The Efficiency of Cross-dialectal Word Recognition

Annelie Tuinman [1,2], *Holger Mitterer* [1], *Anne Cutler* [1,2,3]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; [2] Radboud University Nijmegen, The Netherlands; [3] MARCS Auditory Laboratories, University of Western Sydney, Australia

annelie.tuinman@mpi.nl, holger.mitterer@mpi.nl, anne.cutler@mpi.nl

Abstract

Dialects of the same language can differ in the casual speech processes they allow; e.g., British English allows the insertion of [r] at word boundaries in sequences such as *saw ice*, while American English does not. In two speeded word recognition experiments, American listeners heard such British English sequences; in contrast to non-native listeners, they accurately perceived intended vowel-initial words even with intrusive [r]. Thus despite input mismatches, cross-dialectal word recognition benefits from the full power of native-language processing.

Index Terms: American English, British English, casual speech, [r]-insertion, word recognition.

1. Introduction

Speakers would find it rather hard if all utterances had to consist of a concatenation of citation-form pronunciations. Instead, numerous processes accommodate the sequence of sounds to suit the processes of articulation. Thus in citation form, *to* is pronounced with the vowel [u], *post* with a final [t], and *grand* with final [nd]; but even in an unspeeded utterance of *I gotta post my letter to Grandpa*, the final vowel of *gotta* is not [u] but schwa, the [t] of *post* is usually not pronounced, and the [nd] in *grand* has most likely become [m]. In addition, the intervocalic stop in *letter* and *gotta* is quite probably flapped or glottalised (depending on the speaker's dialect of English). Such processes make uttering a continuous stream of words easy for the speaker, and, importantly, they do not in general cause problems for native listeners [8, 9, 15, 16].

All languages exhibit such casual speech processes. But the processes are not an inevitable consequence of articulation, as speaking can be made easier in many ways. Some casual-speech forms are common across languages (e.g., [t] deletion as in *post my*, place of articulation assimilation as in *Grandpa*), but other forms are relatively rare. Cross-linguistic asymmetries are thus found, even to the extent that a single phoneme sequence may be assimilated in two different ways in two languages (e.g., a [tb] sequence is more likely to assimilate in place to become [pb] in English, but to assimilate in voice to become [db] in French [4]).

The existence of such cross-language asymmetries leads one to expect that, as with cross-language phoneme repertoire differences, second-language [L2] listening may be made more difficult by mismatch in casual speech processes. This is indeed so, and the parallel with phoneme repertoire effects is quite appropriate: When the L2 contains a process that is also found in the L1, it is easy for L2 listeners to deal with ([18, 19] for [t]-deletion in German and Dutch), but when L2 speech is subject to a process that is unknown in the L1, listeners cannot attribute it to the correct source ([18, 20], for Dutch listeners confronted with the cross-linguistically rare process of [r]-insertion in British English). In the latter case, the inability to detect the source can also lead to problems at the word recognition level [21], which can potentially cause a communication breakdown.

The case of [r]-insertion as investigated in [18] is particularly interesting. This process is dependent upon the phonology of the language, in that it emerges in the presence of non-rhoticity, i.e., the absence of [r] in post-vocalic position in citation-form pronunciations. British English is in general non-rhotic, as are some other dialects of English such as Australian, or the variety of American English spoken in Boston. Further, this process contrasts with the processes discussed above of deletion (e.g., of [t]) or transformation (e.g., assimilation to a contextual feature), in that it involves insertion of a phoneme which is not present in the citation form. In no dialect of English does [r] feature in the words *idea* or *saw*. In British English, also, there is no [r] in a citation form utterance of *dear* or *sore*. However, [r] can surface in the word boundary in such phrases as *idea of*, or *saw a*, where the boundary has a vowel on either side of it.

