Title: Reassignment of the flap allophone in rapid dialect adaptation

James S. Germana*, Katy Carlsons, and Janet B. Pierrehumbertc

*Corresponding author: jsgerman@ntu.edu.sg, Tel: +65 6592 1822, Fax: +65 6795 6525

aNanyang Technological University, Division of Linguistics and Multilingual Studies, HSS 03-46, 14 Nanyang Drive, Singapore 637332
bMorehead State University, Department of English, 103 Combs, 150 University Boulevard, Morehead, KY 40351, USA
cNorthwestern University, Department of Linguistics, 2016 Sheridan Road, Evanston, IL 60208-4090, USA

Abstract
In an experiment spanning a week, American English speakers imitated a Glaswegian (Scottish) English speaker. The target sounds were /t/ and /r/, as the Glaswegian speaker aspirated word-medial /t/ but pronounced /r/ as a flap initially and medially. This experiment therefore explored whether speakers could learn to reassign a sound they already produce (the flap) to a different phoneme and to different phonetic contexts. Speakers appeared to learn systematically, as they could generalize to words which they had never heard the Glaswegian pronounce. There was a mix of categorical learning, with the allophone simply switching to a different use, and parametric approximations of the “new” sound. The phonetic context was clearly important, as flaps were produced less successfully when word-initial. And although there was variety in success rates, most speakers learned to produce a flap for /r/ at least some of the time and retained this learning over a week’s time. These effects are most easily explained in a hybrid of neo-generative and exemplar models of speech perception and production.

Keywords
Flap, allophone, dialect, rhotic, learning, generalization, lexicon
Reassignment of the flap allophone in rapid dialect adaptation

1. Introduction

A number of recent sociophonetic studies have provided evidence of dialect adaptation in adult speakers under natural conditions. Munro, Derwing, and Flege (1999) found that Canadian health care professionals who had taken up residency in Birmingham, Alabama partially acquired an American accent, as evidenced by perceptual ratings of extemporaneous speech samples. Harrington, Palethorpe, and Watson (2000a, 2000b), drawing on 40 years of recorded Christmas broadcasts of Queen Elizabeth II, used objective acoustic measures to show that Her Majesty’s pronunciation has been modified in some respects towards the rising Southern British (Estuary) patterns. A post-hoc study by Sankoff (2004) of a longitudinal set of recordings made for the British documentary series *Seven Up* also found dialect adaptation by two speakers who moved from Yorkshire and Liverpool to southern England and elsewhere as adults. In the wake of a substantial body of research associated with the critical period hypothesis, which raised the importance of questions related to late learning, such sociophonetic studies provide key evidence for late plasticity in the phonological and phonetic system. However, diagnostic evidence about the cognitive architecture responsible for this adaptation may be easier to find in controlled experiments.

In this paper, we report the results of a controlled experiment on dialect learning, which we believe to be the first of its kind. In a 20-minute (48-sentence) training session with no feedback, 24 Northwestern undergraduates undertook to imitate a novel dialect of English to the best of their ability. This dialect was Glaswegian English, though the subjects were not told this fact and in debriefing were unable to identify it. The target sounds of interest were /t/ and /r/, though the subjects were not told this fact either. For /t/, we were interested in the allophone that appears intervocally under falling stress (as in the word *pretty*). This is a flap in American English, but is aspirated in the speech sample of our Glaswegian speaker that was used to construct the experimental materials. The Glaswegian /r/, in contrast, was a rhotic flap in all positions. This allowed us to see whether American speakers could avoid flapping intervocalic /t/ and begin flapping intervocalic and initial /r/ instead. In the training phase, subjects heard each training sentence in Glaswegian English before reading it from a printed list. Though the orthographic representation ultimately complicates our interpretation of the results, we found it to be necessary because the speech was unintelligible without orthographic support. The training phase was immediately followed by a test for generalization to novel lexical items. Subjects also returned to the lab a week later, where they were tested for retention of the Glaswegian pattern. The retention testing had three components: the original training set, the original generalization set, and a new generalization set.

Our goal in designing this experiment was to address four related issues. These issues are suggested by prior work on second language learning and on learning of individual speaker traits. Specifically:

1) Lexical versus systematic learning: To what extent do subjects learn general phonological or phonetic patterns, which transfer from words in the training materials to new words?

2) Categorical versus parametric learning: To what extent do learners succeed by exploiting phone categories which they already know from their L1 (or D1, native dialect)? To what extent do they succeed by forming new categories over the parametric phonetic space?

3) Persistent vs. short-term learning: To the extent that speakers learn general phonological or phonetic patterns, do the effects persist over a period of one week?

4) Positional constraints: If existing categories can be reused at all, are they confined to their original D1 context? Or can they be reused in a different context? In short, can D1 positional variants, or allophones, be promoted to independent contrastive status in D2?

The literature on second language (L2) learning has emphasized systematic phonological and phonetic learning. Dialect learning (D2 learning) resembles L2 learning in that
it involves competition between the L1 or D1 phonological system and the novel system. Prior results in L2 learning indicate that a speaker’s success in learning an L2 speech segment depends on the exact nature of its relationship to segments in the L1 inventory. Research on this issue, going back several decades, has been dominated by examination of the L2 phoneme inventory compared to that of L1. Two of the most widely known models, Best’s Perceptual Assimilation Model (Best, McRoberts, & Goodell, 2001) and Flege’s Speech Learning Model (1995), share a number of key assumptions about how the L1 phoneme inventory comes into play during L2 exposure. If an L2 phoneme is phonetically equivalent to an L1 phoneme, it will be processed using the L1 code, hence it will be successfully perceived and produced. If it is phonetically similar to an L1 phoneme, but not equivalent, strong interference is expected: the L2 sound is perceptually assimilated to the L1 phoneme, and hence it is very difficult for the learner to improve beyond his/her initial rapid but partial success with the phoneme. If it is extremely distinct from all L1 phonemes (as Zulu clicks are for English speakers), then there is much less interference, and the phoneme is a candidate for the kind of parametric learning involved in new category formation. The degree of success by adults in such learning would be indicative of the nature of phonetic plasticity that persists into adulthood.

As discussed in Strange (1995), the experimental paradigms employed in work on acquisition of L2 phoneme inventories generally explore only a particular positional variant of the target phonemes (for example: a novel consonant contrast in stressed, word-initial position). It is unclear whether the cognitive units involved are phonemes in the classical sense (which retain their identity across variations in context), or less abstract, allophonic, units. She reviews results on the acquisition of the /r/-/l/ distinction by Japanese learners of English (Mochizuki, 1981; Logan, Lively, & Pisoni, 1991). Though this contrast is generally difficult for Japanese learners, it is much more difficult in some contexts than others, indicating that allophonic units are probably the relevant level of description. Whalen, Best, and Irwin (1997) studied the [p] vs. [pʰ] allophones of English and found that speakers could imitate these sub-phonemic differences even if they could not reliably distinguish them. Polka (1991) explored in detail the ability of English listeners to perceive the dental-retroflex distinction in Hindi as a function of the laryngeal features of the target phoneme pair. Hindi has a four-way distinction amongst plain voiced, plain unvoiced, breathy-voiced, and aspirated-unvoiced stops, which combines with a dental-retroflex place distinction to yield a total of eight distinct obstruent coronal stops. Because English /t/ and /d/ have dentalized contextual variants (as in “put that” or “good throw”), different consonant pairs in the set of eight Hindi consonants were predicted to have different degrees of support from the English allophonic system. Polka did find that perception of the dental-retroflex contrast was significantly above chance for three of the four consonant pairs, suggesting that the English phonetic system in some way supports perception of this contrast. Her specific hypotheses about the Hindi pairs were not supported, possibly because she relied on a model of English allophony which did not, apparently, encompass the stop allophones of [ɻ] and [ɻ̚] observed in emphatic productions and in many dialects (stereotypically, “dat” for “that”). The study we report here supports a more secure interpretation of the allophonic variation, because we obtained baseline recordings in order to estimate the D1 allophonic statistics on a subject-by-subject basis.

