Reassignment of consonant allophones in rapid dialect acquisition

James Sneed German, Katy Carlson, Janet B. Pierrehumbert

Submitted to the Journal of Phonetics (in press)
Manuscript date: 8 February 2013

Title: Reassignment of consonant allophones in rapid dialect acquisition

James S. German[a*], Katy Carlson[b], and Janet B. Pierrehumbert [c]

*Corresponding author: jsgerman@ntu.edu.sg, Tel: +1 65 6592 1822, Fax: +1 65 6795 6525

[a]Nanyang Technological University, Division of Linguistics and Multilingual Studies, HSS 03-46, 14 Nanyang Drive, Singapore 637332
[b]Morehead State University, Department of English, 103 Combs, 150 University Boulevard, Morehead, KY 40351, USA
[c]Northwestern 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 allophones of /t/ and /r/, as the Glaswegian speaker aspirated word-medial /t/ but pronounced /r/ as a flap initially and medially. This experiment therefore explored (a) whether speakers could learn to reassign a sound they already produce (flap) to a different phoneme, and (b) whether they could learn to reliably produce aspirated /t/ in an unusual phonological context. Speakers appeared to learn systematically, as they could generalize to words which they had never heard the Glaswegian speaker pronounce. The pattern for /t/ was adopted and generalized with high overall reliability (96%). For flap, there was a mix of categorical learning, with the allophone simply switching to a different use, and parametric approximations of the "new" sound. The positional context was clearly important, as flaps were produced less successfully when word-initial. And although there was variety in success rates, all 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
allophone, flap, dialect, imitation, learning, rhotic, exemplar

Reassignment of consonant allophones in rapid dialect acquisition

1. Introduction

Ever since the critical period hypothesis raised questions related to late learning, there is growing evidence for late plasticity in the phonological/phonetic system. Various sociophonetic studies, for example, have shown dialect adaptation in adult speakers under natural conditions. Munro, Derwing, and Flege (1999) found that Canadians who had moved to Birmingham, Alabama partially acquired an American accent. Harrington, Palethorpe, and Watson (2000a, 2000b)'s acoustic analysis of 40 years of recorded Christmas broadcasts of Queen Elizabeth II showed that by the late 1980s, Her Majesty's pronunciation had shifted towards a more mainstream variety of RP. A post-hoc study by Sankoff (2004) of recordings made for the British documentary series *Seven Up* also found dialect adaptation by two speakers. Using controlled test materials, Evans and Iverson (2007) similarly showed that young adult speakers from the Midlands, U.K. exhibited shifts in vowel quality after attending university.

While such studies provide key evidence for plasticity in the phonetic and phonological system, the study we present was motivated by the need for diagnostic evidence about the cognitive architecture responsible for such adaptation. Specifically, we conducted a dialect imitation experiment in order to address four key issues suggested by prior work on second language learning and on learning of individual speaker traits:

1) *Lexical vs. systematic learning*: To what extent do subjects learn general phonological or phonetic patterns, which can transfer from specific words in the input to new words?
2) *Categorical vs. parametric learning*: To what extent do learners succeed by exploiting phonetic categories which they already know from their L1 (or D1, native dialect)? To what extent do they succeed by forming new phonetic categories over the parametric (i.e., continuous) phonetic space?
3) *Level of encoding*: Are new phonological patterns learned by substituting one phonemic representation for another, or do allophonic or positional variants have an independent role in the process? Specifically, are existing variants confined to their original D1 context, or can they be reassigned to a different context through modification of the encoding rules? Also, can existing variants of one phoneme be "recycled" to realize another phoneme?
4) *Persistent vs. short-term learning*: To the extent that speakers learn general phonological or phonetic patterns, do the effects persist beyond the period immediately after exposure?

1.1. Systematic and Categorical Learning

The literature on second language (L2) learning has emphasized systematic phonological and phonetic learning; dialect learning (D2 learning) should resemble L2 learning as it involves competition between the native phonological system and the novel system. A speaker's success in learning an L2 speech segment apparently depends on its exact relationship to segments in the L1 inventory. Two of the best known models, Best's Perceptual Assimilation Model (Best, McRoberts, & Goodell, 2001) and Flege's Speech Learning Model (1995), share 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 and 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 difficult for the learner to improve beyond initial rapid but partial success. If it is very distinct from all L1 phonemes (as Zulu clicks are for English speakers), there is much less interference, and the phoneme is a candidate for the kind of parametric learning involved in new category formation. This requires, among other things, that the learner begin to recognize a category based on continuous phonetic properties not usually attended to, and that a new articulatory pattern be implemented in a part of the phonetic space where the learner is unpracticed. The degree of success by adults in such learning would be indicative of the nature of phonetic plasticity that persists into adulthood.

Two studies by psycholinguists used 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 English dialect with categorically lowered target vowels. For example, the substitution of [ɛ] in *witch* yields *wetch*, a non-word in the base dialect. Subjects exposed to the novel dialect significantly increased their endorsement of modified forms as words in a lexical decision task. The effect of specific substitutions (e.g., [ɛ] for [ɪ]) generalized to new words, though the effect of relative lowering or raising did not generalize from front vowel substitutions to back vowel substitutions. 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, since the stimulus materials were created by categorical substitution of phonemes. Peperkamp and Dupoux (2007) used an artificial language learning paradigm to explore categorical feature neutralization in consonants. In their materials, voicing was contextually predictable for stops but not for fricatives, or vice versa. Their experiments also manipulated the degree of semantic support for the phonological patterns. Subjects were tested using a picture-pointing task. When word-learning was semantically supported, learning of the phonological constraint was efficient and generalized to new words.

Results such as those of Maye et al. and Peperkamp and Dupoux suggest a neo-generative architecture following the broad lines of Levelt (1980) as shown in Figure 1. The production system retrieves word forms from the lexicon, assembles the phonological code for the word forms in their phrasal context, and computes the phonetic implementation of the assembled phonological representation. The perception side is more or less analogous in the figure; the acoustic phonetic signal is phonologically parsed, and the phonological parse serves to access the lexicon. Various types of phonetic variability, including social variation, are treated as random noise that is ignored by the encoding rules. Thus, systematic effects of the type that Maye et al. and Peperkamp and Dupoux have demonstrated do not require any modification of the units in the coding level[1]; 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 novel language.

[1] An anonymous reviewer points out that Maye et al.'s result is also consistent with generalized gradient retuning of the perceptual space, given the lexical support for the modified vowels (since the targets were non-words otherwise). Since the materials involved substitution of one phoneme category for another, the study does not distinguish between these two possibilities, and we take category reassignment to be a straightforward account of the findings.

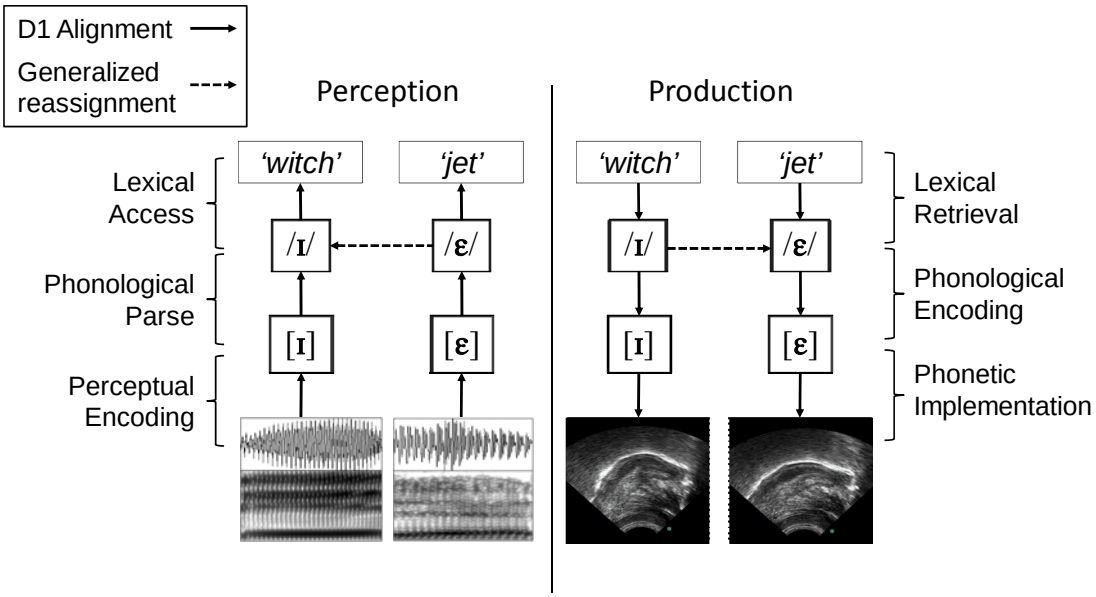

Figure 1. Minimal perception (left) and production (right) architecture consistent with categorical effects found by Maye et al. (2008) and Peperkamp and Dupoux (2007) [2]. Generalization occurs through realignment at the level of phonemic encoding (dashed arrows). The ultrasound images show the outline of the tongue during production of the vowels.

