Interspeech 2007, Aug 27-31, Antwerp Belgium,
International Speech Communication Association.

1486

Effects of Non-native Dialects on Spoken Word Recognition

Jennifer T. Le [1], *Catherine T. Best* [1,2], *Michael D. Tyler* [1], *Christian Kroos* [1]

[1] MARCS Auditory Labs, University of Western Sydney, Milperra, Australia.
[2] Haskins Laboratories, New Haven, CT, U.S.A.

j.le@uws.edu.au, c.best@uws.edu.au, m.tyler@uws.edu.au, c.kroos@uws.edu.au

Abstract

The present study examined the premise that lexical information (top-down factors) interacts with phonetic detail (bottom-up, episodic traces) by assessing the impact of dialect variation and word frequency on spoken word recognition. Words were either spoken in the listeners' native dialect (Australian English: AU), or in one of two non-native English dialects differing in phonetic similarity to Australian: South African (SA: more similar) and Jamaican Mesolect (JA: less similar). It was predicted that low-frequency English words spoken in non-native dialects, especially the less similar dialect, would require more information to be recognised due to systematic phonological and/or phonetic differences from native-dialect versions. A gating task revealed that more gates were required for JA than SA dialect words, with this effect even more pronounced for low than high-frequency words. This suggests that recognition of words is contingent upon both detailed phonetic properties within the mental lexicon, as evident in the effects of goodness of fit between native and non-native dialect pronunciations, and on lexical information.

Index Terms: spoken word recognition, dialect variation, word frequency, forward gating, phonetic differences

1. Introduction

Spoken word recognition entails accurately selecting both the form and meaning of an uttered word from among thousands of candidates in the mental lexicon, which is reflected in the architecture of spoken word recognition models (e.g., [1, 2, 3]). Differing on minor details, all agree that as a given spoken word unfolds, words that start with the same phonemes become partially active (i.e., hearing the /bi/ of *bean* triggers activation of phonologically related words like *bead*) [4]. Speech input therefore automatically activates all words with the same onset, and word recognition occurs via a process of competition among activated candidates. *Word recognition* will be used here to refer to the end-point of the selection phase, when a lexical entry has been determined from the speech input [5]. It is well known that phonological and/or phonetic aspects of native speech guide listeners in their recognition of unfamiliar words [6]. The present study assessed whether and how non-native dialects impact spoken word recognition as a consequence of word pronunciation differences between native and non-native dialects.

1.1. Spoken Word Recognition Models

Contemporary theoretical models posit that word recognition depends on complementary information from the input (bottom-up), which is excitatory, and from word representations in the lexicon (top-down), which involves inhibition of lexical competitors. Thus, the more a spoken word deviates from its representation in the listener's mental lexicon, the more difficult it should be to identify. Spoken word recognition models vary in how well they tolerate initial phonological mismatches in the sensory input. Early versions of the Cohort model [7] showed extreme intolerance to initial mismatch in the speech input via bottom-up inhibition and total lexical deactivation. Alternatively, the TRACE model [2] could account for successful word recognition (even with distorted input e.g., "barty" as "party") by accommodating to minor initial mismatches [8]. Other models such as Shortlist [3] and more recent versions of the Cohort model [1] allow for graded activation that depends upon phonological distance, thus assigning weaker deactivating influences to mismatching information [9].

According to episodic theories (e.g., [10]), the lexicon contains multiple detailed traces of spoken words the listener has previously encountered (e.g., acoustic details specific to the way a given speaker talks). The central question, however, is whether these episodic representations constitute the basis of the mental lexicon or whether they should be considered simply an accessory to representations that are primarily abstract in nature. In episodic models, word recognition entails a comparison of the current input in all its phonetic detail with previously stored lexical episodes. On the other hand, in abstractionist models, word recognition is mediated by phonologically abstract lexical representations. The speech input is mapped onto abstract phonological representations such as phonemes (Shortlist), features (Cohort), or both phonemes and features (TRACE). Therefore extreme abstractionist and episodic models lie at opposite ends of the continuum of possible models of spoken word recognition. The key contrast between these two types of models is whether there is a phonologically abstract representation of the speech signal prior to lexical access [11]. Therefore, a central issue in research on spoken word recognition is the *degree* of phonetic detail represented in the mental lexicon.

