreign or regional accent to the listener. For example, an Australian English speaker may perceive a Dutch accent as less accented, and therefore more comprehensible, compared to a regional accent of English such as Glaswegian.

The Perceptual Assimilation Model (PAM) [4-6] offers an alternative way of using phonetic and phonological differences to categorise accent strength in an unfamiliar accent compared to a perceiver's own accent. Although PAM was designed to account for perceptual assimilation of non-native phones, its principles can also be applied to cross-accent perception [7]. Phonemes in the unfamiliar accent that are perceived as good exemplars of the same phonological category in the native accent are assimilated to the native accent as *nativelike* (NL). A *category goodness* (CG) assimilation occurs when a non-native phoneme is perceived as an acceptable but not ideal exemplar of the same category in the native accent. It should be identified as the same phonological category but deviate in goodness of fit. Finally when a phoneme in the unfamiliar accent is sufficiently disparate from the native accent to be perceived as a different native phonological category, it is a *category shifting* (CS) assimilation. This model proposes that that individual words in an unfamiliar accent may be perceived as more or less accented depending on which phonemes the words contain and how those phonemes are assimilated to native phonological categories. Therefore, the main aim of the present study is to investigate how assimilation type affects cross-accent word recognition. We hypothesise that word recognition will be most accurate and efficient for NL assimilations, followed by CG, then CS types.

Speech from multiple talkers introduces variability that is less systematic and narrower in scope than variability across accents [5]. According to episodic theories of lexical access, variability due to talkers and accents is stored in long-term memory and is used to facilitate word recognition as opposed to the pre-lexical identification of phonemes as proposed by abstractionist theories [5]. Talker variability effects are well established. For example, word identification in a single-item shadowing task was slower and less accurate when words were produced by 15 different talkers, compared to those produced by a single talker [8], indicating that talker variability may not be removed prior to lexical access. In addition, high-frequency words were identified more accurately (but not more quickly) than low-frequency words. Thus, both word frequency and amount of talker variability affect the time course and/or accuracy of lexical access in the native accent, but little is known about how cross-accent assimilation is modulated by the presence of these other types of variation. To assess this, word recognition across accents was tested. Australian participants repeated aloud [following 8] high- and low-frequency words spoken in either their native Australian accent (AusE) or the much less familiar Southeast London accent, by either one or multiple speakers.

2. Method

2.1. Participants

Forty-eight introductory psychology students (41 females) at Western Sydney University participated for course credit.

Participants were aged between 18 and 35 years (M = 21.15, SD = 3.57), had all been exposed to AusE from birth, and reported no hearing or speech disorders at the time of testing. An additional 12 participants were tested but their data were discarded for incomplete responses (n = 1), previous exposure to a UK accent (n = 4), not being exposed to AusE from birth (n = 4), being outside of the 18-35 age range (n = 1), and equipment malfunction (n = 2).

2.2. Research Design

The experiment was a 2 x 2 x (2 x 2 x 3) mixed-design with reaction time (RT) and accuracy as dependent variables. Number of talkers (single vs. multiple) and accent (AusE vs. London) were manipulated between subjects, and word frequency (low vs. high), number of syllables (1 vs. 2), and London word assimilation type (NL vs. CG vs. CS) were manipulated within subjects. Participants were randomly assigned to one of the four between-subject cells (n = 12).

2.3. Stimulus Materials

The target stimuli consisted of 132 words selected from an existing cross-accent spoken word corpus. These words were selected based on assimilation type, frequency and number of syllables. See Table 1 for item totals across these variables.

The cross-accent assimilation types were determined based on published phonetic descriptions of the London and Australian accents. For example, all the phonemes in London-accented *baby* ([bæɪbi]) should be assimilated as good exemplars of the same native phonological categories (an NL assimilation). In the London-accented word *note* ([nəʉʔ]), all phonemes are assimilated as good exemplars of the same native categories, except for /ʔ/, which is assimilated as a poor exemplar of the native /t/ category (a CG assimilation). For London-accented *thorny* ([foːni]), all phonemes are assimilated as good exemplars of their native phonological categories except for the initial consonant which is assimilated to the native /f/ category instead of the intended /θ/ target (a CS assimilation). High-frequency words had a frequency of 50 or more per million (M = 214.5, SD = 127.1) and low frequency items had less than 10 occurrences per million (M = 2.8, SD = 2.4).