Two studies by psycholinguists use artificial language learning tasks to explore the malleability of the coding system in perception. Maye, Aslin, and Tanenhaus (2008) used a speech synthesizer to create an artificial dialect of English with categorical lowering of target vowels. For example, the substitution of [ɛ] into the word witch yields the form wetch, a non-word in the base dialect. Maye et al. found that subjects exposed to the novel dialect significantly increased their endorsement of modified forms as words in a lexical decision task. The effect generalized to new words. It did not generalize to vowel substitutions that were not included in the training phase. Since endorsement of unmodified words was not reduced, the results point to an architecture in which the relation of the phonological code to the lexicon can be systematically augmented in response to novel speech patterns. Parametric learning is not implicated in Maye et al.’s experiment, though, because the stimulus materials were created by categorical substitution of phonemes. Peperkamp and Dupoux (2007) also investigate the
malleability of the system at the categorical level. They use an artificial language learning paradigm to explore a categorical feature neutralization for certain consonants. In their materials, voicing was contextually predictable for stops but not for fricatives, or vice versa in the counterbalancing condition. Their series of experiments also manipulated the degree of semantic support for learning the phonological patterns of the language. Subjects were tested using a picture-pointing task. When word-learning was semantically supported, learning of the phonological constraint was very efficient and generalized to new words.

Results such as those of Maye et al. and Peperkamp and Dupoux suggest a neo-generative architecture, similar to Figure 1. The production system, following the broad lines of Levelt (1980), is involved in retrieving word forms from the lexicon, assembling the phonological code for the word forms in their phrasal context, and computing the phonetic implementation of the assembled phonological representation. The perception side is portrayed as analogous in the figure; the acoustic phonetic signal is phonologically parsed, and the phonological parse serves to access the lexicon. Systematic effects of the type that Maye et al. and Peperkamp and Dupoux have demonstrated do not involve any modification of the units in the coding level; the adaptation resides in the relationship of these units to the lexicon, with Maye et al.’s experiment involving the subjects’ existing lexica, and Peperkamp and Dupoux’s experiment involving novel lexical items in a putatively novel language. Comparing Maye et al.’s and Peperkamp and Dupoux’s results to Polka’s tends to support Strange’s suggestion that the relevant units at the coding level may be positional variants of phonemes rather than phonemes themselves. On this interpretation, Peperkamp and Dupoux’s subjects succeeded in learning equivalence classes of these positional variants.

Regardless of the role of orthography, a neo-generative architecture, such as Figure 1, readily captures categorical, across-the-board effects. That is, if the phonological coding level is systematically modified in production by any means, then this modification will be reflected in the phonetic realizations of all words. No words—whether in the training set or not, whether frequent or rare—will have any privileged status with respect to the new coding pattern. If the coding system is modified in perception, it will likewise affect all words equally. The architecture is also consistent with certain word-by-word effects. We all know that some words have more than one pronunciation. If the subjects in an experiment simply memorized the new

![Figure 1](image_url)
pronunciations for the words in the training materials as categorical alternatives, then the model would capture this fact by listing multiple word-forms for these specific words in the lexicon. A mixed situation, in which words used in training materials show an effect most reliably, but the effect also generalizes to new forms, can also be described by assuming that subjects both remember examples and update their coding systems through statistical generalizations over known examples, as suggested in Pierrehumbert (2003). If we assume Bayesian updating (e.g., modifying prior probabilities in the light of new statistical evidence), then the grammar statistics will lag the lexical statistics until the learning is complete. This is exactly what Maye et al. (2008) and Peperkamp and Dupoux (2007) report. Given the brief training and variable outcomes in these studies, the claim that the experiments ended before the learning was complete is plainly justified.

A different architecture has been proposed by many researchers working on voice recognition and social identity, such as Goldinger (1998) and Johnson (2006). Dialect recognition is similar to voice recognition, because an idiolect can be viewed as a one-person dialect. Recognizing a dialect means recognizing something about the speaker’s social identity, just as recognizing gender or sexual orientation does. Learning to produce a dialect means learning to project a particular social identity, and modern sociophonetic theory indeed explores dialect learning in the context of social identity construction (Mendoza-Denton, Hay, & Jannedy, 2003). Experiments on speech processing in relation to individual speakers and social identity have revealed some surprising interactions, which are problematic for the neo-generative architecture in its basic form. Such effects include: Shifts of category boundaries as a function of gender and gender typicality (Johnson, 2006); effects of speaker identity on word recall (Goldinger, 1996; Goldinger, Pisoni, & Logan, 1991; Palmeri, Goldinger, & Pisoni, 1993; inter alia); effects of speaker identity on novel word recognition (Nygaard, Sommers, & Pisoni, 1994); and unconscious imitation effects, which are more significant for low frequency words than for high frequency words (Goldinger, 1998).

The discovery of such effects has fueled the rise of exemplar-based models of speech perception. These models presuppose that experiences of speech are stored in memory in very considerable detail. Each memory can be indexed in multiple ways; for example, a memory of the utterance [beɪbi] can be indexed as an example of the word baby, as an example of my mom’s speech, and as an example of a female voice. In the simplest exemplar models (e.g., Hintzman’s (1986) MINERVA, Johnson’s (1997) XMOD), phonological structure emerges epiphenomenally from the similarity space defined by the remembered experiences. Exemplar models readily capture interactions between social variables and lexical access. They also capture findings of novel patterns being more robustly applied to the exact materials used in the training than to novel materials.

The simplest exemplar models, such as MINERVA and XMOD, encounter difficulties in explaining the extreme reliability of lexical access under changes in speech rate or prosodic position. If lexical access is attempted from the parametric representations of entire words, the alignment of the speech signal with the stored representations can be problematic. Reduction of a segment or syllable early in the word can induce misalignment of the entire remainder of the word with the stored representations. This can lead to a poor match even if phonologically aligning word subparts in the optimal way would have yielded a very good match. This problem is noticeable in calculations using XMOD presented in Baker (2004). Clearly, this same problem is compounded when word recognition in connected speech is considered. A further issue for exemplar models is the mechanism for speech production. Pierrehumbert (2001) takes as a point of departure the idea that production targets are picked by random selection of the exemplar space for the word. Goldinger (1998), taking a position reminiscent of direct realists, proposes that the combined effect of all exemplars activated by a lexical choice creates a production plan. But both positions are regrettably vague about how novel words can be produced. Productions of novel words do not average the properties of all similar real words. If they did, [bɹɑɡ] would average bog, blog, frog, broad, brought, and so forth, leading to a hybridized sonorant in the onset and a hybridized obstruent in final position. Instead,
productions of [bɹɑg] begin with the [bɹ] of *brought or broad*, and end as in *frog*.

Such issues have led to the development of hybridized models, with some such models already reviewed in Goldinger (1998). Pierrehumbert (2002) adopts the neo-generative claim that production of all words involves programming a categorical phonological representation, and that executing this plan is the only way to produce speech. (See Levelt, 1980). This means that lexical representations of individual words include both a phonological parse, needed to compute alignment and sequencing in speech processing, and a phonetic trace, needed to capture the individual speaker and sociostylistic effects which led to the original rise of exemplar models. She suggests that specific words or social situations can influence phonetic realizations by biasing the selection of phonetic exemplars used as realization targets for phonological plans. Since these biases are within-category, they are expected to yield secondary effects. This hybrid model, then, supports four different mechanisms altogether for imitating a new accent:

1) Learning new phonetic categories. Such learning is predicted to be possible only after high levels of exposure and practice, and is subject to strong interference from existing categories.
2) Learning situationally-appropriate biases within existing categories.
3) Learning alternative pronunciations for known words, encoded using existing categories.
4) Learning generalizations about these alternative pronunciations, encoded as generalizations about phonological representation.