1.2. Potential dialect effects on word recognition

An effective manipulation in the investigation of phonological mapping processes of spoken word recognition in isolation is phonological mismatch [8]. Phonological properties involve structural transformation, whereby a difference in a critical feature of a consonant or vowel in one dialect categorically transforms one word into another (or a nonword) in a different dialect. Phonetic details entail structural invariance, whereby a word's identity is preserved despite structurally-irrelevant though sometimes striking surface variation from one dialect to another [12]. In the present study, we probed the role of both phonological and sub-categorical phonetic mismatch on word recognition, using a novel approach: dialect variations in the pronunciation of English words. Such variations offer naturally-occurring phonological and phonetic mismatches to listeners' lexical representations in their native dialect.

When encountering a talker with an unfamiliar regional dialect, certain phonemes produced by the talker may not map directly onto the listener's existing phonemic categories.

Some dialects of the native language are so phonologically and/or phonetically dissimilar, however, that they can only be understood with considerable experience (e.g., speakers of British English require experience with American English in order to learn that an alveolar flap [ɾ] in intervocalic position is an instance of the phoneme /t/) [13]. Vowel differences are likely to outnumber consonant differences across dialects in the English language [14]. For example, the three-way distinction (in most English varieties) among *look, luck* and *Luke* collapses to two in other varieties (e.g., *luck* vs. *look/Luke* in Scottish English, and *Luke* vs. *look/luck* in Yorkshire English). Perceptual confusions across dialects generally occur when two categories that are distinct in the input are merged in the listener's dialect. As a result, certain words when pronounced in another dialect may be "misperceived" as other phonetically similar words in the native dialect due to the merger or overlapping of some vowels and/or consonants in native phonemic categories.

In particular, the conflicts that can arise in different-dialect pronunciations of a given word are potentially quite useful for testing the level of information that is stored in the representation of a word in a listener's mental lexicon. Is it represented only in its very abstract phonological structure? If so, then presumably people should recognise words in other, even unfamiliar dialects quite well because it must be the case that word's most abstract phonological form has to be comparable across dialects or we would not be able to converse with people from other dialects of our own language.

However, if people's lexical representations of words incorporate lower-level phonological details, or even richer and more detailed phonetic properties, then their recognition accuracy and speed for "foreign"-dialect words *should* be notably affected. The more lower-level the sorts of details that are included in the lexical representation of a word, the more recognition should be influenced by degree of phonetic (and/or lower-level phonological) *similarity* between the native and non-native dialect pronunciation of words. If this is the case, then for two dialects with differing degrees of phonetic distance from the native dialect word pronunciation, recognition accuracy and speed should be more hindered for the more dissimilar dialect than for the less dissimilar one, even if the unfamiliarity of the two non-native dialects is comparable.

1.3. Present study

The present study examined the impact of dialect variation on naïve listeners' identification of spoken words. To achieve accurate performance in an unfamiliar dialect, listeners must be able to somehow recognise the relationship between the input phonetic details and their native-dialect lexical representations. The gating paradigm was used here as it allows precise control over the amount of speech input on which a response is based [15]. If the acceptance point (AP: the gate at which the target word's pronunciation was identified correctly without any change in response thereafter) is indicative of the moment at which a word becomes recognised, then it should be sensitive to top-down factors influencing word recognition, such as word frequency. Indeed, low-frequency words have later APs than high-frequency words. Furthermore, long words (bisyllabic, polysyllabic) provide more bottom-up evidence than short words (monosyllabic), and short words are subject to greater inhibition due to the existence of more similar words [16]. These word frequency and word length effects reflect top-down (lexical) influences. Complementary influences from bottom-up, phonetic details (episodic traces) should be most evident in *low-frequency* words, because the overall lower

amount of exposure to those words over the listener's lifetime should result in less well-established abstract phonological forms in the lexicon. Bottom up details would be more beneficial for mono- than bi-syllabic words because the former have a greater number of lexical competitors.