The sudden appearance of [r] with no obvious source certainly caused word recognition problems for Dutch L2 listeners to English. Dutch is a rhotic language, so post-vocalic [r] is pronounced; vowel-vowel sequences never trigger the insertion of an intrusive [r]. In [18], Dutch listeners could not use acoustic cues to distinguish intended from intrusive [r]; presented with a seven-step continuum from *saw ice* to *saw rice*, they responded *rice* even if the evidence for [r] was very weak. British English speakers, however, produced a typical categorical pattern of response to the same materials. Moreover, the Dutch listeners' responses were strongly influenced by the probability of *ice* versus *rice* in the sentence contexts; again, British native speakers were impervious to this, and based their responses on acoustic factors only.

Unsurprisingly, then, listeners from the same Dutch population who heard sentences containing ambiguous sequences such as *extra (r)ice* could not be sure which they had heard, with the result that both of the words became momentarily available to them [21]. The experiment that showed this used the cross-modal priming task, in which listeners hear spoken words or sentences, while watching a screen and deciding whether letter strings that appear are real words or not. The rationale of the task is based on the robust phenomenon known as repetition priming: hearing or reading a word for a second time is easier than it was the first time [22]. If a participant hears *My brother likes extra rice...* and then responds to the letters RICE on the screen, the YES response will typically be significantly faster than the same response to RICE following a control prime such as *My brother likes extra pages...* In [21], Dutch listeners' decisions for words like RICE were appropriately speeded by preceding matched primes such as *My brother likes extra rice...*, but were also inappropriately speeded by mismatching primes such as *My brother likes extra ice...*with an intrusive [r] before *ice*.

These results show that casual speech processes that are unfamiliar can mislead L2 listening in the same way as occurs with unfamiliar phoneme categories; spurious word candidates can be activated during speech comprehension.

Casual speech processes, however, can differ, as we saw, not just from language to language but also across varieties of the same language, and the [r]-insertion process is a case in point; it is found only in the non-rhotic English dialects. This raises the obvious question of whether not only L2 listeners, but also L1 listeners from a rhotic dialect, could experience significant spurious lexical activation as a consequence of hearing a speaker produce an inserted [r] where their native dialect would have no [r] sound – e.g., in *saw aces*, or *extra ice*. This is the question addressed in the present study, and we tackle it by presenting rhotic-dialect English listeners with the British English materials that demonstrated the L2 effects of spurious word activation in [21].

The results from previous cross-dialectal studies of spoken word recognition do not motivate a definitive prediction of the pattern of results, since earlier studies have rarely addressed either mismatch in casual speech processes or the activation of spurious lexical candidates. Numerous indications suggest, though, that cross-dialectal listening is in general quite robust.

Thus in the one study of casual speech processes that we are aware of, British listeners proved able to correctly interpret syntactic boundary cues in American-English speakers' pronunciation of intervocalic plosives (whereby the stop at the boundary between *eat* and *early* is flapped in *If you want to eat early, dinner is served*, but not in *If you want to eat, early dinner is served*), even though their own speech showed no such patterning [14]. Word recognition by Japanese listeners whose own accent did not include pitch accent variation likewise exhibited efficient use of the accentual cues to word identity in the speech of speakers of the standard Tokyo variety [13]. Further, speech-in-noise recognition is quite robust to dialect variation, both for meaningful input [1] and for phonemes in meaningless syllables [3]. When a dialect difference does cause processing difficulty, problems are short-lived and removed by minimal experience [6].

All this suggests that pronunciation differences between dialectal varieties can be compensated for by knowledge of the underlying common language. As argued in a recent review of L2 speech recognition in noise [7], extensive resources are available for native-language processing at all processing levels (phonetic, lexical, syntactic, semantic, pragmatic), and these resources allow recovery from the effects of noise-masking which in principle affect both L1 and L2 listeners similarly. On this explanation, dialect mismatch would, like noise, degrade the initial perception of spoken input to a certain extent, but native listeners should be well able to recover from its effects. If this is so, the rhotic-dialect users of English should be able to adapt to the speech in the other variety, and correctly identify the intended words.