A range of recent findings support such a hybrid view more directly. Several studies (surveyed in Cutler, Eisner, McQueen, & Norris, 2010) have found that listeners adjust the position of phonemic boundaries between sounds after short exposures to speech which uses ambiguous sounds for one end of a continuum. For example, after hearing words that usually end in /ʃ/ pronounced with a sound in between /ʃ/ and /s/, listeners will accept more s-like sounds as /ʃ/ than they otherwise would. Crucially, most research suggests this is talker-specific, so if a different speaker produces the target sounds than produced the words, the perceptual boundary is not shifted. Kraljic and Samuel (2006) did show transfer across talkers and sounds for stop perception, however. Kraljic, Brennan, and Samuel (2008) showed that a sound shift (on an [s]-[ʃ] continuum) which is restricted to a particular phonological context did not change the perceptual boundary for listeners, while the same change applied more generally did. Their study also showed that listeners would not spontaneously produce sound variants like they had heard (so production did not change when perception did), though they could imitate the sounds when directly asked to. Cutler et al.’s review of similar research points out that if a shift in perceptual boundaries generalizes to perception of new words, then some abstract phonemic representation must exist in addition to episodic traces of word pronunciations. They further show that a model based on MINERVA-2 cannot replicate the human perception data and actually predicts a reversed effect of exposure to the shifted sounds. They argue for a hybrid model in which talker-specific, episodic information about speech does get stored, but not in the lexicon; exemplars of different words can retune abstract phonetic categories instead.

Hay, Warren, and Drager (2010) show interesting cross-talker effects of recent accent exposure, with the dialect of an experimenter affecting perception of vowel contrasts in New Zealand English. Differences emerge between listeners who do or do not have certain vowels merged. They suggest that when a listener whose vowels are merged hears a dialect which preserves a distinction, the phoneme-level data becomes noisier and the distinction is harder for the listener during the test phase. This is explained by a theory in which specific exemplars of a word are stored, but also linked to phoneme categories. Listeners who have a single phoneme category for the merged vowels are not helped by hearing a dialect with the distinction unless more lexical processing is evoked, whereas listeners with distinct phoneme categories are aided in making a distinction by hearing a dialect which makes one.

---

1 Similar interactions of phonological generalization with lexical items can also be captured in cascading connectionist models (Goldrick & Blumstein, 2006; Baese & Goldrick, 2009).
Sumner and Samuel (2009) also study the issue of how dialect variants of words are stored, finding that the answer differs for people who use the dialect variant in question versus those who do not. Words like *baker* pronounced with the ‘dropped r’ of some New York dialects show priming effects just like the standard pronunciation only for listeners familiar with the dialect, whether or not they drop r’s themselves. However, those who hear but do not produce this variant lose priming over a 20-minute time span. They suggest that these non-producers are able to access the appropriate lexical entry during immediate processing, but seem to abstract away from the variant pronunciation over time, possibly not storing the phonetic details in the same way that producers of this dialect do.

The existence of Pierrhumbert’s model, and others like it, means that the comparison between exemplar models and neo-generative models is not dichotomous. Rather, one can define a theoretical spectrum of models, ranging from pure exemplar models (such as Hintzman’s (1986) MINERVA model, which guided Goldinger 1998) to neo-generative models such as Levelt (1980), and the issues which guided our experimental design allow us to locate the cognitive system with respect to this spectrum of models. Insofar as we find fast, systematic, categorical learning, we need key features of the neo-generative models. If positional constraints are important, this provides evidence for the kind of units needed at the phonological coding level. In contrast, pure exemplar models, with their epiphenomenal phonology deriving from a less abstract description of speech, do not provide for the same degree of plasticity in the phonological encoding, a point developed in Cutler et al. (2010). But key features of exemplar models can capture detailed phonetic learning, as well as lexical gang effects in generalization.

2. Background

2.1. American English flapping and /r/

Most sources agree that post-stress intervocalic /t/ is most frequently realized as a flap in American English in conversational speech. Zue and Laferriere (1979) conducted a production study on /t/ and /d/ in various environments which found flapping of /t/ occurring in 99% of post-stress intervocalic cases. Fisher and Hirsh (1976) found more variable results in their production study, with subjects ranging from 36% to 97% flap production, though perhaps some of those subjects were speaking more carefully or formally than others. Patterson and Connine (2001), who examined two corpora of conversational speech, found that 94% of post-stress intervocalic /t/’s were flapped; they found lower levels of flapping in low-frequency and morphologically complex words. This work thus suggests that the flap [ɾ] is the most common reflex of /t/ in this particular phonological environment, having been found both in experimental settings and in natural conversation.

The American flap differs phonetically from other allophones of /t/ by its short duration and voicing. Zue and Laferriere (1979) reported an average duration of 26 ms for flapped /t/. Fukaya and Byrd (2005) recorded word-final flaps as having an average duration of 20 ms, compared to voiceless stops in the same positions averaging 43 ms, and the majority of sounds they transcribed as flaps were voiced.

The normal realization of /r/ in American English is a voiced alveolar approximant [ɹ], which can vary by speaker in the degree of retroflexion (Ladefoged, 1993). This approximant appears on spectrograms with clear formants and lowering of F3 (Stevens, 1998; Foulkes & Docherty, 2000). There is no tendency for the flap to occur as an allophone of /r/ in American English, either intervocally or elsewhere.

2.2. Glaswegian English and our speaker

The speaker whose dialect our American English speakers were adapting to spoke Glaswegian Standard English. He was a native Glaswegian who had lived in Scotland up until he came to America for graduate study. At the time of this experiment, he was engaged in graduate study in Chicago, and he had lived there for 2 years. He had a strong Scottish personal identity, including
active involvement in Scottish political and cultural groups. His retention of his native dialect was very marked and when speaking fast, he could be quite unintelligible to American ears.

There are certainly different varieties of Scottish English and Glaswegian English, some differing from American Standard English in lexicon and grammar as well as pronunciation (Chirrey, 1999), but our experiment only involved Glaswegian pronunciation (we provided the lexical material). Our speaker used a flap or tap articulation for /r/, which Scobbie, Gordeeva, and Matthew (2006) describe as particularly likely in intervocalic post-stress contexts. His pronunciations did not show signs of the derhoticization described in Stuart-Smith (2007) and Lawson, Stuart-Smith, and Scobbie (2008), nor did he generally trill his /r/ (Scobbie et al., 2006 list this as an older pronunciation). The phoneme /t/ was primarily realized with aspiration by our speaker in all positions. In initial recordings, a glottal stop also occurred in medial positions (as would be expected, according to Stuart-Smith (1999) and Scobbie et al. (2006)), but this was infrequent and seemed to be in free variation with the aspirated /t/. To create the stimuli, we made selections from a larger set of recordings so as to present uniform allophonic patterns to the subjects. Utterances with a glottal stop for /t/ were discarded and only aspirated productions were used.

There are many other differences between Glaswegian and American English in addition to the /r/ and /t/ realizations, of course. Many of the vowels differ, for example, with the characteristic American [æ] replaced by a vowel much further back in the mouth. Additionally, Glaswegian English has different prosodic patterns, some of which were imitated by subjects (German, in press).

3. Methods

3.1. Stimuli
The sound patterns under investigation were represented by four conditions, with /t/ and /r/ each appearing in both prosodically strong (i.e., pre-stress), word-initial positions and prosodically weak (i.e., post-stress), word-medial positions (Fougeron & Keating, 1997; Pierrehumbert & Talkin, 1992). A total of 192 sentences were created, 48 of each type, with the constraint that no allophone of /r/ or /t/ appeared anywhere except in the target word of the appropriate condition. The target words were always sentence final, so as to be both prosodically prominent and easy to remember for participants. Sample items are shown in (1):

(1) /t/, word-initial (strong) position: He gave away his only token.
   /t/, word-medial (weak) position: The damp wind made him all sweaty.
   /r/, word-initial (strong) position: All the family’s belongings lay beneath the rubble.
   /r/, word-medial (weak) position: The boy swallowed mud because he was curious.

The items were grouped into four blocks, each containing twelve items of each type for a total of 48 per block. Items within each block were pseudo-randomized such that no two consecutive sentences were from the same condition. The four blocks of items were rotated through the task conditions in a counterbalanced order to avoid extraneous lexical effects. All of the blocks of items were recorded by the Scottish English speaker and put on CD. An additional group of three 12-item blocks was created and recorded for general re-familiarization with the accent. These blocks contained only non-target items, so the sentences had no /r/ or /t/ allophones in them at all (e.g., A display of the dig can be seen in the lobby). All of the items in the experiment are listed in Appendices 1-2.