Strange (1995) noted that studies of the acquisition of L2 phonemes 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 units involved are phonemes in the classical sense (which retain their identity across variations in context), or less abstract, allophonic units. Studies of the acquisition of the /r/-/l/ distinction by Japanese learners of English (Mochizuki, 1981; Logan, Lively, & Pisoni, 1991) find that this contrast is much more difficult in some contexts than others, indicating that allophonic units are probably the relevant level of description. Similarly, 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 in perception. Polka (1991) explored whether experience with specific allophonic variants of /t/ in English (e.g., [t̪] as in *cartridge* and [t̪] as in *eighth*) would support the ability to distinguish them perceptually in Hindi, as compared to other sounds involving the same Hindi contrast which do not appear in English (e.g., [ɖʱ] and [d̪ʱ]). Indeed, the voiceless unaspirated sounds were distinguished more reliably, suggesting that the English phonetic system supports perception of the Hindi contrast in a way that is not predicted by the phoneme system alone.[3] If Strange is correct that the relevant units at the coding level are positional variants of phonemes (allophones) rather

[2] This model portrays only the aspects of a model needed to capture categorical realignment of the type found by Maye et al. (2008) and Peperkamp and Dupoux (2007). The arrows represent the overall direction of feeding ultimately needed to go from acoustic input to word-level representations. Certain details of encoding are not represented, including various top-down and expectation-based effects, such as those found by Harrington, Kleber & Reubold (2008), that feed counter to the direction of the arrows shown here.

[3] The comparison was made for all four Hindi voicing types. Polka's specific predictions about how the difficulty of the task would differ across all four pairs were not supported, though; she concludes that this was likely due to listeners' prior experience with stop variants of English dental fricatives ([d̪æt] for *that*).

than classical phonemes, then this raises the possibility that systematic learning in a model like that in Figure 1 may involve not only substitutions between phonemes, but also systematic realignments between positional variants and the lexicon. A learner should be able to adjust his or her coding system so that a particular variant of some phoneme may (i) be used outside of its usual phonological context or (ii) be reassigned as the realization of an entirely different phoneme.

The architecture outlined so far readily captures categorical, across-the-board effects. 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. Some words have more than one pronunciation. If subjects in an experiment memorized the new pronunciations for the training words as categorical alternatives, then the model would capture this by listing multiple word-forms for these words in the lexicon. A mixed situation, in which words used in training show an effect most reliably, but the effect also generalizes to new forms, can 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 justified.

1.2. Parametric Learning

A different architecture has been proposed by 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, like recognizing gender or sexual orientation. 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 a basic neo-generative architecture. 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).

Building on Goldinger's finding of imitation effects, several recent studies have established that speakers make gradient phonetic adjustments to speak more like a speaker they are exposed to. Schockley, Sabadini and Fowler (2004), for example, showed that speakers modified their voice onset times in word-initial stops during shadowing when those of the target speaker had been artificially lengthened or shortened. Similar results have been found for vowel formants (Tilsen 2009, Babel 2012) and F0

(Babel & Bulatov, 2011). Such findings support the relevance of phonetic detail in the adaptation that is typically associated with convergence phenomena, including accommodation (Giles & Coupland 1991, *inter alia*; Babel 2010), and a few recent studies have shown similar effects that cross dialects. In Delvaux and Soquet (2007), for example, participants heard ambient speech from a French regiolect different from their own (Liège vs. Brussels) during a word naming task, and showed gradient effects of vowel quality and vowel duration tending towards the pattern of the regiolect they heard. Babel (2010) showed that speakers of New Zealand English tended to converge with the vowel quality of an Australian speaker during shadowing, though this tendency was conditioned by social factors like the participants' implicit positive or negative attitudes towards Australia.

Such effects have fueled the rise of exemplar-based models of speech perception. These models assume that experiences of speech are stored in memory in considerable detail. Each memory can be indexed in multiple ways; a memory of the utterance [beɪbi] can be indexed as an example of the word *baby*, as an example of my mother'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. Since exemplar models explicitly provide for links between phonetic, lexical, and contextual variables, they readily capture word-specific phonetic effects and interactions between social variables and lexical access. By comparison, neo-generative models treat social variation as random noise that is ignored by the phonological parse, and therefore have difficulty explaining such effects.

However, models like MINERVA and XMOD, which do not explicitly encode segmental or positional information, encounter difficulties in explaining the extreme reliability of lexical access by human listeners under changes in speech rate or prosodic position. If lexical access is attempted from the parametric representations of entire words, alignment of the speech signal with the stored representations can be problematic. Reduction of segments early in a word, for example, can induce misalignment of the rest of the word with the stored representations. This can lead to a poor match, even in cases where aligning word subparts in the optimal way would have yielded a very good match[4]. This problem is noticeable in calculations using XMOD presented in Baker (2004). Clearly, this would be compounded when word recognition in connected speech is considered, and the issue highlights the importance of an abstract level of phonological encoding.

A further issue for exemplar models is the mechanism for speech production. Pierrehumbert (2001) starts from 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 (Fowler, 1986, 1990; Fowler & Rosenblum, 1990, 1991), 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, [bɹɑg] would average *bog*, *blog*, *frog*, *broad*, *brought*, etc., leading to

[4] If *ventilation* is reduced to a phonetic form like [vɛ̃lɛɪʃən], then [vɛ̃l] can provide a relatively good match for the first part of the stored representation *ven-*. In the absence of a syllable parse to correct for temporal misalignment, the attempted match between [ɛɪʃən] and the remainder of the stored representation (i.e., *-tilation*) will then be poor, even though it would be a good match for just the last part (i.e., *-ation*).

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*.

1.3. Hybrid Models

Such issues have led to the development of hybrid models, with some already reviewed in Goldinger (1998). Pierrehumbert (2002) adopts the neo-generative claim (see, for example, Levelt, 1980) that production of all words involves programming a categorical phonological representation, and that executing this plan is the only way to produce speech. 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 rise of exemplar models. A production plan for a specific phonological category is generated by sampling over existing exemplars of that category. This sampling is probabilistic, so very frequent patterns should have greater influence on the final target. It is also activation-weighted, so not only do very recent experiences have more influence than older ones, but specific words or social situations can influence phonetic realizations by biasing the selection of phonetic exemplars used as targets for phonological plans. Pierrehumbert argues that these biases are within phonetic categories, and they are therefore expected to be secondary to any categorical adjustments associated with specific lexical entries or modifications to the encoding rules.[5]

Such a hybrid model supports four different mechanisms for imitating a new accent. First, since individual words may have distinct phonological representations listed in the lexicon, the model provides for learning alternative pronunciations for known words, encoded using existing phonetic categories. Second, speakers can update their coding system through statistical generalization over known examples (of word-forms) in the lexicon. Thus, the model provides for learning of generalizations about these alternative pronunciations, encoded as generalizations about phonological representations. Since a new word-form can be learned from just a few examples, and generalization can proceed from just a few examples, learning under such a mechanism is expected to progress quickly in comparison with exemplar-based processes. Third, the exemplar component of the model provides for learning social, situational, contextual, and word-specific biases, realized as gradient differences within existing phonetic categories. Finally, the model provides for learning of new phonetic categories. This occurs as exemplars with a novel phonetic category index begin to accumulate in a specific region of the phonetic space, and can therefore be independently accessed for selecting a production target. We assume, following Best et al. (2001) and Flege (1995), that listeners can recognize certain sounds as distinct from those in the D1 inventory, and that this prompts them to introduce a new phonetic category index during perception and practice. The relative sparseness of the nascent exemplar cloud implies a large noise factor during sampling, predicting that implementation of a novel phonetic category should be subject to high phonetic variability until high levels of experience have been achieved.

While numerous studies have demonstrated exemplar effects in gradient, within-category changes, recent findings suggest a hybrid view more directly. Several studies (surveyed in Cutler, Eisner, McQueen, & Norris, 2010) have found that listeners adjust their perceptual boundaries between sounds after short exposures to speech that uses

[5] Similar interactions of phonological generalization with lexical items can also be captured in cascading connectionist models (Goldrick & Blumstein, 2006; Baese & Goldrick, 2009).

ambiguous sounds for one end of a continuum. For example, after hearing words that usually end in /f/ pronounced with a sound in between /f/ and /s/, listeners accept more s-like sounds as /f/ than they otherwise would. 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 one 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 that they had heard (so production did not change when perception did), though they could imitate the sounds when asked to.

Cutler et al. point 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. Ultimately, 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. This view is further supported by the findings of a Bayesian model simulation reported in Norris and McQueen (2008). In that study, word identification from phonetically atypical pronunciations was facilitated by even very small levels of experience with the "mispronounced" phonemes involved. The training data consisted of diphone-diphone confusions obtained from a listening study, and words containing pairings that were not instantiated in the training materials could not be identified unless all diphone confusions were assigned a non-zero prior probability. By comparison, for pairings that had at least one instantiation in the training materials, even those representing a very poor phonetic match (e.g., [pianti] for /kianti/ "chianti"), the word was reliably identified regardless of the minimum prior probabilities. This suggests that small levels of experience with a pattern may greatly facilitate a shift to that pattern, compared with patterns that are entirely novel.