Table 1: *Mean duration (ms), mean frequency per million, and number of items by number of syllables, assimilation type, frequency, and accent.*

Syllables	Assimilation	Frequency	Accent	M_{dur}	M_{freq}	n_{items}
1	Nativelike	High	AusE	590	243	12
			London	549		
		Low	AusE	568	4	12
			London	529		
	Category Goodness	High	AusE	575	167	12
			London	470		
		Low	AusE	581	3	12
			London	524		
	Category Shifting	High	AusE	582	256	8
			London	532		
		Low	AusE	586	3	8
			London	509		
2	Nativelike	High	AusE	615	216	12
			London	632		
		Low	AusE	653	2	12
			London	660		
	Category Goodness	High	AusE	645	196	12
			London	640		
		Low	AusE	644	3	12
			London	642		
	Category Shifting	High	AusE	573	210	10
			London	585		
		Low	AusE	631	2	10
			London	657		

There were six stimulus lists per between-subjects condition, and each was used twice across the 12 participants in each group. In the single-talker condition, there was one list for each of the six speakers. In the multiple talker condition, each speaker's words were distributed across the six lists using a Latin square design, such that each token was presented an equal number of times across the single- and multiple-talker groups. The presentation was blocked by assimilation type with high- and low-frequency words presented randomly within each. NL and CG blocks were presented first in counterbalanced order. The CS block was always presented last, as the smaller number of items meant it might need to be excluded at a later stage.

2.4. Apparatus

The experiment was controlled using PsyScope X [9] running on a MacBook with an Edirol sound card. Participants listened to the stimuli through headphones and responded into a headset microphone (Beyerdynamic DT290). Trials were progressed using a voice key, and then reaction time was calculated manually from the recorded waveform, from the onset of the stimulus to the onset of the participant's response.

2.5. Procedure

Each trial began with the word *ready* displayed on the screen followed by the auditory presentation of the test item. Participants were instructed simply to repeat the word that they heard as quickly and as accurately as possible. If they took longer than 3.5 s to respond they were instructed to respond more quickly. There were no breaks between blocks. At the completion of the shadowing task, participants completed a language background questionnaire.

3. Results

RTs were analysed for correct responses only. Trials were scored as incorrect if a response was a word other than the test item, except for plurality errors (e.g., *papers* instead of *paper*; 8.33% of all trials). Responses that timed out were also considered incorrect (0.28% of all trials). RT and accuracy data were analysed using analysis of variance with planned contrasts (see 2.2 for the research design). The assimilation type variable was analysed using three non-orthogonal contrasts: 1) NL vs. CG, 2) NL vs. CS, and 3) CG vs. CS. An alpha level of .019 was used to adjust for using multiple contrasts [see 10]. For brevity only main effects and significant interactions involving the accent variable are reported below.

3.1. Reaction Time

There were four significant main effects for reaction time. Reaction times were shorter when words were presented from a single speaker (812 ms) compared to multiple speakers (918 ms), $F(1, 44) = 8.55$, $p = .005$, $\eta_p^2 = .16$, and for words in their native accent (788 ms) compared to the London accent (942 ms), $F(1, 44) = 18.00$, $p < .001$, $\eta_p^2 = .29$. High-frequency

words (819 ms) were responded to more quickly than low-frequency words (911 ms), $F(1, 44) = 233.83$, $p < .001$, $\eta_p^2 = .84$, and participants responded more quickly to 1-syllable (840 ms) than 2-syllable (890 ms) words, $F(1, 44) = 125.38$, $p < .001$, $\eta_p^2 = .74$. There were no significant effects of assimilation type on RT.