3.2. Procedure
Each participant produced all four blocks of items in some task condition, and the blocks were counterbalanced to appear equally often in each condition. One block was produced as a Baseline. Before a participant heard any Glaswegian English recordings, they were asked to read a block of items in a normal conversational style from a script. This set served as an example of
the participant’s American production of /r/ and /t/. We did not ask subjects to produce a baseline block of items in a Scottish accent as we did not wish to reveal which accent was being used in the study. If we had identified the geographical origin of the accent, the results would most likely have been contaminated with subjects’ impressions of more familiar Scottish accents.

Another block of items was used for the Training tasks. Participants were told that this was a training session in which they were attempting to learn the accent of the speaker, and that they should try to imitate the way he said each sentence. The participants were given a script and a personal CD player with the relevant CD. The participant would listen to the Scottish English speaker producing each sentence in this block while following along on the written script, stop the CD, and then imitate the sentence into the microphone. This Training session was repeated once with the same procedure immediately after its first iteration. The two Training sessions together took under 20 minutes to complete, on average.

The final task in the first week was the Generalization1 task. The participant was given the script of a third block of items, which they had not previously seen nor heard the Glaswegian English speaker produce, and asked to continue imitating the accent. They did not have a CD to imitate.

Each participant returned to the lab a week after their first session. In this session, three blocks of items were recorded: the Training block again (making the third time through this block), the Generalization1 block again, and a fourth block of items for the Generalization2 task. The order of these three task types was counterbalanced so that each was recorded first, second or third by an equal number of participants. Before each of the target blocks, participants refreshed their memory of the speaker and accent using one of the Non-target re-familiarization blocks of items. They would listen to the Scottish English speaker on CD and imitate him, as in the first week’s Training sessions, except that these 12-item blocks did not contain any /t/ or /r/ sounds. Therefore the accent in general was re-familiarized, but the specific pronunciations of /t/ and /r/ were not repeated for participants. Participants did not hear the speaker produce any of the target items from the Training or Generalization blocks during Week 2. The full set of recordings is summarized in Table 1.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recording tasks by week. Tasks that share a row involve identical blocks for a given speaker.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Week 1 (fixed order of tasks)</th>
<th>Week 2 (rotating order of tasks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Training 3</td>
</tr>
<tr>
<td>Training 1, Training 2 (with CD)</td>
<td>Generalization 1R</td>
</tr>
<tr>
<td>Generalization 1</td>
<td>Generalization 2</td>
</tr>
<tr>
<td></td>
<td>Non-target (with CD, one block preceding each task above)</td>
</tr>
</tbody>
</table>

The recordings were made using a Shure SM 81 microphone connected through an Ariel Proport, an Earthworks preamp, and an Apogee PSX 100 A/D into a Macintosh G4 computer running ProTools. The microphone and participants were located inside a sound-attenuated recording booth. The recordings were saved as mono sound files sampled at 22050 Hz and burned onto CDs.
3.3. Participants

The participants in this study were 24 undergraduate students at Northwestern University enrolled in lower-division linguistics classes. They received course credit for their participation. Data from bilingual and non-native participants was excluded from analysis, as was that of students who were unable to return for the second session. The students ranged in age from 19 to 38, and their average age was 22. All but three of the participants had studied at least one foreign language, and twelve of them had studied Spanish. Eight of the participants were male.

3.4. Data Analysis

Each of the recorded sound files from participants was inspected and labeled by one of the first two authors, while both of the first two authors examined the Glaswegian English speaker’s productions and a small set of evenly distributed other files to assess intercoder agreement. Labelers listened to the final word of each sentence while examining the waveform and spectrogram using Praat (Boersma & Weenink, 2011). Initially, waveform, spectrogram and auditory evidence was used to determine whether the target was within the set of alveolar gestures targeted by the study (i.e., [t], [tʰ], [ɹ] or [ɾ]), or whether it involved some other place of articulation (e.g., velar) or manner of articulation (e.g., trill) not predicted by the allophony patterns of the two dialects involved. For the relevant alveolar consonants, if the acoustic evidence supported the presence of a closure, then the endpoints of the closure were labeled. The point of voicing onset was also labeled if it differed from the end of the closure. For voiced gestures, F3 was measured by inspection at the point in or near the target gesture where it reached a minimum. Closure duration and voice onset time were later extracted automatically using Praat.

The central goal of our study is to test whether speakers successfully reproduced the Glaswegian pattern of phoneme realization associated with /t/ and /r/. We therefore used an objective method for categorizing any given production of a target segment as a “success” or not. In essence, the method decides, for each production of /t/, whether or not it is realized as [tʰ], and for each production of /r/, whether or not it is realized as [ɾ]. This procedure is described in Section 5. In the next section, we first report the results of the phonetic measurements in order to give a sense of the variety of productions.

4. Phonetic Results

4.1. Phonetic Outcomes for /t/

The observed productions of /t/ included voiceless alveolar closures followed by a voicing onset delay (suggesting [tʰ]), voiced alveolar closures generally short in duration (suggesting [ɾ]), voiced alveolar gestures lacking complete closure but resembling [ɾ], and a very few other sounds. In cases where the speaker was clearly aiming at a different target sound, as in the fairly common mispronunciation of the initial segment of Thames as [θ], the data were excluded as being irrelevant.

The acoustic results for /t/ reflect our broad expectations. As the summary data in Table 2 shows, the American subjects pronouncing initial /t/ in the Baseline task nearly always included a long voiceless closure (averaging over 40 ms) followed by a voicing onset delay averaging over 70 ms, consistent with [tʰ]. The Glaswegian speaker’s initial /t/s were very similar, as were the imitated versions in the Training and Generalization tasks. These results are consistent with, though slightly longer than, the findings of Lisker and Abramson (1967), who report a mean VOT of 48 milliseconds for /t/ in initial positions produced by American speakers of English.
Table 2
Summary of closure duration, VOT and F3 minimum for production of /t/ for native Glaswegian model, Baseline American, and imitation tasks.

<table>
<thead>
<tr>
<th>Speaker/Trials</th>
<th>Initial /t/</th>
<th>Medial /t/</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Closure</td>
<td>100%</td>
<td>95%</td>
</tr>
<tr>
<td>Average Closure Duration, ms (SD)</td>
<td>53 (15)</td>
<td>43 (23)</td>
</tr>
<tr>
<td>% of Trials with Voicing Onset Delay</td>
<td>100%</td>
<td>99.7%</td>
</tr>
<tr>
<td>Average VOT, ms (SD)</td>
<td>70 (11)</td>
<td>74 (20)</td>
</tr>
<tr>
<td>Average F3 minima, Hz (SD), females</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Average F3 minima, Hz (SD), males</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

When less than 5% of the data fit into a category, averages were not calculated, because the small number of tokens are likely to be unevenly distributed across speakers or items.

Summary of closure duration, VOT and F3 minimum for production of /t/ for native Glaswegian model, Baseline American, and imitation tasks.

Voiceless aspirated closures of a slightly shorter duration were observed for the Glaswegian pronunciations of medial /t/. In the imitated Training and Generalization tasks, participants also produced mainly voiceless aspirated stops medially, thereby approaching the Glaswegian dialect.

Medial /t/ in the Baseline task was most often realized with a relatively short, voiced closure, consistent with [ɾ], the expected American English allophone. Figure 2 shows the frequency distribution for these productions, with the most frequent closure duration being around 20 ms and the average 23 ms. These measures are consistent with the findings of Zue and Laferriere (1979), who report a range of 10-70ms for ‘flapped’ /t/ in a falling stress context, though Fukaya and Byrd (2005) found a somewhat smaller range of 5-40 ms for the three speakers in their study.
A smaller proportion of Baseline medial /t/ s were produced with no identifiable closure, though these sounded to the labelers like instances of [ɾ]. Stone and Hamlet (1982) reported similar ‘less closed’ [ɾ]-like variants of /d/ in American English that “appeared as a momentary decrease in the intensity of the preceding and following vowels and during which there was occasionally a small burst” (404-405). Finally, a small number of Baseline medial /t/ s were produced with the voicing onset delay characteristic of [tʰ].