Hay, Drager and Warren (2010) found differences between New Zealand listeners who do or do not have certain vowels merged after exposure to a dialect that preserves the distinction. Listeners with merged vowels showed a reduced ability to perceive the contrast compared to listeners with unmerged vowels. This can be explained if specific exemplars of words are stored but also linked to phoneme categories. For listeners with merged vowels, experience with the contrast led to phoneme-level data that was noisier and thus perception of the contrast was not aided unless more lexical processing was evoked. Sumner and Samuel (2009) studied the effects of speaker experience with respect to the 'r-dropping' of certain New York City dialects. In a set of word form priming and semantic priming tasks, New Yorkers who normally produce r-ful variants behaved similarly to those who produce r-less variants. In long-term repetition priming, however, the r-ful New Yorkers behaved more like speakers raised outside of New York, showing no priming for r-less variants. The authors suggest that because of their experience with r-less variants, the New York-raised r-producers are able to access the appropriate lexical entry during immediate processing, but abstract away from the variant pronunciation over time, possibly not storing the phonetic details in the same way as r-less New Yorkers.

At least one study supports a hybrid model in speech production. Nielsen (2011) showed that speakers exposed to lengthened VOTs of word-initial /p/ during word

shadowing produced longer VOTs for novel words beginning with both /p/ and /k/. The fact that such gradient effects of experience generalized beyond words in the input suggests an important role for abstract units. Additionally, the fact that the effect generalized to new sounds indicates that the size of the units involved are smaller than phonemes (i.e., sub-phonemic features).

Finally, Mitterer and Ernestus (2008), taking a position against a hybrid model, showed that Dutch speakers in a speeded shadowing task tended to produce the variant of /r/ (either alveolar or uvular) that matched the speaker they were shadowing, regardless of what their habitual pattern was. Crucially, they matched only the categorical aspects of the target speaker (i.e., place of articulation), but did not match the gradient within-category aspects of the targets (the timing of prevoicing), suggesting that the tendency to imitate was being mediated by an abstract level of representation in the perception-production loop. Jesse and McQueen (2011), however, show that experience-driven gradient retuning of perceptual boundaries along the /f/-/s/ continuum was restricted to non-word-initial position. Such gradient retuning effects are therefore likely to be lexically guided, and listeners may not encode sub-phonemic detail if lexical support for the phoneme category is not available at the time the sound is processed. Since the targets in Mitterer & Ernestus' study were all word-initial, it is possible that speakers simply were not able to remember enough detail about the target speaker's prevoicing to reproduce it accurately. Additionally, the speeded nature of the task may have reduced participants' ability to attend to subphonemic detail.

1.4. The Present Study

Pierrehumbert's model and other hybrid models exist on 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). Our experimental design allows us to locate the cognitive system with respect to this spectrum. Insofar as we find fast, systematic, categorical learning, we need key features of the neo-generative models. In contrast, pure exemplar models, with their epiphenomenal phonology deriving from a less abstract description of speech, require much larger amounts of experience and 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 the kind of detailed phonetic learning required for learning entirely new categories, as well as lexical, speaker-specific, and social effects that are now empirically well-documented.

To address these issues, we tested the ability of American English speakers to reproduce a novel dialect of English, namely Glaswegian English. The target sounds of interest were allophones of /t/ and /r/. For /t/, we were interested in the allophone that appears intervocalically under falling stress (as in the word *pretty*). This is usually a flap in American English, though sometimes it is aspirated (Zue & Laferriere, 1979; Fisher & Hirsch, 1976; Patterson & Connine, 2001). In the sample of Glaswegian English in our experimental materials, it is always aspirated. The challenge for our speakers was therefore to learn to recruit a rare, but familiar, variant of /t/. The Glaswegian /r/ was a flap in all positions. Since /r/ never appears as a flap in American English, participants needed to learn to produce an entirely unfamiliar realization of /r/. In the training phase, subjects heard each training sentence in Glaswegian English before reading it from a

printed list.[6] The training phase was immediately followed by a test for generalization to novel lexical items. Subjects were tested for further retention of the Glaswegian pattern a week later. The retention testing had three components: the original training set, the original generalization set, and a new generalization set.

If speakers can learn to transfer the patterns of the target dialect to words not in the training set, then learning must involve representations more abstract than words. We also explore the extent to which speakers exploit existing phonetic categories for the realization of patterns in D2 (i.e., [ɾ] for /r/), or begin forming a new phonetic category by trying to approximate known examples parametrically. To the extent that speakers make use of existing categories systematically, we can learn about the size of the units involved. If adaptation to D2 only involves modifying the relation of the phonological code (phonemes) to the lexicon, then recruited phonemes are expected to obey the same prosodic conditioning that they do in D1. Thus, if /t/ were to be substituted across the board for /r/, /r/ would be correctly realized as [ɾ] in word-medial position but as [t^h] in word-initial position. If, on the other hand, allophones can be produced outside of their D1 positions (i.e., [t^h] in word-medial positions, and [ɾ] in word-initial positions), then this suggests a model in which phonetic categories (allophones) are themselves abstract units that can be referenced independently by novel encoding rules. Given that [t^h] is sometimes used for medial /t/ in American English, learning of that pattern should progress more quickly than learning to produce [ɾ] for /r/. Finally, the comparison between performance immediately following learning and after one week provides an indication of the extent to which learning depends on the recency of exposure, and therefore the type of mechanism that is likely to be involved.

2. Background

2.1. Dialect Imitation

Several studies have explored conscious speech imitation from the perspective of voice impersonation, though these typically involve few speakers and the emphasis is on perceived similarity of the target and imitation (e.g., Markham, 1999; see Eriksson (2010) for an overview). At least two studies explored conscious imitation of dialect specifically. Van Dommelen, Holm and Koreman (2011) asked Norwegian speakers to speak with an accent different from their own based on a small speech sample, and found that they could match the pre-aspiration timing of the target dialect. Kim and de Jong (2007) studied the imitation of F0 contours for Korean speakers whose dialect either included (Kyungsang) or did not include (Cholla) lexical pitch accent. Kyungsang speakers responded with a categorical shift in their F0 pattern corresponding to their own perceptual category boundary, while Cholla speakers responded gradiently, reflecting the absence of a category distinction in their native phonological system. We are not aware of any study that explores categorical modification of the phonological system in conscious dialect imitation.

Most recent studies on plasticity in speech production are based on word shadowing or similar tasks (e.g., spoken word identification, Delvaux & Soquet 2007), in which the participants are instructed to say a word after an auditory prompt without being told to attend to dialectal or speaker-specific aspects of the word. The effects of exposure are

[6] Though the orthographic representation ultimately complicates our interpretation of the results, we found it necessary because the speech was potentially unintelligible without this support.

largely assumed to be unconscious and automatic. Nielsen (2011), however, argues against the automaticity of such effects on the basis of her finding that speakers imitated lengthened VOTs of English stops, but not shortened ones, suggesting that they were deliberately avoiding overlap with the voiced versions of those stops. This issue is developed more fully in Babel (2010, 2012), which show that phonetic convergence effects are sensitive to implicit social factors such as cultural bias (Babel 2010), gender of the listener, and the ethnicity and perceived attractiveness of the speaker (Babel 2012). On that basis, Babel argues that convergence effects must involve some combination of low-level automatic processes and socially guided processes.

By comparison, in our study we explicitly informed speakers that the target sentences were produced in another dialect, and we instructed them to try to imitate that dialect. The overall changes in speech observed during training and generalization trials are therefore straightforwardly interpretable as the result of a conscious effort. The primary behavior of interest is not *whether* our speakers modify their speech (as it generally is in word-shadowing tasks), but the extent to which they are successful, how rapidly they achieve success, and how any success is influenced by factors such as training (experience), time delay, and the relationship between the D1 and D2 phonological systems. Thus our study has more in common with perception studies like Maye et al. (2008), in which listeners heard speech involving a saliently atypical pattern and performed a task that required them to make systematic adjustments to their coding system. Maye et al. used a lexical decision task, though the measure was in fact off-line, since the main results were the lexical decisions themselves and not reaction times for correct responses. Since the lexical information of target words was readily recoverable from the story and sentence context, listeners could recognize that certain vowel phonemes were being pronounced differently in the experiment, and they adjusted the set of pronunciations they would consider as instances of words containing those phonemes.

2.2. American English flapping and /r/

Post-stress intervocalic /t/ is most frequently realized as a flap in conversational American English. Zue and Laferriere's (1979) production study found flapping of /t/ in 99% of post-stress intervocalic cases, while Fisher and Hirsh (1976) found from 36% to 97% flap production, as perhaps some subjects were speaking more formally than others. Patterson and Connine (2001) found that 94% of post-stress intervocalic /t/ in corpora of conversational speech were flapped, with lower levels of flapping in low-frequency and morphologically complex words. Steriade (2000), building on Withgott (1982), found that [tʰ] sometimes appears for intervocalic /t/ between two unstressed syllables, where phonologically [ɾ] would normally be expected. This occurred in certain derived contexts where /t/ is normally aspirated in the stem (e.g., [ˌmɪlətʰəˈɹɪstɪk], *militaristic* from [ˈmɪlɪˌtʰæɹi], *military*), and is accounted for in terms of paradigm uniformity.

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 usually being voiced and having an average duration of 20 ms, compared to voiceless stops in the same positions averaging 43 ms.

The normal realization of /r/ in American English is a voiced alveolar approximant [ɹ], which varies widely in its articulatory characteristics (Delattre & Freeman, 1968), but is often characterized by two general patterns involving either a somewhat retroflex tongue position or bunching of the tongue (Stevens, 1998; Ladefoged, 1993). In either

variety, this approximant appears on spectrograms with clear formants, smooth transitions from surrounding vowels, 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 intervocalically or elsewhere.