The present study compared native adult speakers of Australian English in their perception of high versus low-frequency and mono- versus bi-syllabic words as spoken in Australian English (AU) versus two non-native dialects, South African English (SA) and Jamaican Mesolect English (JA). The two dialects were chosen to be fairly equally unfamiliar to most Australians tested (in Sydney's south suburban region), but at the same time to differ in their degree of phonetic and phonological similarity to AU. Words were selected such that JA differed notably from AU pronunciation in vowels, consonants, and prosody (e.g., JA has a merger of AU low vowels; /æ, a, ɒ/ which become /a/, creating homophones like *black-block*; intervocalic /t/ becomes /k/, *little = likkle*, /θ/ becomes /t/, *thing = ting*; occasional metathesis, *ask = aks*; deletion of word-initial /h/, *hear = ear*, and /s/, *stomach = tomak*, and certain polysyllabic words take word-final stress, *reaLISE, celeBRATE*) [17], while SA words bore more similarity to AU on all these dimensions (the most noticeable difference in pronunciation is the centralisation of /ɪ/ [17]). It was expected that the greater the pronunciation differences between dialects, the more difficult it would be for adults to recognise words due to differences in dialect phonological structure and/or phonetic details: this effect should be magnified for low-frequency, monosyllabic words.

2. Method

2.1. Participants

L1 Australian English first-year Psychology students at the University of Western Sydney participated in the study for course credit (10 males and 11 females, mean age = 22.2, range = 18-39). Data from two additional participants were excluded due to language, speech or hearing impairments. All participants were screened to insure minimal experience with the selected non-Australian dialects.

2.2. Stimulus materials

Eighteen high-frequency (9 monosyllabic, 9 bisyllabic), and 18 low-frequency (9 monosyllabic, 9 bisyllabic), English content words selected from the CELEX database were produced by three male speakers: AU, an SA English-Afrikaans bilingual and a JA Patois bilingual. They were selected for similar voice quality and fundamental frequency range. Speech stimuli were recorded in a sound-attenuated room at a sampling rate of 22 kHz and 16-bit resolution. Each word was recorded eight times in random order and the best versions of each were selected, excised, and trimmed using Cool Edit. Word duration, mean pitch and contour were calculated using Praat [18], and used to select the best matches of each word across the three dialects. The durations of matched identical words from each dialect were equalised and resynthesised with Praat's PSOLA algorithm.

2.3. Gating task

In forward gating tasks, a spoken stimulus is presented in segments of increasing duration from word onset over a series of trials, and participants are asked to identify the target after each gate [15]. A successive, blocked by duration variant of the gating task was used to avoid response perseveration and negative feedback [19]. Gated versions of the target words

were prepared using Praat [18]. Because items differed in length (monosyllabic vs. bisyllabic), the total number of gates per word ranged from 9 to 17. The first gate for each word was 60 ms from word onset. Each subsequent gate added another 60 ms from word onset (e.g., gate 2 = 120 ms; gate 3 = 180 ms etc.), until the complete word was presented. To avoid click artifacts resulting from an abrupt cut-off of the signal, the end of each gate was faded out over 10 ms using a cosine function. Each block contained 36 randomised stimuli corresponding to the same gate number for each item. Successive blocks corresponded to increasingly longer gates.

2.4. Procedure and apparatus

All participants were tested individually. Six practice trials using a talker and stimuli unrelated to the experiment were followed by the test, lasting approximately 30-45 minutes. Participants received, in a single session, 412 gates blocked by duration. Each listener was assigned to one of three counterbalanced groups, in which all 36 words were presented in all three dialects (12 from each), without a given listener hearing the same word twice. Alvin software [20] controlled stimulus presentation and response collection. The experiment was self-paced, and auditory stimuli were presented binaurally through AKG K271 closed-ear headphones at a comfortable loudness of 79.8 dB SPL.

3.   Results

Two dependent variables were analysed. One was the AP, expressed as a percent through the word (i.e., AP gate/total gate number x 100). AP can be used as a predictor for the amount of information needed and sufficient for word recognition. For target words that were never correctly identified, the AP was entered as the total number of gates plus one more (i.e., AP > 100%). To test whether recognition among listeners differed as a function of our dependent variables, participants' mean AP scores were submitted to a three-way repeated measures analysis of variance (ANOVA) for word frequency, syllable, and dialect.