4.2. Phonetic Outcomes for /r/

The observed productions of /r/ were more varied, including voiceless alveolar closures with a short duration (suggesting [ɾ]), voiced alveolar gestures lacking a closure (suggesting either [l] or [ɾ], see discussion below), trilled [ɾ]s, and voiced uvular or velar fricatives (resembling [ɤ] or [ɣ], Figure 3). Some participants produced an interesting retroflex palato-alveolar fricative resembling [ʐ] (Figure 4) and occasionally an [l]- or [w]-like sound.
Figure 4. *New road* as produced by speaker B1a in Generalization 1R: /r/ → [ʐ]

In some productions, the auditory evidence suggested a brief, flap-like closure, but the waveform and spectrogram showed an event which had a closure onset but a release too gradual for the end to be marked definitively. A typical example is illustrated in Figure 5.

Figure 5. Waveform and spectrogram showing a sudden acoustic transition that suggests a closure onset with no observable corresponding offset or release. The speckled trace represents intensity.

In the Baseline task, /r/ was almost exclusively produced with no closure and a marked lowering of F3, consistent with a normal American [ɹ] (Stevens, 1998). This is reflected in the lower F3 minima for Baseline /r/ compared with Baseline medial /t/ as illustrated in Figure 6 and Table 3.
Figure 6. Boxplot of F3 minimum for three combinations of phoneme and position in the Baseline task. Baseline initial /t/ was always produced with a voiceless closure so F3 was not measured for any tokens in that condition.

Table 3. Summary of closure duration and F3 minimum for production of /ɾ/ for native Glaswegian model, Baseline American, and imitation tasks.

<table>
<thead>
<tr>
<th>Speaker/Trials</th>
<th>Initial /ɾ/</th>
<th>Medial /ɾ/</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Closure</td>
<td>77%</td>
<td>90%</td>
</tr>
<tr>
<td>Average Closure Duration, ms (SD)</td>
<td>24 (13)</td>
<td>15 (6)</td>
</tr>
<tr>
<td>Average F3 minima, Hz (SD), females</td>
<td>NA</td>
<td>1910 (202)</td>
</tr>
<tr>
<td>Average F3 minima, Hz (SD), males</td>
<td>1971 (216)</td>
<td>2123 (244)</td>
</tr>
</tbody>
</table>

b When less than 5% of the data fit into a category, averages were not calculated, because the small number of tokens are likely to be unevenly distributed across people or items.

The majority of /ɾ/s produced by the Glaswegian speaker had a short, voiced closure with little or no discernible dip in F3, consistent with [ɾ]. Glaswegian initial and medial /ɾ/s were similar to each other, with the medials shorter on average. Importantly, both Glaswegian flap duration means are similar to that for medial /t/ in the Baseline American pronunciations (Table 2), suggesting that participants are likely to make an association between the two categories. There were also more Glaswegian tokens lacking clear acoustic closure for initial than for medial /ɾ/, but in both cases these resembled [ɾ] auditorily.

The Training and Generalization imitation tasks were where participants produced a variety of sounds for /ɾ/, including some with short, voiced closures and some with voicing but
no closure. Clear closures were present for less than half of the tokens for both initial and medial /r/. Similarly to the Glaswegian speaker, initial /r/ s with clear edges were slightly longer on average than medials. The closure durations showed substantial overlap with those for Baseline medial /t/ and for Glaswegian /r/, consistent with categorization of the tokens with closure as instances of [ɾ]. For tokens with measurable formants, F3 minima exhibited a wide range of values, suggesting a mixture of both rhotacized and non-rhotacized tokens, but were higher overall than F3s for American Baseline /r/ s.

4.3. F3 and Interspeaker Variability

There was significant variation in the range of F3 minima among speakers in a way that was independent of the phonetic category being produced. This is particularly apparent in Figure 7, which shows the distribution of F3 for baseline /r/ for all 24 speakers organized by median score. There is a relatively abrupt transition that roughly corresponds to a division between male speakers (lower F3 distributions toward the left) and female speakers (higher F3 distributions toward the right). These results are consistent with known effects of vocal tract length on absolute formant values (Stevens, 1998), and highlight the importance of taking such individual differences into account when trying to distinguish between [ɾ] and [ɹ] in the non-Baseline tasks, for which the target categories are not predictable a priori.

The Glaswegian speaker used in our study is a particularly tall individual, who likely has a very long vocal tract and low overall formant frequencies. The F3 associated with his production of [ɾ] is therefore expected to be very low independently of whether it involves an actual lowering. This likely explains, for example, why the mean F3 for medial [ɾ] produced by the Glaswegian speaker is only slightly higher than the mean of F3 for [ɹ] by female speakers in the Baseline task, as shown in Table 3.

![Figure 7](image-url). Boxplots of F3 minima for all Baseline /r/ for individual speakers arranged by increasing median.
5. Categorization Procedure

5.1. Identification of \[ tʰ \]

The \[ tʰ \] allophone is most readily distinguished from other alveolar categories on the basis of its voiceless closure and relatively long, positive voice onset delay. For the purposes of our analysis, we categorized as an instance of \[ tʰ \] any alveolar gesture that included a voiceless closure followed by a delay in voicing onset. Since the unaspirated \[ t \] allophone of \(/t/\) is also predicted to be voiceless with a short but positive voice onset delay, this method potentially misclassifies any instances of \[ t \] as \[ tʰ \]. Such errors are unlikely for three reasons. First of all, the minimum VOT for \(/t/\) produced by the Glaswegian speaker was 39 ms, so there were no clear instances of \[ t \] in the auditory stimuli. Second, none of the targets included \(/t/\) in a phonological environment that is associated with \[ t \] in American English (e.g., following \(/s/\) in an onset). Finally, as Figure 8 shows, the frequency distribution for VOT in the imitation tasks suggests that our data follow a single distribution with a range and mode that are much higher than would be expected for \[ t \] (Lisker & Abramson, 1967).

![Figure 8](image.png)

**Figure 8.** Relative frequency density distribution of VOT for \(/t/\) in all imitation (i.e., non-baseline) tasks.

5.2. Identification of \[ ɾ \]

In our study, all targets that included a voiced closure were categorized as \[ ɾ \]. Although this method potentially includes any instances of \[ d \], note that speakers had access to the orthographic representations of the targets, which never included \(/d/\). Additionally, the frequency distribution for closure durations shown in Figure 9 suggests that these observations follow a single distribution centered on a value that is typical for \[ ɾ \]. Thus, it is unlikely that this method would result in a large number of such errors.

As noted above, \[ ɾ \] is occasionally produced without a full closure both in the Baseline American productions of medial \(/t/\) and in the Glaswegian productions of \(/r/\). Since \[ ɾ \] is also realized without closure, some other measure must be used to distinguish between the two categories for those productions lacking a closure. A widely recognized acoustic correlate of \[ ɾ \] is a marked lowering of the third formant (Stevens, 1998). Thus, \[ ɾ \] is predicted to have a lower F3 as compared to \[ ɾ \]. While Figure 6 above shows that F3 is overall very different for these two categories in the expected direction, the significant overlap in the distributions by speaker...
for the Baseline task suggests that classification based on a single F3 threshold would result in a substantial amount of error.

Therefore, we calculated a separate F3 threshold for each speaker based on their Baseline productions of medial /t/ and /r/, for which the underlying phonetic categories are known. Specifically, we used optimal discriminant analysis to find, for each speaker, the single way of dividing the combined F3 distribution for [t] and [r] into two categories, such that the total number of errors (i.e., [t]s categorized as [r] plus [r]s categorized as [t]) is minimized. Since such a method yields a pair of observations that span the optimal cutpoint, whereas the use of a threshold requires a single value on the F3 scale, we therefore took the midpoint (mean) of these two points following Yarnold and Soltysik (2005). The reliability of this discriminant can be evaluated by calculating the proportion of successes out of the total number of relevant observations in the Baseline task, where we knew whether participants were producing an allophone of /t/ (the flap) or /r/. The mean reliability score across the 24 speakers is 0.97 with a standard deviation of 0.036, suggesting that this normalization procedure is an effective method for distinguishing between [t] and [r] on the basis of F3.²

![Figure 9. Relative frequency density distribution of closure duration for all tokens categorized as [r] in the imitation tasks.](image)

**5.3 Inter-labeler Reliability**

In order to assess the consistency of this method across labelers, a series of analyses was performed on the classification results using Cohen’s Unweighted Kappa. For the Glaswegian speaker, category agreement between the labelers was perfect (Kappa = 1). For 7 of the /r/-initial tokens, and 5 of the /r/-medial tokens, the labelers disagreed on whether a closure was present, though in all such cases the labelers agreed that the phonetic category produced was [r].