2.3. 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 the U.S. 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 because 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/s (Scobbie et al., 2006 list this as an older pronunciation).[7] 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. Additionally, Glaswegian English has different prosodic patterns, some of which were imitated by subjects (German, 2012).

3. Methods

3.1. Stimuli

The sound patterns under investigation appeared in four conditions, with /t/ and /r/ in both prosodically strong (pre-stress), word-initial positions and prosodically weak (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):

[7] An anonymous reviewer points out that not all Glaswegians use a flap for /r/, that this usage can vary with social class, and that flaps are more frequent after vowels. We acknowledge that there may be considerable variation in Glaswegian English accents which we do not explore in this paper, as we are focused on the speech of a single Glaswegian speaker.

(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 Glaswegian English speaker and put on CD. An additional group of three 12-item blocks was created and recorded for 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.

The lexical frequencies of the target words in the Celex2 database were collected for use in analyzing the results. They ranged from 0, for morphologically complex but transparent words like *unhittable* and rare words like *rhombus*, to 35,351 for the common word *time*. Words which did not appear in the database were considered to have a frequency of 0. The average frequency of /t/-initial words was 1478, for /t/-medials was 649, for /r/-initials was 693, and for /r/-medials was 672.

Due to an oversight during stimulus generation, a subset of the r-initial words occurred after words with final consonants instead of vowels. Thus, although /r/ was intervocalic in all r-medial words, this was not true for all of the r-initial words. There were 33 r-initial words with intervocalic /r/, and 15 with post-consonantal /r/. These subsets are analyzed together and then separately in the results. We would expect lower performance on production of non-intervocalic /r/ as a flap than the intervocalic /r/, because flaps are usually intervocalic in American English. Thus the phonetic routine for producing a flap would be more practiced in this environment.

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 productions of /r/ and /t/. We did not ask subjects to produce a baseline block of items in a Scottish or Glaswegian 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 could 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 Glaswegian 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 Glaswegian 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 1. Recording tasks by week. Tasks that share a row involve identical blocks for any given speaker. Blocks were counterbalanced to appear equally often in each task across speakers.

Week 1 (fixed order of tasks)	**Week 2** (rotating order of tasks)
Baseline	----
Training 1, Training 2 (with CD)	Training 3
Generalization 1	Generalization 1R
	Generalization 2
	Non-target (with CD, one block preceding each task above)

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.

3.3. Participants

There were a total of 43 participants in this study, all undergraduate students at Northwestern University enrolled in lower-division linguistics classes. They received course credit for their participation. Data from nine bilingual and non-native participants was excluded from analysis, as was that from three students who were unable to return for the second session. An additional seven students were excluded in order to correct for counterbalancing errors. The remaining 24 students used for the analysis 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. Acoustic Data Analysis

Each of the recorded sound files from participants was inspected and annotated by one of the first two authors, while both of the first two authors examined all of the Glaswegian English speaker's productions and a small set of evenly distributed participant files to assess intercoder agreement. Labelers listened to the target word of each sentence while examining the waveform and spectrogram using Praat (Boersma & Weenink, 2011). Initially, auditory, waveform, and spectrogram evidence were used to determine whether the target either (a) fell within the set of alveolar sounds targeted by the study (i.e., [t], [t^h], [ɹ] or [ɾ]), or (b) involved a place of articulation (e.g., velar) or manner of articulation (e.g., trill) not expected for the dialects involved. For tokens in the former group, if the acoustic evidence supported the presence of well-defined consonant boundaries (or edges), then the endpoints of the consonant were labeled. An example is shown in Figure 2. The point of voicing onset was also labeled if it differed from the end of the closure, as in Figure 3. For voiced sounds, F3 was measured by inspection at the point in or near the target where it reached a minimum. Consonant duration and voice onset time were later extracted automatically using Praat (Boersma & Weenink, 2011).

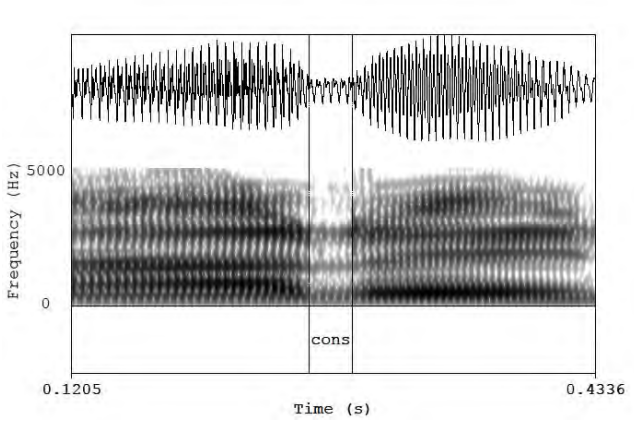

Figure 2. Example of an annotated token of medial /r/ (in "marriage") showing placement of consonant boundaries.

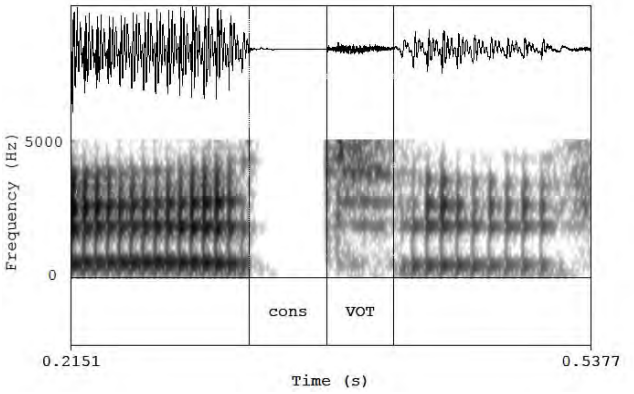

Figure 3. Example of an annotated token of medial /t/ (in "fetish") showing placement of consonant boundaries and the onset of voicing.

3.5. Categorization Procedure

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 a method based on acoustic evidence that decides, for each instance of /t/, whether it is produced as [t^h] or [ɾ], and for each instance of /r/, whether it is produced as [ɾ] or [ɹ].

For our analysis, we categorized as [t^h] any alveolar sound that included a voiceless closure and a delay in voicing onset. Since the unaspirated [t] allophone of /t/ is also voiceless with a short voice onset delay, this method potentially misclassifies [t] as [t^h]. Such errors are unlikely, however, since none of the targets included /t/ in a phonological environment associated with [t] in American English (e.g., following /s/ in an onset).

In our study, all targets that were voiced with clear consonantal edges were categorized as [ɾ]. Although this method potentially includes instances of [d], speakers in our study had access to the orthographic representations of the targets, which never included /d/ as the target phoneme. Additionally, Zue and Laferriere (1979) report a range of 10-70 ms for "flapped" /t/ in a falling stress context, and we compared the range and frequency distribution for consonant durations against those findings in order to assess whether [d] may have played a role.

A preliminary inspection of our data revealed that [ɾ] was sometimes produced without evidence of a full closure or acoustically well-defined consonantal boundaries, both in the Baseline American productions of medial /t/ and in the Glaswegian productions of /r/. Stone and Hamlet (1982) similarly reported '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). Since [ɹ] is also often realized without well-defined boundaries, some other measure was needed to distinguish between the two categories for those productions lacking such acoustic evidence. We used F3.

A widely recognized acoustic correlate of the American [ɹ] is a marked lowering of the third formant (Stevens, 1998), where [ɹ] is predicted to have a lower F3 than [ɾ]. However, since differences in vocal tract length among speakers lead to different overall formant distributions, the use of a single F3 threshold for deciding between [ɹ] and [ɾ] would result in substantial error. We therefore calculated a separate F3 threshold for each speaker based on his or her 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 [ɹ] and [ɾ] into two categories, such that the total number of errors (i.e., [ɹ]s categorized as [ɾ] plus [ɾ]s categorized as [ɹ]) is minimized. To obtain a scalar value for the threshold, we took the mean of the two data points surrounding the optimal cutpoint, following Yarnold and Soltysik (2005).

In the absence of detailed articulatory data, this method is an effective way to objectively classify outcomes while accounting for speaker variability. One consequence of the method, however, is that the F3 means of the resulting groups are predicted to be biased away from the center of the overall distribution, relative to the underlying population means. In fact, this is a property of any method that forces classification of tokens in the overlapping portion of the tails of two distributions. Thus the estimate of the mean F3 for [ɹ] is predicted to be too low relative to the baseline mean, and that for [ɾ]

to be too high. For this reason, consonant duration provides a more reliable way to compare categorized tokens against those in the baseline data.

In summary, our procedure initially used labeler inspection to classify productions according to whether or not they could broadly be considered one of the possible realizations of /t/ or /r/, namely [tʰ], [ɹ], [ɾ] or [t]. Productions that were determined not to be in this set were placed into a single category, which we refer to as "innovations". Productions within the set were further classified as [tʰ] if they had a voiceless closure and a positive VOT, and as [ɾ] if they were voiced and had clear consonantal edges (and possibly full closure). The remaining productions, having no clear consonantal edges, were classified as [ɹ] if the measured F3 was below the speaker-specific threshold and as [ɾ] otherwise. This method exhaustively classified all tokens in our study.