Nonetheless, the [r]-insertion process could pose special difficulty for listeners, precisely because it involves insertion of a phoneme into an environment in which, in the listeners' variety, the same phoneme can occur, and where it infallibly signals a lexically canonical segment rather than the result of a casual speech process. In rhotic English, *saw + r + aces* can only be *saw races*; it is not ambiguous as it can be held to be for British English users. In this case, rhotic-variety listeners may be as misled as were the L2 listeners tested in [21].

The listeners tested in the present study were native users of a rhotic variety of American English, resident in a rhotic-variety-speaking area. Note that degree of exposure, a problem that in general bedevils research on dialect perception, is not possible to resolve definitively; individuals' media exposure, use of internet materials, etc. can never be determined, and no proficiency tests exist for cross-dialect listening as they do for cross-language listening. We will return to this issue below.

2. Experiments

Basing the design on [21], we conducted two experiments, with the comparison of vowel-initial versus [r]-initial targets for the critical pairs being between-subjects. This design was chosen to avoid drawing attention to the existence of potentially ambiguous sequences in the materials.

2.1. Participants

A total of 72 native speakers of American English took part, 36 in each of Experiments 1 and 2. All were undergraduates recruited from the participant pool of the Institute for Research in Cognitive Science at the University of Pennsylvania, with no hearing impairment and normal or corrected-to-normal vision. They received a small remuneration for taking part.

2.2. Materials

Twenty-seven pairs of English sentences were constructed, each based on a minimal pair of words (nouns, verbs or adjectives) that differed only in starting with a vowel or with [r], such as *ace/race*, *ice/rice*, *ejected/rejected*. In all sentences, the member of such a minimal pair followed a word ending in a non-high vowel, e.g, *saw*, *extra*, *Emma*. This vowel can then in each case be followed by [r]; trivially, the [r]-initial member of the pair will be uttered with [r], but more importantly, the vowel-initial member of the pair preceded by a non-high vowel creates the requisite context for an intrusive [r] to occur. As a result, potentially ambiguous sequences arise: examples are *saw (r)aces*, *extra (r)ice*, *Emma (r)ejected the cassette.*

A further 180 sentences were constructed, so that the total of sentences in each experiment was sufficiently large that the 27 experimental items should not stand out. These further sentences were: (a) 27 control sentences containing sequences matched for English frequency to those in the experimental sentences (e.g., *extra pages…*); these were used as control primes against which the experimental prime conditions could be compared; (b) 18 question trials, in which a complete sentence was followed not by a visual target but by a yes/no question; these served as a check on whether participants were paying attention to the prime sentences as well as to the visual target words; (c) 108 filler prime sentences, 54 with unrelated-word targets and 54 with nonword targets; (d) 27 filler prime sentences with nonword targets starting with either [r] or a vowel, included to prevent an association of phonological relatedness between prime and target with YES responses in the visual lexical decision task; these contained sequences with [r] such as *your explanation,* or *my neighbour refused,* or they contained no [r] at all, e.g., *the news about taxes.*

All sentences were recorded by a female native speaker of British English, from London, who was unaware of the study's purpose and normally produces intrusive [r] in casual speech. Each sentence was recorded at least twice with no disfluencies.

Measurements were made of the critical minimal pairs as produced by this speaker in the sentence contexts. As expected on the basis of the phonetic literature on this topic [2], onset [r] was always longer and had a larger intensity decrease than intrusive [r]. The mean duration for onset [r] was 89 ms, for intrusive [r] 69 ms; the mean intensity decrement was 7.9 dB and 2.2 dB respectively. In each case the effect size across the set of 27 items was high enough to be perceptually relevant (Cohen's $d > 1.5$).

The prime materials for the experiment were created by truncating the sentences directly after the (potential) prime word, so that participants heard no complete sentences, but always fragments such as *My brother likes extra rice…*

2.3. Procedure

Participants were tested one at a time in a sound-attenuated booth. They received instructions in English on a computer screen, informing them that on each trial they would hear a portion of an English sentence, directly after which an English word or nonword would appear on the screen. They were instructed to press a response button labeled YES with their dominant hand if they thought the visually presented item was an English word, and a button labeled NO with their other hand if they thought it was not an English word. Participants were asked to respond as rapidly as possible without errors.