An experimentally balanced and evenly distributed subset of the participant data was labeled by both labelers, comprising 672 tokens taken from each task of each speaker. With four labeling categories ([tʰ], [t], [r] and “other”), the interlabeler reliability was found to be Kappa = 0.92, with a 95% confidence interval of 0.894 to 0.946.

² The Glaswegian productions did not include any tokens of [t], so it is not possible to apply this method to those data.
[ɪ] and [ɾ] represent the largest source of interlabeler differences, accounting for 95% of all disagreements. This fact likely reflects the high variability associated with the production of /r/ in the various imitation tasks. Thus, a lower bound on inter-labeler reliability may be approximated by considering only tokens involving /r/ in a non-baseline task. This lower bound was found to be Kappa = 0.83, 95% CI (0.763, 0.894), which is considered “excellent” or “nearly perfect” according to commonly cited guidelines (e.g., Landis & Koch 1977, Fleiss 1981).

6. Categorization Results

The categorization results are shown first in Figure 10 and Figure 11, which display the percentage of Glaswegian-like outcomes for /t/ and /r/ respectively.

**Figure 10.** Mean percentage of [tʰ] outcomes by task for /t/ in word-initial and word-medial positions.

**Figure 11.** Mean percentage of [ɾ] outcomes by task for /r/ in word-initial and word-medial positions.
It is clear from Figure 10 that participants came close to 100% success in producing aspirated /t/ in the word-initial position. For /t/ in word-medial position, all participants fluently produced flaps in the Baseline condition at an average rate of over 95%, though eight speakers produced at least one instance of [tʰ] here, including two that produced 25% and 33% of tokens as [tʰ].

The condition with /t/ in word-initial position served as a control, with participants producing the aspirated allophone expected for both native and imitated targets in all tasks. The condition with /t/ in word-medial position tested whether speakers could suppress the usual flap allophone in favor of the aspirated allophone typically found in a different environment. Excluding medial /t/ in the Baseline task, the average rate of [tʰ] production for both conditions was approximately 98%. The difference between the initial and medial /t/ conditions, though small, was significant in a between-items ANOVA exploring the effects of training on lexical items and time, containing the Training2, Generalization1, Training3, and Generalization2 tasks (F2(1,94) = 32, p < 0.001); the test could not be conducted by speakers due to insufficient variability in the initial /t/ data. This analysis by items also showed significant effects of practice on lexical items, since performance was better in the Training tasks (F2(1, 94) = 6, p < 0.05). An ANOVA by speakers on only the medial /t/ results showed a similar effect of lexical items (F1(1, 23) = 6, p < 0.05). Neither analysis showed any significant effects of time, as participants’ performance did not drop noticeably in the second week, nor interactions of time with training on lexical items.

The flapped /r/s were clearly more difficult for the participants, with average percentages below 50% for /r/ in initial position and below 80% for /r/ in medial position. There was variation in performance, too, with some individual subjects who achieved 100% performance on /r/ conditions as early as the Training1 task, and others whose highest success rate in any imitated /r/ condition was 8%. This may be related to participants’ innate ability to mimic, which has been shown to affect the degree of foreign accent (Flege, Yeni-Komshian, & Liu, 1999; Piske, MacKay, & Flege, 2001; Purcell & Suter, 1980; Thompson, 1991). This may also be related to participants’ previous language experience, since Spanish, for example, uses flapped and rolled /r/s. The rest of the statistics below will focus on the /r/ conditions as being of most interest and variability.

The two first-week Training tasks were examined to see whether participants improved their imitation with additional exposure to the Scottish speaker. An ANOVA on /r/ performance in initial and medial positions in Training1 vs. Training2 was conducted; the factor of r-position was within-subjects but between-items, while the task factor was within-subjects and within-items. There was a significant main effect of r-position, with better performance for /r/ in medial position than in initial position (F1(1, 23) = 37, p < 0.001; F2(1, 94) = 45, p < 0.001). There was also a significant main effect of additional training, such that participants’ performance improved in Training2 relative to Training1 (F1(1, 23) = 12, p < 0.005; F2(1, 94) = 31, p < 0.001). The interaction between these factors was non-significant. In general, then, participants improved their rate of flapping for /r/ on the second time through the Training task, though performance on words with /r/ in medial position was better than for words with /r/ in initial position from the very start.

In order to examine the effects of time and training on specific lexical items, an ANOVA was conducted on /r/-initial versus /r/-medial items in the Training2, Generalization1, Training3 and Generalization2 tasks. There was a significant effect of position, with higher rates of flapping in medial position than in initial position (F1(1, 23) = 29, p < 0.001; F2(1, 94) = 78, p < 0.001). There was a significant main effect of time, with a small performance drop between the first and second week’s sessions (F1(1, 23) = 7, p < 0.05; F2(1, 94) = 18, p < 0.001). There was a significant main effect of practice on lexical items, since the Training tasks showed higher levels of success than the Generalization tasks in both weeks (F1(1, 23) = 10, p < 0.005; F2(1, 94) = 11, p < 0.001). Finally, there was a significant interaction between r-position and time, with a larger performance difference between weeks for /r/ in medial position than for /r/ in initial position (F1(1, 23) = 6, p < 0.05; F2(1, 94) = 5, p < 0.05). No other interactions
approached significance. Figure 10 and Figure 11 clearly show that performance during Week 2 did not fall back to Baseline American English levels, meaning that speakers retained much of the dialect adaptations they had learned during the first week’s training. Also, although performance in the Training tasks was better than in Generalization tasks, the Generalization results were also far above the Baseline results, showing extension of the flap allophone to new lexical items.

Because of counterbalancing, different subjects encountered different block types in week two in different orders. Figure 12 and Figure 13 summarize the data by block type and block order respectively.

**Figure 12.** Percentage of successful Glaswegian allophones, Week 2, by tasks.

**Figure 13.** Percentage of successful Glaswegian allophones, by position of block in the Week 2 sequence.
An ANOVA on the three blocks of items by order of recording (First, Second, and Third) showed a significant main effect of r-position, with medials showing higher rates of flapping than initials (F1(1,23) = 18, p < 0.001; F2(1,94) = 53, p < 0.001), but the main effect of order was only significant by items (F1(2,46) = 1.5, p = 0.233; F2(2,188) = 4, p = 0.014). There were no significant interactions. Therefore, the order of block types in the second week did not reliably affect performance.

To fairly test whether task affected second week performance, an analysis compared only the Training3 and Generalization2 tasks (since Generalization1R was a set of items which were in between practiced and new items, having been new in Week 1 but repeated in Week 2). In this ANOVA, the effect of /r/ position was robustly significant (F1(1, 23) = 13, p < 0.005; F2(1,94) = 29, p < 0.001), and the effect of training on lexical items was also significant (F1(1,23) = 5, p < 0.05; F2(1,94) = 4, p < 0.05). Thus there was a small advantage for lexical items which were trained in the first week.

All of these statistics have showed a strong effect of word-initial versus word-medial position for /r/. However, there was a minority of word-initial /r/ targets (15 out of 48) in which /r/ followed a consonant, as the preceding word was consonant-final (e.g., good reason). Since the usual environment for flap in American English is intervocalic, it could be that the group of items with non-intervocalic /r/ in initial position accounts for the difference between initial and medial position data. We therefore carried out a post-hoc analysis to evaluate this issue. Figure 14 shows the percentages of success for the intervocalic vs. non-intervocalic items with /r/ in initial position as well as the items with /r/ in medial position.