These main effects were moderated by two significant interactions. An accent × frequency interaction, $F(1, 44) = 44.57$, $p < .001$, $\eta_p^2 = .50$, indicates that the mean difference between low- and high-frequency words was significantly greater in the London accent ($M_{diff} = 133$ ms) compared to the Australian accent ($M_{diff} = 52$ ms). An accent × syllable interaction, $F(1, 44) = 26.96$, $p < .001$, $\eta_p^2 = .38$, indicates that the mean difference between 1- and 2-syllable words was significantly greater in the London accent ($M_{diff} = 73$ ms) compared to the Australian accent ($M_{diff} = 27$ ms).

3.2. Accuracy

The accuracy results are presented in Figure 1. They are collapsed across the number of talkers variable because it did not interact significantly with any other factors. Overall mean accuracy was higher for the Australian accent (96%) compared to the London accent (87%), $F(1, 44) = 134.22$, $p < .001$, $\eta_p^2 = .75$, for the single-talker (92%) than the multiple-talker condition (90%), $F(1, 44) = 11.24$, $p = .002$, $\eta_p^2 = .20$, for high-frequency (93%) than low-frequency items (89%), $F(1, 44) = 66.78$, $p < .001$, $\eta_p^2 = .60$, and for 2-syllable words (95%) than 1-syllable words (87%), $F(1, 44) = 100.30$, $p < .001$, $\eta_p^2 = .70$. The three non-orthogonal contrasts on the assimilation type variable were all significant: NL vs. CG: $F(1, 44) = 27.50$, $p < .001$, $\eta_p^2 = .38$; NL vs. CS: $F(1, 44) = 46.93$, $p < .001$, $\eta_p^2 = .52$; and, CG vs. CS: $F(1, 44) = 7.33$, $p = .01$, $\eta_p^2 = .14$. These results show that accuracy was significantly different across all levels of assimilation type with NL items having the highest accuracy (95%) followed by CG (90%) and then CS (88%). Note, however, that this is collapsed across the accent factor, and that the assimilation types are only relevant for the words when they are spoken in a London accent.

The main effects were moderated by four significant two-way interactions and a three-way interaction. The first two-way interaction was between accent and frequency, $F(1, 44) = 7.70$, $p = .008$, $\eta_p^2 = .14$. As can be seen in Figure 1, the difference in accuracy between high- and low-frequency words was greater for the London accent compared to the Australian accent. The interaction between accent and syllable, $F(1, 44) = 24.09$, $p < .001$, $\eta_p^2 = .35$, indicates that the difference in accuracy for words spoken in AusE and London accents is greater for 1-syllable than 2-syllable words. Additionally, these two effects were moderated by a three-way interaction of accent, frequency and syllable, $F(1, 44) = 6.91$, $p = .012$, $\eta_p^2 = .14$. Figure 1 shows that the greater effect of low frequency than high frequency words on accuracy in the London versus AusE accent was more pronounced for the 1-syllable than the 2-syllable words.

Finally, two of the assimilation type contrasts interacted significantly with accent: NL vs. CG, $F(1, 44) = 33.80$, $p < .001$, $\eta_p^2 = .43$; and NL vs. CS, $F(1, 44) = 26.16$, $p < .001$, $\eta_p^2 = .37$. These interactions show that both CG and CS words were recognised less accurately than NL words in the London accent compared to the AusE. However, the absence of a significant interaction between accent and the CG vs. CS contrast indicates that the mean difference in accuracy for the CG vs. CS items in AusE (2%) is not significantly different from the mean difference of the accuracy for the same items in the London accent (3%). This suggests that the significant main effect contrast between CG words and CS words is due to inherent properties of the words themselves rather than to pronunciation differences when those words are spoken in AusE versus London accents.