Each participant received each visual target only once, with nine targets in each of the three prime conditions (match, mismatch, control). In Experiment 1, the targets were vowel-initial words (e.g., *ice*), the vowel primes matched the target and the r-primes mismatched. In Experiment 2, the target words were r-initial (e.g., *rice*), r-primes thus matched and vowel primes mismatched. In each control condition, targets were preceded by a phonologically and semantically unrelated prime. Besides the experimental trials, participants in each experiment received all filler word and all filler nonword trials, and all yes/no question trials. Each experiment began with seven practice trials and one practice question trial.

2.4. Results

Lexical decision response times (RTs) were measured from onset of visual presentation of the target words. Overall mean RT was 749 ms., and the baseline (unrelated control-prime) RT to vowel-initial words 748 ms, to [r]-initial words 783 ms. (Such a difference is expected for visually presented sets of words of differing length – in this case, always by one letter.)

Crucially, however, the predictions concern not the overall cross-experiment comparison, but cross-condition comparisons within each experiment: RTs in conditions with matching primes (*rice* preceded by *rice* in the sentence, *ice* preceded by ice) versus mismatching primes (*rice* preceded by *ice*, *ice* preceded by *rice*), relative to the control-condition RTs. Figure 1 shows the mean RTs for the correct responses in experimental trials of each experiment, expressed as amount of facilitation (i.e., RT differences when responses to targets preceded by a matching or mismatching experimental prime are subtracted from responses to the same targets preceded by an unrelated control prime).

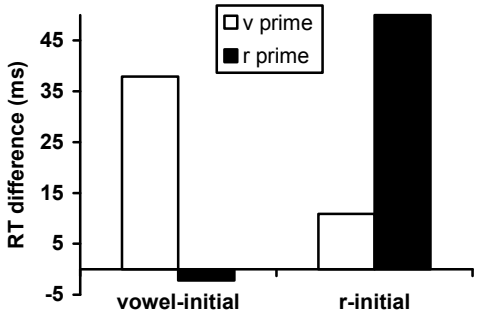

Figure 1: *Results of Experiments 1 and 2: Amount of priming (control condition RT minus primed condition RT, in ms) for decisions on vowel-initial target words (e.g.,* ice*, at left) versus [r]-initial target words (e.g.,* rice*, on the right), as a function of prime sentences with vowel-initial words preceded by intrusive [r] (clear bars), versus [r]-initial words (dark bars).*

It can be easily seen from the figure that these listeners showed significant, and, importantly, appropriate priming: when the target word on the screen was vowel-initial (*ice*, *aces*, *ejected*), there was strong priming from sentences containing that word (*extra ice*, *saw aces*, *Emma ejected*), despite the intrusive [r] that was always present in these cases. When the target was r-initial (*rice*, *races*, *rejected*), there was strong priming from sentences containing that matching word (*extra rice*, *saw races*, *Emma rejected*). In neither experiment was there any significant priming from a mismatching prime.

The RTs were log-transformed to reduce RT distribution skew, and analysed with a linear mixed-effects model with items and participants as simultaneous random factors. In Experiment 1 (vowel-initial targets) the main effect of prime type was significant ($F\ [2,757] = 4.39$, $p < .05$; F-values are based on model comparison using the anova function in R). Separate comparisons across prime conditions showed that matching primes led to significantly faster RTs than control primes ($p < .05$), but mismatching and control prime conditions did not significantly differ ($p > .1$). In Experiment 2 (r-initial targets), essentially the same pattern appeared: the effect of prime type was significant ($F\ [2, 793] = 5.99$, $p < .01$), matching primes produced significantly faster RTs than control primes ($p < .01$), but mismatching and control conditions did not differ ($p > .1$).

The regression-based analysis with a linear mixed-effects model also allows us to examine the nature of the effects across an experiment, by testing for an effect of trial number. Trial number was indeed significant, with faster responses later in the experiment than earlier ($p < 0.001$), but this effect did not differ between conditions ($p > 0.2$). That is, there was an overall increase in participants' lexical decision speed, but this was true even in the control condition, not, for instance, only with vowel-initial primes with intrusive [r]. This pattern was constant across both Experiment 1 (vowel-initial targets) and Experiment 2 (r-initial targets).