![Figure 14](image.png)

**Figure 14.** Percentage of flaps for /r/ items in word-initial position, intervocalic (33 items) vs. non-intervocalic (15 items), plus percentage for /r/ in word-medial positions.

Indeed, the intervocalic set of /r/-initial items showed higher percentages of flapping than the non-intervocalic items in all of the tasks except the Baseline. The difference between the intervocalic and non-intervocalic word-initial items was significant in within-subjects and between-items ANOVAs including the Training2, Generalization1, and Training3, and Generalization2 blocks (F1(1, 23) = 16, p < 0.001; F2(1, 46) = 14, p < 0.001). Nevertheless, similar ANOVAs on the items with medial /r/ vs. only the intervocalic initial /r/ items showed that there was still a fully significant main effect of prosodic position, with greater success for medials (F1(1,23) = 19, p < 0.001; F2(1, 79) = 44, p < 0.001). Thus the advantage for /r/ in word-medial position persists even when compared to only the subset of items with /r/ in word-initial position which were also intervocalic. Additionally, the lexical factor had a significant effect in the analysis using only the intervocalic initial /r/ items, as the Training 2 and 3 blocks
had more flaps than the Generalization 1 and 2 blocks (F1(1,23) = 12, p < 0.005; F2(1,27) = 8, p = 0.005).

The reasonable success of subjects in producing medial flapped /r/s suggests that they were able to map this existing allophone to a different phoneme, but the remapping did not always guarantee the target results or was not always successful. That is, in addition to completely non-adapted American responses, most subjects also produced phonetic innovations or approximations. These were sounds which could not be categorized as either a flap or an American /r/, sharing some features of both. As we do not know whether the participants were intending to produce a flap but failing or simply exploring the phonetic space for a new sound, we cannot settle whether these should rightly be called innovated or approximated sounds; for ease of exposition, they will be called innovations. Figure 15 shows the percentage of successful flaps and of additional innovated outcomes for both /r/ conditions (the level of success in the /t/ conditions meant that there were very few innovative or non-adapted responses).

![Figure 15. Mean percentage of flap recruitment and innovated outcomes, /r/ in word-medial and word-initial positions.](image)

The proportion of innovated trials was highest for the /r/s in word-initial position and lowest for the /t/ conditions. Looking at phonetic innovations by subjects, we found that the vast majority of subjects who produced phonetic innovations also produced successful flaps, rather than particular speakers producing only these non-target sounds and not the Glaswegian targets. The intervocalic vs. non-intervocalic word-initial items were also examined. The rate of phonetic innovations for the non-intervocalic word-initial /r/s equaled or exceeded the rate of phonetic innovations for the intervocalic word-initial /r/ items in most blocks. That is, the more difficult environment following a consonant resulted in more innovated outcomes instead of successful flaps. Another interesting phonetic outcome found in the non-intervocalic word-initial /r/ data was the epenthesis of a short unstressed vowel. Most of the speakers, including even the Scottish speaker, used this strategy at least once during the experiment in order to place the /r/ in an intervocalic context.

The phonetic innovation data could suggest that subjects intended to produce something other than an American /r/ in the partially successful trials for word-initial /r/, but failed to fully execute either the phonetic remapping or the articulatory maneuvers necessary for a flap. This highlights a possible connection between the innovation data and the intervocalic/non-intervocalic data in Figure 14, suggesting that American English speakers are only practiced at articulating a flap sound in a medial intervocalic position, and thus have difficulty producing it in any other environments. The resulting variety of other sounds produced, including a retroflex palato-alveolar fricative, a uvular sound, and a trill, suggests that they were exploring their
phonetic resources. Frequent use of trill by subjects who had taken Spanish suggests that speakers were accessing and utilizing a range of available resources including those acquired through an L2.

The above explanation for the innovation data is supported by data from Munson (2001) on error rates in the production of phonological patterns as a function of frequency. He found that infrequent sequences of sounds were more likely to be produced slowly or incorrectly than frequent sequences, even though all of the sequences did occur in grammatical English words. It would not be surprising, then, for our speakers to have difficulty producing the flap in a word-initial post-stress context (especially a post-consonantal context).

7. Discussion

The dominant effect in our study was that speakers were able to modify their phonological coding system in order to approximate the speech of an unfamiliar speaker in an unfamiliar dialect. In particular, they were able to produce [tʰ] for /t/ reliably in contexts where that phoneme is realized by [ɾ] in their native dialect, and most speakers were able to produce some [ɾ]s in place of [t] for the phoneme /t/. This learned ability was systematic to the extent that it generalized strongly to words not in the training materials. In that sense, our main finding represents the production counterpart to the perception results of Maye et al. (2008) and Peperkamp and Dupoux (2007). In addition, our results show that the ability persisted over a period of one week. Thus, to the extent that speakers can learn a modified coding system, they can do so semi-permanently.

Subjects in our study were able to reassign [ɾ] to /r/ in both prosodically weak, word-medial contexts and prosodically strong, word-initial contexts, though their performance was better in word-initial positions where [ɾ] typically occurs in D1. This may be partly due to a difference in articulatory difficulty between the two contexts, since the airflow required to produce [ɾ] was reduced in word-initial positions, where /r/ was preceded by an unstressed vowel or consonant, relative to word-medial positions, where /r/ was always intervocalic and preceded by a primary stressed vowel. Many languages including Glaswegian English, however, use rhotic flaps or taps in such initial positions without difficulty, so this effect is most likely a secondary one. The discrepancy is perhaps best accounted for in terms of the speakers' experience. Motor patterns, such as the articulation of a flap, are learned in context and learned more robustly with a large number of examples. Speakers of American English have experience producing [ɾ] in medial, falling stress, intervocalic contexts across a large number of words, whereas their experience producing [ɾ] in other contexts is very limited. The generalization that is most readily available to them, therefore, is for producing [ɾ] in prosodically weak, word-medial positions. Edwards, Beckman, and Munson (2004) showed that children’s repetition accuracy of phoneme sequences in non-words was correlated most strongly with the frequency of the sequence in the lexicon, thereby demonstrating the importance of sequential practice in a variety of cases. While it is not possible in our study to determine the exact relationship between [tʰ] and [ɾ] as they occur in the imitative speech and variants of those sounds in the subjects’ native dialect, the high degree of phonetic similarity in terms of closure duration,

---

3 [ɾ] is commonly described as an allophone of /t/. Comparing the classical notion of the allophone to its coverage in the modern literature, a flap is minimally a highly routinized variant of /t/. In a classical linguistic approach, the transfer of an allophone from one context to another is a type of abstraction or generalization. Meanwhile, the interpretation of “transfer” from a motor perspective is also a type of generalization. Since this paper is concerned primarily with the generalized productivity of the system, we set aside the issue of the
VOT, and F3, combined with the pattern of success across prosodic contexts provide strong evidence that subjects were able to (a) extend their D1 knowledge of the [tʰ] variant of /t/ to prosodically weak (medial) positions in D2, and (b) transfer their knowledge of the [ɾ] allophone of /r/ in D1 for realizing /r/ in D2.

Note that subjects did have some success at producing [ɾ] in prosodically strong, word-initial contexts, a fact which indicates that sub-phonemic variants, rather than phonemes in the classical sense, are the relevant level of coding for the learning that took place. To see why, consider that if the adaptation to D2 involved only modification of the relation of the phonological code (phonemes) to the lexicon, then we would expect recruited phonemes to obey the same prosodic conditioning that they do in D1. Thus, if /t/ were being substituted across the board for /ɾ/, /tʰ/ would be realized as [ɾ] in word-medial position and [tʰ] in word-initial position. Further support for sub-phonemic encoding derives from the fact that speakers were able to replace an allophone from one prosodic context (the flap) with an allophone from another prosodic context ([tʰ]) within the same phoneme category (/t/). This is not predicted by a model that only permits realignments at the level of phonemic encoding. Both results converge with Polka’s (1991) suggestion that listeners are able to use their implicit knowledge of English allophony as a resource for processing non-native phonemic contrasts.