The main effect for syllable was not significant, indicating that these word length differences did not affect the word's AP. However the main effect for word frequency was significant, $F(1, 20) = 255.11$, $p < .001$, $\eta_p^2 = .93$, indicating that, overall, high-frequency words were recognised with less information than low-frequency words. The main effect for dialect was also significant, $F(2, 40) = 141.38$, $p < .001$, $\eta_p^2 = .88$. The word frequency by syllable interaction was also significant, $F(1, 20) = 10.10$, $p = .005$, $\eta_p^2 = .34$, as was the word frequency by dialect interaction, $F(2, 40) = 14.99$, $p < .001$, $\eta_p^2 = .43$, and the word frequency by syllable by dialect three way interaction, $F(2, 40) = 6.48$, $p = .004$, $\eta_p^2 = .25$ (see Figure 1). No other interactions were significant.
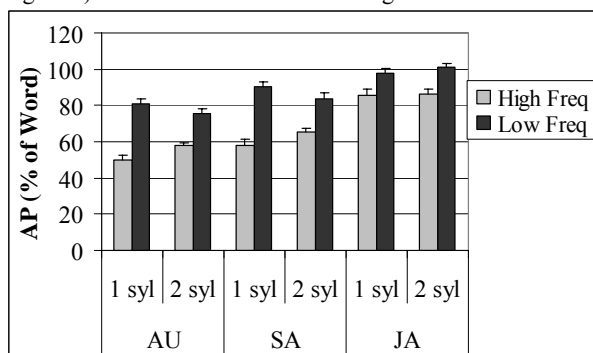

Figure 1: *Mean AP scores for the interaction of Dialect x Syllable x Word Frequency. Error bars represent standard errors of the mean.*

To determine whether the expected order of dialect difficulty was as predicted, three dependent-samples *t*-tests were conducted on all dialect pairs using a Bonferroni adjusted alpha of .02, collapsed across syllable and word frequency. The results indicated significant differences for all three dialect comparisons: listeners performed progressively poorer as the stimulus words differed from the native dialect (see Table 1).

Table 1. *Paired samples t-tests for AU, SA, and JA AP scores*

Pairs	*M*	*SD*	*t*	*p*
AU/SA	14.29	11.83	5.54	.000
SA/JA	36.91	17.59	9.61	.000
AU/JA	51.19	12.99	18.06	.000

The second dependent variable was the mean number of correctly identified items, calculated as percentages. Scores were analysed using a three-way repeated measures ANOVA with factors of word frequency, syllable, and dialect. The main effect for word frequency was significant, $F(1, 20) = 65.93$, $p < .001$, $\eta_p^2 = .77$, indicating that, overall, high-frequency words were more accurately recognised than low-frequency words. The main effect for dialect was also significant, $F(2, 40) = 139.27$, $p < .001$, $\eta_p^2 = .88$. The word frequency by syllable interaction was also significant, $F(1, 20) = 4.61$, $p = .044$, $\eta_p^2 = .19$, as was the word frequency by dialect interaction, $F(2, 40) = 6.51$, $p = .004$, $\eta_p^2 = .25$, and the word frequency by syllable by dialect three way interaction, $F(2, 40) = 5.52$, $p = .008$, $\eta_p^2 = .22$ (see Figure 2). No other interactions were significant.
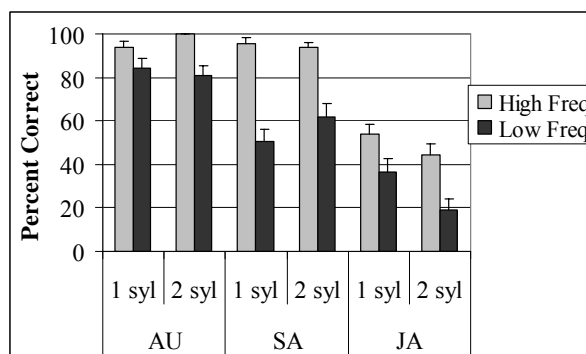

Figure 2: *Mean percent correct scores for the interaction of Dialect x Syllable x Word Frequency. Error bars represent standard errors of the mean.*

Dialect order of difficulty was tested again using three dependent samples *t*-tests, with a Bonferroni adjustment, on the mean percentage identification scores collapsed across syllable and word frequency. The results indicated significant differences for all pairs (see Table 2). Again, this suggests that participants performed worse when words were spoken in non-native dialects, especially in the dissimilar dialect, than in the native dialect.

Table 2. *Paired samples t-tests for AU, SA, and JA percentage correct scores*

Pairs	*M*	*SD*	*t*	*P*
AU/SA	14.37	12.02	5.48	.000
SA/JA	36.89	17.64	9.58	.000
AU/JA	51.25	13.62	17.25	.000

4.   Discussion

As hypothesised, low-frequency words as spoken in the more dissimilar non-native dialect were most difficult for participants to recognise. The gating task showed that fewer