3. Discussion

The results of our study show robust priming by matching primes but no significant priming by mismatching primes. This finding contrasts dramatically with the results for the same materials presented to proficient L2 listeners [21], whose responses to r-initial targets showed that they were misled by mismatching vowel-initial primes pronounced with intrusive [r]. Clearly, a dialect mismatch is not equivalent to a language mismatch.

The results are in line with predictions based on previous studies of cross-dialectal recognition [1, 3, 13, 14], that broadly showed a robust ability by listeners to adjust to a mismatch in variety within the same language. They are not in line with the assumption that phoneme insertions would cause particular perceptual problems, differing from those caused by phoneme deletions or transformations.

Recent studies have shown that native listeners are in fact extremely skilled in perceptual learning about speech [12], and that such learning is speaker-specific [5] and hence serves to adapt listening efficiently to talker variation. This perceptual learning is, moreover, sensitive to likely dialectal features [10]. We suggest that there is no principled difference between adjusting to a new speaker with an unusual articulatory setting and adjusting to a speaker from another native-language variety; as long as the speech is in the native language, it is efficiently processed despite mismatches to the listener's own speech. The resources of native listening – in clear contrast to the difficulties of second-language listening – support very rapid learning about newly encountered input.

As noted above, it is impossible to control for degree of exposure to varying input. A newly encountered speaker may have a voice similar to other people known to one listener but be quite dissimilar to everyone ever heard by another listener. Listeners can never have been away from their home town but can still differ in their exposure to other dialects as a result of the teachers they have encountered, the vacation jobs they have held, the television they watch, and their taste in YouTube videos. We can be sure that our University of Pennsylvania undergraduate participants have all had some experience of British English, and possibly other non-rhotic English dialects; but we cannot establish how much for each individual. Our results showed, however, no significant within-group variation in the response patterns. Had the [r]-intrusion been completely unfamiliar to them, we might have expected adaptation across the experiment, in the form of increasingly faster responses to vowel-initial targets primed by vowel-initial words with intrusive [r]. While responses did indeed get faster across the experiment, this effect was unspecific and hence indicative only of a practice effect. Note that as few as 10 examples of an unfamiliar speech pattern provide sufficient evidence to support robust perceptual learning [11]; it could even be that exposure to the practice trials at the outset of an experiment is sufficient to re-set dialect perception parameters for native listeners.

The British English process of [r]-intrusion only occurs in a restricted phonetic environment, and as a result is not so frequent that it is constantly mimicked or derided; it may in fact rarely become obvious to listeners, either those who use it or those who do not, even though the presence versus absence of rhoticity in a dialect as a whole is highly noticeable [17]. The acoustic cues that differentiate intrusive from intended [r] are sufficiently marked to be perceptually useful – just as, indeed, is the case with other casual speech processes in which a phoneme is transformed or is effectively inserted. So has it been established that native speakers of French rely on phoneme duration to distinguish between a sound that has arisen by liaison and the same sound uttered as an intended word onset [15], and native speakers of English can likewise distinguish an intended velar or bilabial place of articulation from an assimilated coronal [9].

Given that such subtle acoustic differences are present in the speech of most languages, and thus are potentially used productively by most listeners, the next question for researchers must be to establish why native listeners can use this information, and apparently listeners from another variety of the same language can use it too, but, as [21] showed, non-native listeners, even with considerable proficiency in their L2, cannot. Between native listening, so highly efficient even when cross-dialectal, and listening in the non-native case, looms a gap that seems to be both large and very difficult to bridge.

4. Acknowledgements

This research was supported by a Max Planck Society doctoral fellowship and a Fulbright Foundation award, both to the first author. We thank Delphine Dahan, University of Pennsylvania, for enabling the testing of the participants.