As mentioned in Section 1, we found it necessary to provide subjects with orthographic transcriptions of the speech they were attempting to imitate. Unlike some other studies (Weber and Cutler, 2004), then, we were unable to deconfound the effects of orthography and categorical learning. There is, in fact, a line of research going back to Jaeger (1980, 1984) suggesting that orthography is relevant to phonology for literate speakers (see also Steinberg & Krohn, 1975; Armbruster, 1978). Most notably, in the L2 literature, orthography is demonstrably activated during speech production. Kaushanskaya and Marian (2007), for example, found an interference effect (i) between L1 orthography and L2 phonology and (ii) between L2 orthography and L1 phonology in a picture naming task. Since our study is concerned with dialect learning rather than second language acquisition, it is broadly related to results of this type, though not closely parallel. Our results are entirely consistent, however, with the idea that learning at the categorical level was being facilitated by knowledge of orthography. In other words, it is quite plausible that the presence of orthography enhanced speakers’ ability to both access an intermediate (i.e., sub-phonemic) level of representation and learn remapping relative to it. Our results nevertheless support the need for a model with two levels of representation (abstract categorical and phonetic), where learning can take place at each of the levels, and we leave it to future research to address the question of whether categorical effects would have predominated to the same degree in a study involving only auditory stimuli from a more accessible dialect.

In addition to systematic effects of the kind discussed above, our results also showed certain word-by-word effects. That is, subjects performed better on items from the Training task than on new items, both immediately and after a period of one week. Because speakers can simply memorize new word-forms for specific words in the lexicon, effects of this type are not inconsistent with neo-generative models. They are explicitly provided for, however, by a hybrid model in which generalized coding principles project from known examples. According to that model, then, our subjects began the learning task by storing new forms in the lexical entries for individual words (for example, [bɛɾiz] for berries), and then rapidly (i.e., by the beginning of the first Generalization set in Week 1) generalized from those examples to coding rules for realizing the target phonemes in the relevant phonological contexts (e.g., that /ɾ/ is realized as [ɾ] in both strong (initial) and weak (medial) environments).

In a hybrid model, exemplar effects operate at a second level as well. In other words, not only do grammar rules project from known examples in the lexicon, but the coding units themselves project from known phonetic experiences. In addition to recruiting [ɾ] for the realization of /ɾ/, subjects in our study realized /ɾ/ with sounds not found in American English.

precise formal relationship between the two instances of [ɾ] (i.e., as a realization of /t/ in D1 and a realization of /ɾ/ in D2), and continue to treat them as instances of the same allophone.
Several subjects, for example, produced /r/ with some variant of a retroflex alveolar fricative [ʐ], and others with variants of [ś], [ř], and [ʒ]. Innovations of this type represent a larger proportion of all substitutions (i.e., non-[ɹ] productions) in prosodically strong, word-initial positions than in word-medial positions, suggesting that recruitment of preexisting categories is preferred whenever possible. In other words, only when a D2 category cannot be identified with a preexisting one, or when implementation or articulation of a preexisting category is inhibited, does a speaker begin to explore the phonetic space. Compare this to the L2 learning models in Best et al. (2001) and Flege (1995), which predict that speakers will form a new perceptual category only when the phonetic target in L2 cannot be identified with a preexisting L1 target. To the extent that category formation projects from experiences, this result also reinforces the claim in Pierrehumbert (2002) that the categorical behavior of the perception and production system is basic to the architecture, and that exemplar effects play a secondary role.

The total picture is thus illustrated by Figure 16. Dashed arrows in the figure show how learning begins when individual lexical items become associated with alternative pronunciations independently of any phonological encoding or parse rules present in D1 (represented by solid arrows). The central feature of our model, however, is that systematic realignment may occur between phonemes and individual allophonic categories. This is represented as a deviation from solid arrows to dotted arrows at the level of phonological encoding in production, and at the level of the phonological parse in perception.

![Figure 16](image_url)

Figure 16. Production (left) and perception (right) architecture for proposed model in which realignment occurs at the level of subphonemic encoding. Lexical learning occurs when specific lexical items are associated with new subphonemic variants (dashed arrows). Categorical realignment may occur at the level of phonological encoding or the phonological parse (dotted arrows). The ultrasound images on the left show the outline of the tongue during the production of the three corresponding phones.

In our experiment, it was not practical to carefully control for the amount and type of language experience that subjects brought with them to the trials. It would have been impossible to determine, for example, whether a given subject had ever heard Glaswegian English, perhaps even unknowingly, in their lifetime. It would have been even less practical to rule out any subject who had prior experience or practice either with a different dialect of English that includes similar phonological features (e.g., Southern British with regard to /t/), or with an entirely different language that has similar phonetic categories in similar phonological contexts (e.g., Spanish with regard to /r/). What we do know, and what was verified by our Baseline...
condition, is that all subjects were native, first-language speakers of a dialect of American English in which the relevant features of our study are not present. Furthermore, we know that there were no native speakers of Glaswegian English in our study. In fact, informal exit interviews suggest that most of our subjects could not identify the dialect they heard as a variety of English spoken in Scotland, and several could not even narrow its origin to the British Isles.

To give one example of the diversity of experience level among our subjects, one subject’s productions sounded remarkably like those of a native speaker of an Indian variety of English, both on the training and generalization tasks. Her subject questionnaire, however, indicates that she is a native speaker of American English, and her baseline recordings confirm this. In an exit interview, the subject reported having had significant contact with the India-born mother of a childhood friend, whom she had learned to imitate through repeated practice. In the subject’s attempt to access the experimentally targeted dialect, the social index associated with the experientially robust Indian English features provided a strong competitor for the relatively weakly reinforced social index associated with the experimental dialect. The result is that she was quite successful both at producing [ɾ] for /r/ and at producing [tʰ] in medial position, in addition to other features more typical of Indian English than Glaswegian English. In keeping with our theoretical claims, then, this subject’s behavior was both systematic and categorical, yet it displays the contextual and social biasing effects that are indicative of an exemplar component to the model.

Whatever the maximum level of speech experience was that our subjects brought to the experiment, any success they demonstrated in the tasks required one of two abilities. Either they replaced a preexisting category with a new one which they were able to generate parametrically, or they activated a preexisting category in a novel lexical and social context. Either way, the learning was systematic to the extent that it applied to both familiar and unfamiliar word and sentence contexts, and it was long-term, since it persisted over a period of one week. Comparing our results to Pierrehumbert’s (2002) hybrid model then, we find support for the relevance of all four proposed mechanisms. To the extent that subjects in our experiment succeeded at replacing [ʃ] with the flap from their native dialect or from another language, they were able to modify their pronunciation of known words using preexisting categories, while those who succeeded by learning a novel articulation of /r/ demonstrated the ability to form new phonetic categories parametrically through exposure and practice. In both cases, subjects encoded these new pronunciations as generalized phonological principles. All subjects were able to produce [tʰ] in prosodically weak, word-medial contexts. Some subjects accomplished this by learning to produce a preexisting category in novel lexical contexts, while those whose native dialect includes [tʰ] in weak position demonstrated an ability to learn the situational bias associated with that pattern. In sum, our results show that systematic effects strongly dominate the learning mechanism, though exemplar effects play a secondary role to the extent that speakers (a) were able to generalize from known examples to systematic principles and (b) showed an ability to form new categories.

8. Conclusions

The ultimate question is what these results suggest about the speech production system. We suggest that the dominant effects, which show systematic transfer of an existing allophone to a new phoneme, accord best with neo-generative models such as those discussed in Maye et al. (2008) and Peperkamp and Dupoux (2007). However, there are small lexical effects which recall the effects found by exemplar theorists (Goldinger, 1998, 2000; Johnson, 2006). Therefore, the total picture might be captured best in a hybrid model (Pierrehumbert, 2002).

Acknowledgements
We would like to acknowledge the support of a grant from the James S. McDonnell Foundation and an Andrew W. Mellon Postdoctoral Fellowship at Northwestern University. This article has
benefitted significantly from discussions with Matt Goldrick, Christine Meunier and Robert Espesser. We would like to thank Matt Bauer for generously providing the ultrasound images in Figure 1 and Figure 16.

References