Finally, in order to assess the consistency of the categorization method across labelers, a series of analyses was performed on the classification results using Cohen's Unweighted Kappa. For the Glaswegian speaker, the entire set of productions was analyzed by both labelers and compared. For the participants' productions, an experimentally balanced and evenly distributed subset of the data (672 tokens taken from each task of each speaker) was labeled by both labelers. Agreement was found to be "excellent" to "nearly perfect" (see Section 4.2).

4. Results

The results of the categorization procedure are the crucial concern of this study and are presented in Section 4.3. Since that procedure ultimately depends on phonetic measurements, however, we first present a summary of the phonetic results in 4.1, followed by the results of an analysis addressing the reliability of the categorization procedure in 4.2.

4.1. Phonetic Summaries

The observed productions of /t/, based on acoustic examination, included voiceless alveolar consonants with evidence of closure followed by a voicing onset delay (suggesting [tʰ]), voiced alveolar consonants with short duration (suggesting [ɾ]), and a few other sounds. In cases where the speaker intended a different sound, as in the mispronunciation of the initial segment of *Thames* as [θ], the data were excluded.

The data in Table 2 show the percentage of /t/s with clear consonantal edges in the acoustic signal, as well as the durations of those consonants, voice onset times, and F3 data for voiced sounds. (The results for all imitation tasks are combined here because they had the same target sounds; they are analyzed separately in the categorization results.) The American subjects nearly always pronounced initial /t/ in the Baseline task with a long voiceless closure (averaging over 40 ms) followed by a voice onset delay averaging over 70 ms, consistent with previous findings for [tʰ] (e.g., Lisker & Abramson, 1967). The Glaswegian speaker's initial /t/s were similar, as were the imitated versions by American speakers in the Training and Generalization tasks.

Table 2. Summary of consonantal duration, VOT, and F3 minima for production of /t/ for native Glaswegian model, Baseline American, and imitation tasks.

	Initial /t/			Medial /t/		
Speaker/Trials	Glasweg.	Baseline Am.	Training/ Generaliz.	Glasweg.	Baseline Am.	Training/ Generaliz.
% of Trials with Consonantal Edges	100%	95%	97%	97%	87%	97%
Average Consonantal Duration, ms (SD)	53 (15)	43 (23)	57 (27)	35 (11)	23 (12)	55 (24)
% of Trials with Voicing Onset Delay	100%	99.7%	98%	100%	4%	96%
Average VOT, ms (SD)	70 (11)	74 (20)	70 (22)	71 (11)	----[a]	50 (18)
Average F3 minima, Hz (SD), females	NA	NA	NA	NA	2747 (263)	----
Average F3 minima, Hz (SD), males	NA	NA	NA	NA	2460 (185)	----

[a] 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.

Voiceless aspirated consonants with a slightly shorter average 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, shifting towards the Glaswegian dialect. Medial /t/ in the Baseline task was most often realized with a relatively short, voiced consonant with clear edges and visible F3, consistent with [ɾ], the expected American English allophone. The average duration was 23 ms, consistent with Zue and Laferriere's (1979) finding. Finally, some Baseline medial /t/s were produced with the voicing onset delay characteristic of [t^h], showing that aspiration in this position is occasionally produced naturally by these American English speakers.

The observed productions of /r/ were more varied, including voiceless alveolar closures with a short duration (suggesting [ɾ]), voiced alveolar sounds lacking evidence of closure (suggesting either [ɹ] or [ɾ]), trilled [r]s, and voiced uvular or velar fricatives (resembling [ʁ] or [ɣ]). Some participants produced a retroflex palato-alveolar fricative resembling [ʐ̻] and occasionally an [l]- or [w]-like sound. In other productions, the auditory evidence suggested a brief, flap-like closure, but the waveform and spectrogram showed an event which had a clear consonantal onset but a release too gradual for the end to be marked definitively.

The data in Table 3 show the average phonetic properties of /r/ productions. In the Baseline task, /r/ was almost exclusively produced with no evidence of consonantal edges or closure and with lowering of F3, consistent with normal American [ɹ] (Stevens, 1998). The majority of /r/s produced by the Glaswegian speaker had a short, voiced closure with little discernible dip in F3, consistent with [ɾ]. There were also some Glaswegian tokens lacking clear acoustic closure for initial and medial /r/, but these all resembled [ɾ]

auditorily. The Training and Generalization imitation tasks were where participants produced the largest variety of sounds for /r/. Clear consonantal edges or closure were present for less than half of the tokens for both initial and medial /r/. The consonantal duration means were quite short. For tokens with measurable formants, F3 minima exhibited a wide range of values.

Table 3. Summary of consonantal duration and F3 minima for production of /r/ for native Glaswegian model, Baseline American, and imitation tasks.

	Initial /r/			Medial /r/		
Speaker/Trials	Glasweg.	Baseline Am.	Training/ Generaliz.	Glasweg.	Baseline Am.	Training/ Generaliz.
% of Trials with Consonantal Edges	77%	3%	37%	90%	0%	44%
Average Consonantal Duration, ms (SD)	24 (13)	----[b]	24 (25)	15 (6)	----	19 (11)
Average F3 minima, Hz (SD), females	NA	1910 (202)	2073 (312)	NA	2110 (196)	2424 (336)
Average F3 minima, Hz (SD), males	1971 (216)	1610 (172)	1992 (300)	2123 (244)	1781 (146)	2163 (290)

[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.

4.2 Reliability

The reliability of the discriminant analysis based on F3 of tokens lacking consonantal edges was 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/.[8] The overall mean score for the Baseline productions is 0.97, with a standard deviation of 0.036, suggesting that the method is effective for distinguishing between [ɹ] and [ɾ].

The items analyzed by both labelers give an estimate of the reliability of the overall categorization procedure. 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 consonantal edges were present, though in all such cases they agreed that the phonetic category produced was [ɾ]. For the participant data, interlabeler reliability using four categories ([t^h], [ɹ], [ɾ] and "innovation") was found to be Kappa = 0.92 (95% confidence interval: 0.894, 0.946). Two sounds, [ɹ] and [ɾ], represent the largest source of interlabeler differences, accounting for 95% of all disagreements. Thus, a lower bound on inter-labeler reliability was estimated by considering only tokens involving /r/ in a non-baseline task. This was found to be Kappa

[8] The Glaswegian productions did not include [ɹ], so it is not possible to apply the method to those data.

= 0.83, 95% CI (0.763, 0.894), which is considered "excellent" or "nearly perfect" according to commonly cited guidelines (Landis & Koch, 1977; Fleiss, 1981).

The VOT for tokens classified as [tʰ] followed a single distribution with a median (58 ms) and interquartile range (43-76 ms) much higher than would be expected for [t], confirming our assumption that [t] was rare. Note that Lisker & Abramson (1967) found that nearly 10% of tokens for /t/ in a stressed context were produced with a VOT less than 25 ms, so it is not surprising that some of our speakers' tokens (3.3%) fall in that range, especially given the larger number of speakers in our study. The distribution for duration in [ɾ]-coded tokens is also largely consistent with previous findings. A small proportion of tokens (2.6%) had durations longer than the 70 ms upper range reported by Zue and Laferriere (1979), though again it is expected that the tails of the distribution would be extended in our study given the much larger number of speakers and tokens.

To further assess our procedure, we compared the consonant duration of imitated productions of /r/ categorized as [ɾ] against those flaps produced for medial /t/ in the Baseline task. The imitated flaps had a mean duration of 22 ms (SD = 6 ms) and the Baseline flaps a mean duration of 25 ms (SD = 8 ms). These very similar values suggest that the two groups of sounds belong to the same phonetic category, and indeed the difference between the durations was not fully significant in within-subjects and between-items ANOVAs ($F1(1,22) = 2.6$, $p = 0.124$ [one subject produced no measurable duration and was excluded]; $F2(1,142) = 3.9$, $p = 0.051$). As predicted, the mean F3 is higher for imitated [ɾ] (2897 Hz, SD = 328 Hz) than for baseline tokens (2616 Hz, SD = 303 Hz), likely due to the incidental removal of some tokens from the lower tail of the distribution. Overall, however, the phonetic characteristics of the categorized imitations suggest that participants were exploiting their knowledge of [ɾ] for producing /r/ in D2.

4.3 Categorization Results

The overall categorization results are shown first in Figure 4 and Figure 5, which display the percentage of Glaswegian-like outcomes for /t/ and /r/, respectively.

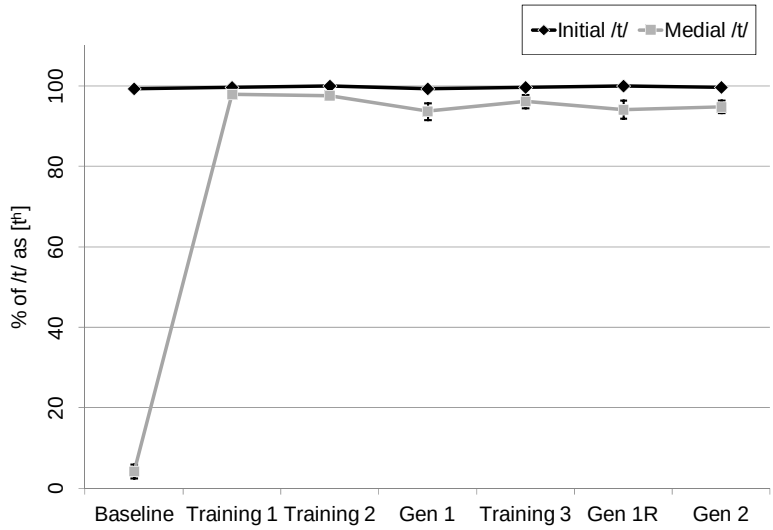

Figure 4. Mean percentage of [tʰ] outcomes by task for /t/ in word-initial and word-medial positions.