5. References

[1] Clopper, C.G., and Bradlow, A.R., “Effects of dialect variation on speech intelligibility in noise”, J. Acoust. Soc. Amer., 119, 3424, 2006.

[2] Cruttenden, A., and Gimson, A. C., Gimson's pronunciation of English (5th ed.), London, Arnold, 1994.

[3] Cutler, A., Smits, R. and Cooper, N., “Vowel perception: Effects of non-native language versus nonnative dialect”, Speech Commun., 47, 32-42, 2005.

[4] Darcy, I., Peperkamp, S., and Dupoux, E., “Plasticity in compensation for phonological variation: the case of late second language learners”, In J. Cole and J. Hualde [Eds], Laboratory Phonology 9, Mouton de Gruyter, 2007.

[5] Eisner, F., and McQueen, J.M., “The specificity of perceptual learning in speech processing”, Perc. & Psychophys., 67, 224-238, 2005.

[6] Floccia, C., Goslin, J., Girard, F., and Konopczynski, G., “Does a regional accent perturb speech processing?”, J. Exp. Psychol.: Hum. Perc. and Perform., 32, 1276-1293, 2006.

[7] Garcia Lecumberri, M.L., Cooke, M., and Cutler, A., “Non-native speech perception in adverse conditions: A review”, Speech Commun., 52, 864-886, 2010.

[8] Gaskell, M.G., and Marslen-Wilson, W.D., “Mechanisms of phonological inference in speech perception”, J. Exp. Psychol.: Hum. Perc. and Perform., 24, 380-396, 1998.

[9] Gow, D.W., “Does English coronal place assimilation create lexical ambiguity?” J. Exp. Psychol.: Hum. Perc. and Perform., 28, 163-179, 2002.

[10] Kraljic, T., Brennan, S.E., and Samuel, A.G., “Accommodating variation: Dialects, idiolects, and speech processing”, Cognition, 107, 51-81, 2008.

[11] Kraljic, T., Samuel, A.G., and Brennan, S.E., “First impressions and last resorts: How listeners adjust to speaker variability”, Psychol. Sci., 19, 332-338, 2008.

[12] Norris, D., McQueen, J.M., and Cutler, A., “Perceptual learning in speech”, Cognit. Psychol., 47, 204-238, 2003.

[13] Otake, T. and Cutler, A., “Perception of suprasegmental structure in a nonnative dialect”, J. Phon., 27, 229-253, 1999.

[14] Scott, D.R. and Cutler, A., “Segmental phonology and the perception of syntactic structure”, J. Verb. Learn. Verb. Behav., 23, 450-466, 1984.

[15] Spinelli, E., McQueen, J.M., and Cutler, A., “Processing resyllabified words in French”, J. Mem. Lang., 48, 233-254, 2003.

[16] Sumner, M., and Samuel, A. G., “Perception and representation of regular variation: The case of final-/t/”, J. Mem. Lang., 52, 322–338, 2005.

[17] Sumner, M. and Samuel, A. G., “The effect of experience on the perception and representation of dialect variants”, J. Mem. Lang., 60, 487-501, 2009.

[18] Tuinman, A. and Cutler, A., “Casual speech processes: L1 knowledge and L2 speech perception”, Proc. 6th Int. Symp. on the Acquisition of Second Language Speech: New Sounds 2010, Poznan, Poland, 2010.

[19] Tuinman, A., and Mitterer, H., “Transfer of L1 knowledge helps in doubly adverse conditions: Perception of casual speech in L2”, under revision, Lang. and Cog. Proc, 2011.

[20] Tuinman, A., Mitterer, H. and Cutler, A., “Cross-language and cross-dialectal differences in perception of intrusive /r/ in English.”, under revision, J. Acoust. Soc. Amer., 2011.

[21] Tuinman, A., Mitterer, H. and Cutler, A., “Speakers differentiate English intrusive and onset /r/, but L2 listeners do not”, Proc. 16th Int. Cong. Phon. Sci., Saarbruecken, 1905-1908, 2007.

[22] Zwitserlood, P., “Form priming”, Lang. and Cog. Proc., 11, 589-596, 1996.