JA words were recognised than AU or SA words, and for those that were identified, recognition occurred later in time (required longer gates). This was more the case for low-frequency target words. We chose our dialect comparisons to tease apart the familiarity factor from the similarity factor: Both SA and JA are equally unfamiliar dialects to our south-Sydney region AU listeners, but SA is more phonetically similar to AU than is JA. Thus, dialect order of difficulty indicates that the phonological and phonetic distance of an unfamiliar dialect from an individual's native dialect more significantly impacts upon the word recognition process than does mere lack of familiarity. We did not find a simple overall advantage for the bisyllabic words as predicted; instead a 3-way interaction was found that indicated earliest recognition for high-frequency monosyllabic words in the native dialect. Our set of high-frequency monosyllabic words were among the earliest ones learned in life, and were therefore ones with which listeners would have had the most extensive experience. This advantage in experience may have aided identification at early gates.

The number of erroneous word candidates that listeners proposed after the full presentation of a word can provide us with important information about word recognition, because they allow us to track the paths followed by individual listeners in the process of narrowing down various candidates to arrive at a single word. The reason why some words in the Jamaican dialect were never identified correctly might be attributable to the nature of JA versus AU pronunciation differences. Sometimes the pronunciation of words can be so phonologically different from the native dialect that their vowels or consonants may be "misperceived" as other vowels/consonants in the native dialect (e.g., JA "bear" was often misperceived by listeners as AU "beer": different native-dialect vowel; similarly for JA "bottle", which listeners often reported as AU "buckle": different native-dialect vowel and consonant). These findings suggest that sometimes the phonemic organisation of non-native vowels/consonants may overlap and/or be misperceived as other, contrasting vowels/consonants of the native dialect.

The results of this study appear to support the notion that listeners are sensitive to detailed phonetic as well as abstract phonological aspects of spoken words, and use such cues to recognise lexical candidates from different dialects. What we have shown here is that abstraction makes word recognition more efficient when one encounters a talker speaking in an unfamiliar way. The fact that low-frequency words were more difficult to recognise than high-frequency words, with predicted order of dialect difficulty supported, suggests that the physical characteristics of words (i.e., phonological/phonetic similarities) may be more important than familiarity of non-native dialects, and that correct identification of spoken words depends on the exact, familiar phonetic details and phonological structure of native-dialect words. These arguments, and our results, suggest that an adequate model of spoken word recognition must include flexible and abstract representations as well as fine episodic traces.

5. Acknowledgements

6. References

[1] Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25,* 71-102.

[2] McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1-86.

[3] Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189-234.

[4] Yee, E., & Sedivy, J.C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1-14.

[5] Frauenfelder, U.H., & Tyler, L.K. (1987). The process of spoken word recognition: An introduction. *Cognition, 25,* 1-20.

[6] Best, C.T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed). *Speech perception and linguistic experience: Theoretical and methodological issue in cross-language speech research,* pp. 167-200. Timonium, MD: York Press.

[7] Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology, 10,* 29-63.

[8] Liu, Y., Shu, H., & Wei, J. (2006). Spoken word recognition in context: Evidence from Chinese ERP analyses. *Brain and Language, 96,* 37-48.

[9] Frauenfelder, U.H., Scholten, M., & Content, A. (2001). Bottom-up inhibition in lexical selection: Phonological mismatch effects in spoken word recognition. *Language and Cognitive Processes, 16,* 583-607.

[10] Goldinger, S.D. (1998). Echoes of echoes?: An episodic theory of lexical access. *Psychological Review, 105,* 251-279.

[11] McQueen, J.M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30,* 1113-1126.

[12] Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.). *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research*, pp. 167-200.

[13] Norris, D., McQueen, J.M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204-238.

[14] Cutler, A., Smits, R., & Cooper, N. (2005). Vowel perception: Effects of non-native language vs. non-native dialect. *Speech communication, 47,* 32-42.

[15] Grosjean, F. (1996). Gating. *Language and Cognitive Processes, 11,* 597-604.

[16] Pitt, M.A., & Samuel, A.G. (2006). Word length and lexical activation: Longer is better. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 1120-1135.

[17] Wells, J. (1982). *Accents of English* (Vol. 1). Cambridge: Cambridge University Press.

[18] Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9-10), 341-345.

[19] Hallé, P. A., Segui, J., Frauenfelder, U., & Meunier, C. (1998). Processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of Experimental Psychology: Human Perception and Performance, 2*, 592-608.

[20] Hillenbrand, J. M., & Gayvert, R. T. (2005). Open source software for experiment design and control. *Journal of Speech, Language, and Hearing Research, 48*, 45-60.