4. Discussion

Consistent with previous research, participants were less accurate and took longer to recognise words in a non-native regional accent than in their own accent. This is consistent with episodic theories of lexical access that suggest that items falling to the edges of categories (pronunciations that deviate

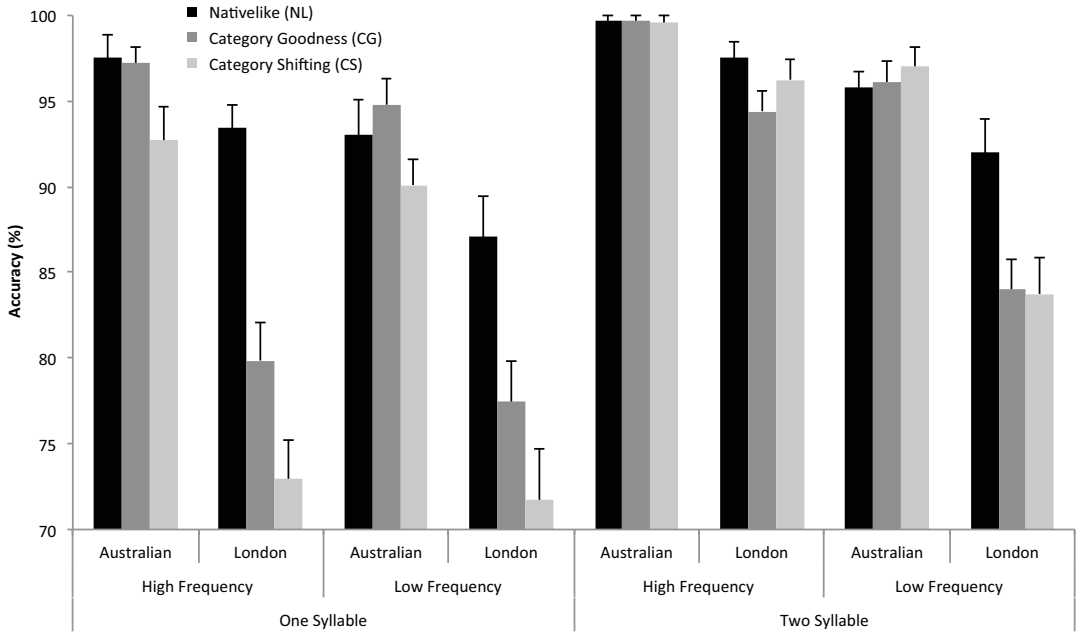

Figure 1: *Accuracy data collapsed across number of talkers, separated by accent, word frequency, number or syllables, and assimilation type. Error bars represent standard error of the mean.*

from the native accent) will take longer to be identified than items that are more typically encountered as in the native accent. Crucially, however, the effect of accent was moderated by assimilation type. Word recognition in the London accent was perturbed less for words that were assimilated as nativelike than for words that are phonetically and/or phonologically different from the native pronunciation (CG and CS assimilations). Contrary to our hypothesis, however, we observed no relative difference in accuracy or reaction time between category-goodness and category-shifting assimilations. However, as there were a smaller number of CS items compared to CG and NL items, due to the constraints of the existing corpus, it is possible that the number of observations was insufficient to provide a reliable estimate of the accuracy difference between CG and CS items. Future research should investigate this with a larger stimulus sample. Nevertheless, our results show a clear effect of perceptual assimilation on cross-accent word recognition between the NL items and the CG and CS items. This indicates that not all accent differences are equal and that accuracy is likely to be higher for items that are assimilated as good exemplars of native phonological categories, than phonetically and/or phonologically mismatching exemplars. We suggest, therefore, that perceptual assimilation may also play a role in recognition of foreign-accented speech.

The main effect of speaker number on reaction time replicates the well-established effect of number of speakers on word recognition [8], consistent with episodic models. However, the lack of an interaction between number of speakers and accent for either the accuracy or RT data suggests that the mechanisms for normalising across multiple speakers and normalising across accents may be operating at different stages of processing. For example, it has been suggested that speaker normalisation may occur at a pre-lexical stage and accent normalisation at a lexical stage [11], the latter being more consistent with abstractionist models.

The main effect of frequency as well as the frequency by accent interaction for RT also provides support for episodic theories of lexical access. That is, the robust frequency effects found in previous studies (e.g., [8]) have been replicated. It is also apparent that this frequency effect is moderated by accent. This is likely to correspond to the number of typical (native) versus deviant (other-accent) episodes held in memory, and that this number is reduced more so for low-frequency words compared to high-frequency words. This finding is also supported by the main effect of frequency, and the accent by frequency interaction for the accuracy data.