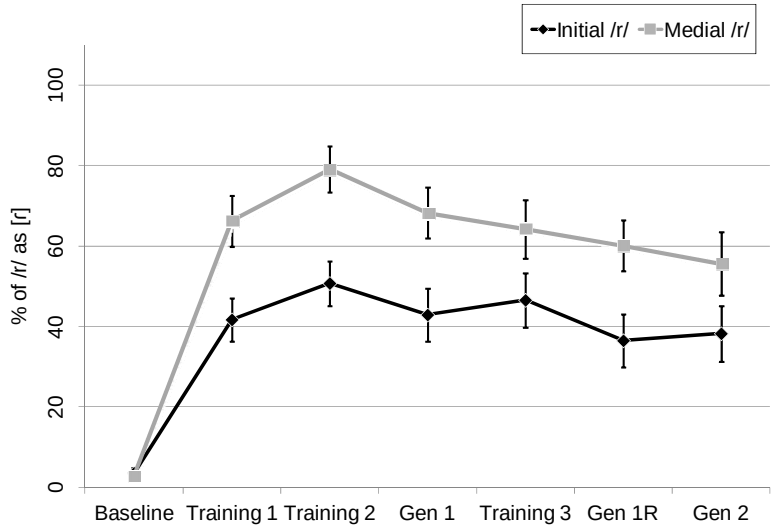

Figure 5. Mean percentage of [ɾ] outcomes by task for /r/ in word-initial and word-medial positions.

It is clear from Figure 4 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 initial Baseline condition at an average rate of over 95%. Consistent with previous findings, some of the speakers (8 out of 24) produced [t^h] here part of the time, including one who produced 33% of tokens as [t^h]. All speakers adjusted to producing aspirated medial /t/s in the imitation tasks.

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 learn to consistently produce the aspirated allophone in an environment where it only rarely occurs in D1. Speaker performance in the latter task was near ceiling, suggesting that speakers were able to exploit their previous experience with this pattern. The difference between baseline and imitation task performance was confirmed by simple one-factor within-subjects and within-items ANOVAs (see Table 4 below for statistics).

Table 4. Statistical difference between Baseline task and each imitation task; F-values shown, all p's < 0.001

	Task	T1	T2	Gen1	T3	Gen1R	Gen2
medial t	F1 (1,23)	2726	2203	1214	1670	1309	2152
	F2 (1,47)	4604	3593	2766	4218	2565	2777
initial r	F1 (1,23)	47	68	33	40	28	50
	F2 (1,47)	116	280	104	164	93	83
medial r	F1 (1,23)	113	197	115	79	56	50
	F2 (1,47)	335	737	353	294	342	171

The difference between the initial and medial /t/ conditions, though small, was significant in a between-items ANOVA with the two factors 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 exposure to and practice on specific lexical items, since performance was better in the Training tasks than in the Generalization 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, with Training performance higher than Generalization performance ($F1(1, 23) = 6$, $p < 0.05$). Neither analysis showed any significant effects of time, as participants' performance did not drop significantly in the second week, nor interactions of time with training on lexical items. Together, these results show that speakers learned to produce [t^h] in a rare prosodic position, and moreover, that they were able to quickly and robustly generalize that pattern to new words. Performance dropped off slightly after training, so subjects generalized imperfectly to new words, though only slightly. They retained this new pattern easily into the second week.

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 trilled /r/s. Nevertheless, all participants were able to produce [ɾ] for /r/ to some degree. Simple one-factor within-subjects and within-items ANOVAs showed that the percentage of flap productions was significantly higher in each imitation task than in the Baseline task for both initial /r/ and medial /r/ (see Table 4 above).The rest of the statistical discussion 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 Glaswegian speaker. An ANOVA on the percentage of flap production for /r/s in initial and medial positions in Training1 vs. Training2 was conducted; the factor of r-position was within-subjects but between-items, while the training 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 exposure to and 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 4 and Figure 5 clearly show that mean levels of performance during Week 2 did not fall back to Baseline American English levels, meaning that speakers largely retained the new patterns they had learned during the first week's training. Also, although performance in the Training tasks was better than in Generalization tasks, the mean Generalization results were still far above the mean Baseline results, showing extension of [ɾ] to new lexical items, both immediately and after a one-week time interval.

Because of counterbalancing, different subjects encountered the tasks in Week 2 in different orders. 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 exposure and practice affected second week performance, an analysis compared only the Training3 and Generalization2 results (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 during the second week for the specific lexical items which were trained in the first week, suggesting that adaptation involved a combination of both new word-form learning and generalization.

All of these tests have shown a strong effect of word-initial versus word-medial position for /r/. However, there were 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 6 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.

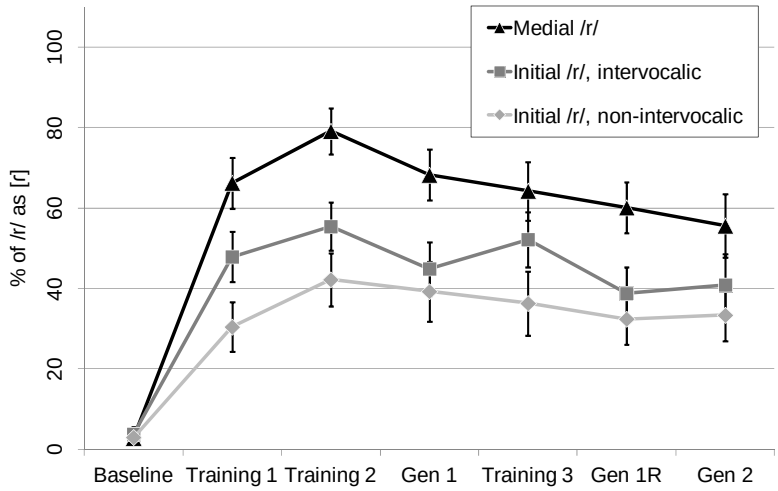

Figure 6. Mean percentage of flaps for /r/ items in word-initial position, intervocalic (33 items) vs. non-intervocalic (15 items), plus percentage for /r/ in word-medial positions.

The intervocalic set of /r/-initial items did show 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, Training3, and Generalization2 blocks ($F1(1, 23) = 16$, $p < 0.001$; $F2(1, 46) = 14$, $p < 0.001$). Nevertheless, similar ANOVAs on all 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 factor of training on lexical items remains significant in the analysis using only the intervocalic initial /r/ items, as the Training 2 and 3 blocks had higher rates of flapping than the Generalization1 and Generalization2 blocks ($F1(1,23) = 12$, $p < 0.005$; $F2(1,27) = 8$, $p = 0.005$).

Turning to word frequency, we included the Celex frequencies of the target words in a set of analyses by items to see whether frequency affected imitative success. The /t/-initial items could not be tested in this way due to insufficient variation in the results. For the /t/-medial items, an ANOVA including time, training, and frequency as a continuous covariate, over the Training2, Training3, Generalization1, and Generalization2 blocks, showed no frequency effect ($F2(1,46) = 0.47$, $p = .5$). The same test with /r/-initial items showed a similar lack of a significant effect ($F2(1,46) = 0.01$, $p > 0.9$). This test with /r/-medial items came closest to showing a significant frequency effect ($F2(1,46) = 3.76$, $p = 0.06$). Overall, though, lexical frequency did not seem to exert a reliable influence on the success of the allophonic reassignment. This is not surprising given the small size of the lexical (training) effect to start with, as any frequency effects would be inside that word-level variability.

In addition to completely non-adapted American responses, most subjects also produced phonetic innovations. These were sounds which shared some features of either

[ɹ] or [ɾ], but which were not intermediate to those sounds. Regardless of whether these represent attempts to approximate a new phonetic category parametrically (innovations), or failed attempts to produce known phonetic categories (due to the unusual phonetic environment), they involve sounds outside of the usual articulatory phonetic space for D1, and we treat them together. Some sounds in this group, such as [ʁ] and [ɣ], almost certainly represent innovations. If some others represent failed implementations of [ɾ] that had been successfully assigned to /r/, then this would only imply that the true rate of successful reassignment is underestimated in our results. Figure 7 shows the percentage of successful [ɾ] and of innovations for both /r/ positions (the level of success in the /t/ conditions meant that there were very few innovated or non-adapted responses).