The main effect of syllable is unremarkable as it is expected that one-syllable words would be identified more quickly than two syllable words. The accent by syllable interaction for the RT data should be interpreted with caution in this case, as stimulus length was not controlled across accents. However, the accent by syllable interaction for the accuracy data is an interesting finding. Although it is expected that one syllable words will have lower accuracy than two syllable words, as shorter words tend to have more lexical neighbours [12], this effect appears to be exacerbated by the phonetic ambiguity introduced by a non-native regional accent. Additionally, the three-way interaction among syllable, accent and frequency for the accuracy data suggests that the ambiguity introduced by the non-native accent cancels out the effects of any frequency benefit for one-syllable words, which are more easily confused with each other due to their larger lexical neighbourhoods, as compared to two-syllable words.

Finally, on a methodological note, the shadowing task is useful because it measures performance at a point after word recognition has occurred. However, the obtained results, particularly RT, may have been affected by the reliance of this task on speech planning and production. Future research should employ a purely perceptual task such as lexical decision or a visual world word recognition paradigm to tease out this issue.

5. Conclusions

This study replicated the well-established effects of talker variability and word frequency on speed and accuracy of lexical access in a single-item shadowing task. It provides support for the idea that the mechanisms for normalising across accents and multiple talkers may operate at different stages of processing. Finally, it produces a new finding that not all accent differences impede lexical access equally. Rather, cross-accent word recognition is moderated by assimilation type: NL items in the London accent are identified more accurately than CG or CS items.

6. Acknowledgements

This research was funded by ARC grant DP120104596.

7. References

[1] D.B. Pisoni & S.V. Levi, “Some observations on representations and representational specificity in speech perception and spoken word recognition,” in *The Oxford handbook of psycholinguistics*, M.G. Gaskell, Ed. Oxford, UK: Oxford University Press, 2007, pp 3-18.

[2] C.M. Clarke & M.F. Garrett, “Rapid adaptation to non-native speech,” *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3647-3658, 2004.

[3] C. Floccia *et al.*, “Does a regional accent perturb speech processing?,” *J. Exp. Psych.: Human Per. and Perf.*, vol. 32, no. 5, pp. 1276-1293, 2006.

[4] C.T. Best, “A direct realist view of cross-language speech perception,” in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Baltimore, MD: York Press, 1995, pp. 171-204.

[5] C.T. Best, “Devil or angel in the details? Perceiving phonetic variation as information about phonological structure,” in *The phonetics-phonology interface*, J. Romero and M. Riera, Eds. Amsterdam, The Netherlands: John Benjamins, 2015, pp. 3-31.

[6] C.T. Best & M.D. Tyler, “Nonnative and second-language speech perception: Commonalities and complementarities,” in *Seconds language learning: The role of language experience in speech perception and production*, M.J. Munro & O.S. Bohn, Eds. Amsterdam, The Netherlands: John Benjamins, 2007, pp. 13-34.

[7] C.T. Best *et al.*, “From Newcastle MOUTH to Aussie ears: Australians’ perceptual assimilation and adaptation for Newcastle UK vowels,” *Interspeech*, Dresden, Germany, 2015.

[8] J.W. Mullennix, D.B. Pisoni, & C.S. Martin, “Some effects of talker variability on spoken word recognition,” *J. Acoust. Soc. Amer.*, vol. 85, no. 1, pp. 365-378, 1989.

[9] L.L. Bonatti, PsyScope X (Build 57). Trieste, Italy: SISSA, 2010.

[10] M.A. Betz & J.R. Levin, “Coherent analysis-of-variance hypothesis-testing strategies: A general model,” *J. Edu. Stat.*, vol. 7, no. 3, pp. 193-206, 1982.

[11] B. Kriengwatana *et al.*, “Speaker and accent variation are handled differently: Evidence in native and non-native listeners,” *PLoS ONE*, vol. 11, no. 6, 2015.

[12] D.B. Pisoni *et al.*, “Speech perception, word recognition and the structure of the lexicon,” *Speech Comm.*, vol. 4, no. 1-3, pp. 75-95, 1985.