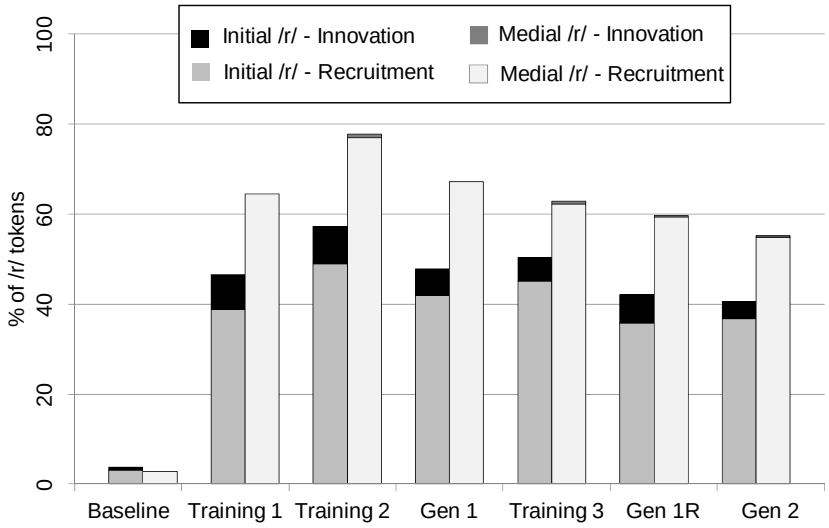

Figure 7. Mean percentage of [ɾ] recruitment and innovations, /r/ in word-medial and word-initial positions.

The proportion of innovated trials was highest for the /r/s in word-initial position and lowest for the /t/ conditions. Looking at innovations by subjects, we found that all subjects who produced 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 /r/ items were also examined. The rate of innovations for the non-intervocalic word-initial /r/s equaled or exceeded the rate of 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 apparent epenthesis of a short unstressed vowel. Most of the speakers, including even the Glaswegian speaker, used this strategy at least once during the experiment, possibly in order to place the /r/ in an intervocalic context.

5. 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 [t^h] for /t/ reliably in contexts where that phoneme is usually realized by [ɾ] in their native dialect, and all

speakers were able to produce some [ɾ]s in place of [ɹ] for the phoneme /r/. This learned ability was categorical since it involved a substitution of one sound in the D1 inventory for another. It was systematic in that it generalized to words not in the training materials, and it was fast, since robust learning occurred after a small number of examples (24 for each condition by the end of Training 2). In that sense, our main finding represents the production counterpart to perception results like those of Maye et al. (2008) and Peperkamp and Dupoux (2007), and reinforces the need for certain neogenerative features in the overall model of speech production.

Speakers in our study were able to produce existing sounds outside of their usual D1 contexts. This is not predicted by a model that only permits realignments at the level of phonemic encoding. Thus, our findings also show that sub-phonemic variants (allophones) are independent units whose role in phonological encoding can be systematically modified. Also, the learning that took place persisted over a period of one week with a slight, but significant decline. Thus, to the extent that speakers can modify their coding system, they can do so over a period longer than can be explained by short-term memory, and the system responsible for adaptation does not appear to be highly sensitive to the recency of exposure.

Speakers 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-medial positions where [ɾ] typically occurs in D1. Where /r/ was preceded by a consonant, this may have been partly due to a difference in articulatory difficulty, since the airflow required to produce [ɾ] was reduced in such cases. This is supported by the observation that even our Glaswegian English speaker occasionally introduced a very short epenthetic vowel before the word-initial flaps following a consonant, suggesting that the intervocalic environment was preferred for him as well.[9] Focusing on the intervocalic cases, however, the positional difference cannot be due to articulatory difficulty, since in American English flaps regularly occur intervocalically in certain word-final contexts (e.g., “at Anne’s”). Instead, the difference is perhaps best accounted for in terms of the speakers’ experience with [ɾ] in specific prosodic positions. 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 word-medial (and word-final) contexts across a large number of words, whereas they have no experience producing [ɾ] in word-initial position. The generalization that is most readily available to them, therefore, is for producing [ɾ] in medial (and final) 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 phonetic relationship between [tʰ] and [ɾ] as they occur in the imitative speech and the variants of those sounds in the subjects’ native dialect[10], the high degree of phonetic similarity in terms of closure duration, VOT, and

[9] A anonymous reviewer suggests that epenthesis before flaps in clusters or post-pausally may be common cross-linguistically, and notes that Baltazani and Nicolaidis (2011) report such effects for Greek.

[10] [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

F3, combined with the pattern of success across prosodic contexts, provides strong evidence that subjects were accessing the D1 sounds in order to imitate the D2 pattern.

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. Since our model assumes that phonological rules project from learned word-forms, it is expected that the combined effects of lexical learning and generalization will be greater than the effect of generalization alone. The model also assumes that learning such generalizations depends primarily on the robustness of the generalization among word-forms in the input, which was perfect in our materials. Any effects of lexical frequency are predicted to be smaller than the word-specific effects at best, and none were detected in our study.

The difference in performance between producing [tʰ] for medial /t/ and [ɾ] for /r/ is expected given that speakers already had some experience with the former pattern going into the experiment. The difference in performance could also be explained if there is a difference in the perceptual salience of the two patterns. At least one study, however, notes that variation in /r/ is a particularly strong dialect marker for English speakers in the U.K. (Llamas, 2010), which would tend to predict the opposite trend. Because of these differences, further research is needed to test whether it is more difficult to reassign a phonetic category to a different phoneme that to reassign it to a new position within the same phoneme.

As noted in Section 1.4, we found it necessary to provide subjects with orthographic transcriptions of the speech they were attempting to imitate. Unlike some other studies (Weber & Cutler, 2004), then, we were unable to deconfound the effects of orthography and categorical learning. There is 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). In the L2 literature, orthography is clearly activated during speech production. Kaushanskaya and Marian (2007), for example, found interference (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 related but not fully parallel to results of this type. Our results are entirely consistent, however, with the idea that learning at the categorical level was facilitated by knowledge of orthography. In other words, the presence of orthography probably enhanced speakers' ability to both access an intermediate (i.e., allophonic) level of representation and learn remapping relative to it. Our results nevertheless support the need for a model with two levels of representation (allophonic and exemplar), where learning can take place at each of the levels, and we leave it to future research to address 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 recruiting [ɾ] for the realization of /r/, subjects in our study realized /r/ with sounds not found in American English. Several subjects produced /r/ with some variant of a retroflex alveolar fricative [ʐ̠], and others with variants of [ʁ], [ɾ], and [ʒ]. This is accounted for in our model on the assumption that in some cases participants failed to assimilate the Glaswegian sound to [ɾ] in D1. This could have been due to slight

---

set aside the issue of the precise formal relationship between the two instances of [ɾ] (i.e., as a realization of /t/ in D1 and a realization of /r/ in D2), and continue to treat them as instances of the same allophone.

differences in the acoustic properties of the Glaswegian sound, which may have been perceived as salient by some participants and not by others. In such cases, participants attempted to implement the new phonetic variant based on the relatively small number of exemplars encountered during the study. The combination of sampling noise during the generation of a production plan and lack of articulatory practice explains why the outcomes were so variable.

Innovations of this type represent a larger proportion of all non-[ɹ] productions in prosodically strong, word-initial positions than in word-medial positions. On the one hand, this could have a perceptual basis: [ɾ] does not normally occur in word-initial positions in American English, and this contributed to a bias against assimilating the Glaswegian sound to the D1 phonetic category. As already mentioned, this type of result could also have an articulatory basis, and highlights a possible connection between the innovation data and the intervocalic/non-intervocalic data: American English speakers are only practiced at articulating a flap in a medial or final intervocalic position, and thus have difficulty producing it in any other environments. The latter explanation 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 pre-stress context (especially a post-consonantal context). Regardless of the cause of the non-[ɹ] tokens, the variety of sounds produced suggest that speakers were exploring their phonetic resources in different ways. 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 total picture is thus illustrated by Figure 8. Dotted arrows in the figure show how learning begins when individual lexical items become associated with alternative pronunciations independently of the encoding 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 phonetic (allophonic) categories. This can occur in one of two ways. An existing phonetic category may be assigned as the realization of a phoneme with which it has no association in D1 (dashed arrow), as when [ɾ] is used to realize /r/. Additionally, an existing variant of a phoneme, which may be rare in some contexts, is assigned as the realization of that phoneme with a higher probability in conjunction with some salient social contextual factor. In Figure 8, this can be viewed as a shift in the relative probability weights associated with the two solid arrows leading from /t/.

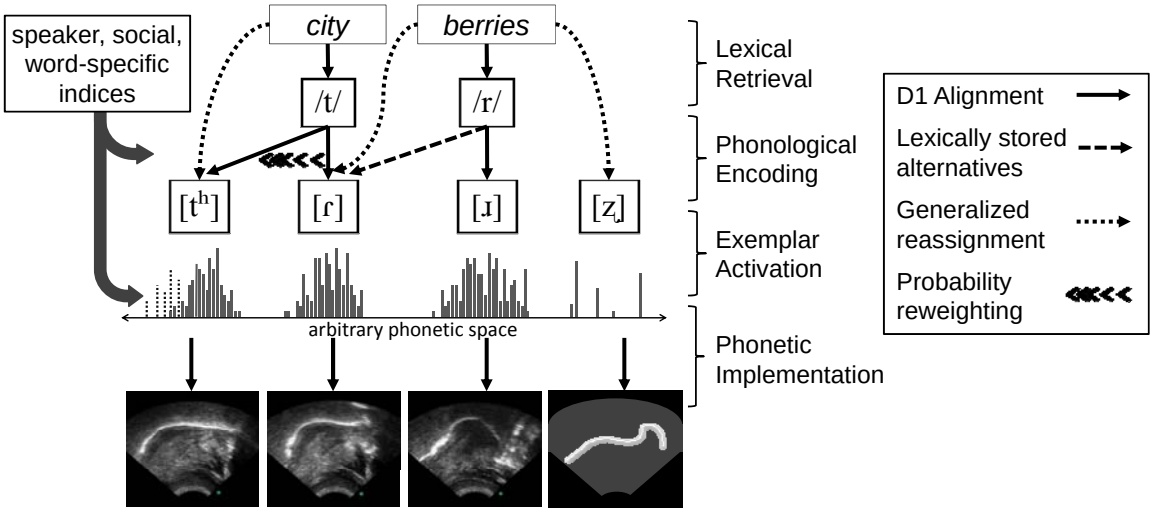

Figure 8. Proposed production architecture in which realignment occurs at the level of sub-phonemic (allophonic) encoding. Key features include word-by-word learning of new pronunciations (dotted arrows), generalization of novel mappings between phonemes and phonetic variants (dashed arrow), and shifting of probability weights associated with multiple existing D1 mappings («««). Speaker, social and word-specific factors influence productions by (i) shifting the probability weights of existing mappings, and (ii) activating exemplars that bias the phonetic distributions used to generate production plans. The images at the bottom show tongue position during production of the corresponding phones; the rightmost outline is hypothetical.

The model also supports the gradient within-category effects needed to capture the various speaker-specific, social, and word-specific biases that have been found in other studies. Phonetic categories induce activation of associated exemplars (shown as vertical bars), and phonetic implementation proceeds by weighted sampling over activated exemplars. Note that nascent categories (arbitrarily depicted as [z̥] in Figure 8) have sparse exemplar clouds, resulting in high variability during phonetic implementation. Exemplars associated with specific contextual indices may receive additional activation during phonetic implementation, which can influence production targets by shifting them away from the center of the exemplar cloud. This implies that individual words can influence production plans in at least four ways: (i) they can feed phonemic representations in the phonological parse, (ii) they can activate phonetic categories directly through lexically stored alternative pronunciations, (iii) they can influence the probability weights associated with multiple existing variants, and (iv) they can influence the production targets for phonetic categories once the phonological parse has been selected. The paths associated with (i) and (ii) are portrayed by arrows in our model. The paths for (iii) and (iv) are included with speaker and social biases, since such effects are equivalent in the model.

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/). One speaker, for example, whose imitations resembled Indian English, reported having had significant contact with the India-born mother of a childhood friend, whom she learned to imitate through practice. 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.

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 phonetic category with a new one which they were able to generate parametrically, or they activated a preexisting phonetic category in a novel (or rare) 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 three of the four mechanisms discussed in the introduction. To the extent that speakers in our experiment succeeded at replacing [ɹ] with the flap from their native dialect or from another language, they were able to modify their pronunciation of specific words using preexisting phonetic categories, and subsequently encoded these new pronunciations as generalized phonological principles. Those who succeeded by learning a novel articulation of /r/ demonstrated the ability to form new phonetic categories parametrically through exposure and practice. Speakers were also able to exploit their knowledge of [t^h] and their prior experience with that sound as a variant of /t/ in medial position. Thus, both phonemes conform to the model we are proposing here. In sum, our results show that systematic effects dominate the learning mechanism, though exemplar-based representations are needed in the model to capture parametric phenomena including new category learning as well as gradient within-category effects found in many other studies.

6. Conclusions

The ultimate question is what these results suggest about the speech production system. The current study demonstrated that speakers can modify their pronunciation through systematic transfer of an existing allophone to a new phoneme, or to a different phonological context within the same phoneme. This finding accords best with neo-generative models (Levelt 1980) such as those exemplified in the results of Maye et al. (2008) and Peperkamp and Dupoux (2007). Categorical and systematic findings of this type must be reconciled, however, with the gradient, within-category effects found in many other studies, which are best accounted for by exemplar models (Goldinger, 1998, 2000; Johnson, 2006). Since an exemplar component also straightforwardly provides a mechanism for new phonetic category formation, the total picture might be captured best in a hybrid model (Pierrehumbert, 2002).

Acknowledgements
We would especially like to thank our Glaswegian speaker, Alistair McGowan, for lending us his time and his voice. We would like to acknowledge the support of the James S. McDonnell Foundation (Award 21002061 to Northwestern University) 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 8.

References

Armbruster, T. E. (1978). *The psychological reality of the vowel shift and laxing rules*. PhD dissertation, University of California, Irvine.

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39,* 437-456.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40,* 177-189.

Babel, M. & Bulatov, D. (2011). The role of fundamental frequency in phonetic accommodation. *Language and Speech, 55(2),* 231-248.

Baese, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes, 24(4),* 527-554.

Baker, K. (2004). Auditory classification of regular and irregular pseudoverbs. Paper presented at McWOP 10, Northwestern University, Evanston, IL. Oct. 29, 2005.

Baltazani, M. & Nicolaidis, K. (2011). The many faces of /r/. Paper presented at R-atics 3, Free University of Bozen-Bolzano, Bolzano, Italy. 2-3 December 2011.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America,109(2),* 775-794.

Boersma, P., & Weenink, D. (2011). Praat: doing phonetics by computer [Computer program]. Version 5.3, retrieved November 1, 2011 from http://www.praat.org/.

Chirrey, D. (1999). Edinburgh: Descriptive material. In P. Foulkes & G. Docherty (Eds.), *Urban voices: Accent studies in the British Isles* (pp. 223-229). London: Arnold.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kuhnert, M d'Imperio, & N. Vallee (Eds.), *Laboratory Phonology 10* (pp. 91-111). Berlin: De Gruyter Mouton.

Delattre, P. & Freeman, D. (1968). A dialect study of American English R's by X-ray motion picture. *Language, 44,* 28-69.

Delvaux, V. & Soquet, A. (2007). The influence of ambient speech on adult speech production through unintentional imitation. *Phonetica, 64,* 143-173.

Edwards, J., Beckman, M. & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research, 47, 421-436.*

Erickson, A. (2010). The disguised voice: Imitating Accents or Speech Styles and Impersonating Individuals. In C. Llamas & D. Watt, *Language and Identities* (pp. 86-96). Edinburgh: Edinburgh University Press.

Evans, B. G. & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *Journal of the Acoustical Society of America, 121(6),* 3814-3826.
Fisher, W. M., & Hirsh, I. J. (1976). Intervocalic flapping in English. In *CLS 12-1*, 183-198. Chicago: Chicago Linguistic Society.
Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-linguistic research* (pp. 233-277). Timonium, MD: York Press.
Flege, J. E., Yeni-Komshian, G., & Liu, H. (1999). Age constraints on second language acquisition. *Journal of Memory and Language, 41*, 78-104.
Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley.
Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3-28.
Fowler, C.A. (1990). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America, 88*, 1236-1249.
Fowler, C.A., & Rosenblum, L. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 742-754.
Fowler, C.A., & Rosenblum, L. (1991). The perception of phonetics gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 33-59). Hillsdale, NJ: Erlbaum.
Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101(6),* 3728-3740.
Foulkes, P., & Docherty, G. J. (2000). Another chapter in the story of /r/: 'Labiodental' variants in British English. *Journal of Sociolinguistics, 4*, 30-59.
Fukaya, T., & Byrd, D. (2005). An articulatory examination of word-final flapping at phrase edges and interiors. *Journal of the International Phonetic Association*, 35, 45-58.
German, J. (2012). Dialect Adaptation and Two Dimensions of Tune. *Proceedings of Speech Prosody 6*.
Giles, H. & Coupland, N. (1991). *Language: Contexts and consequences*. Milton Keynes: Open University Press.
Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166-1183.
Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.
Goldinger, S. D. (2000). The role of perceptual episodes in lexical processing. In A. Cutler, J. M. McQueen, and R. Zondervan (Eds.), *Proceedings of SWAP (Spoken Word Access Processes)* (pp. 155-159). Nijmegen: Max Planck Institute for Psycholinguistics.
Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 152-162.

Goldrick, M., & Blumstein, S. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes, 21, 649-683.*

Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation,/u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *The Journal of the Acoustical Society of America, 123*, 2825-2835.

Harrington, J., Palethorpe, S., & Watson, C. I. (2000a). Does the Queen speak the Queen's English? *Nature, 408*, 927-928.

Harrington, J., Palethorpe, S., & Watson, C. I. (2000b). Monophthongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association, 30*, 63-78.

Hay, J., Drager, K. & Warren, P. (2010). Short-term exposure to one dialect affects processing of another. *Language and Speech, 53(4)*, 447-451.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411-428.

Jaeger, J. J. (1980). *Categorization in Phonology: an experimental approach*. PhD dissertation, University of California, Berkeley.

Jaeger, J. J. (1984). Assessing the psychological status of the Vowel Shift Rule. *Journal of Psycholinguistic Research, 13*, 13-36.

Jesse, A. & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin and Review, 18*, 943-950.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145-166). San Diego: Academic Press.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34*, 485-499.

Kaushanskaya, M., & Marian, V. (2007). Non-target language recognition and interference: Evidence from eye-tracking and picture naming. *Language Learning, 57*, 119-163.

Kim, J. & de Jong, K. (2007). Perception and production in pitch accent system of Korean. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1273-1276.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition, 107*, 54-81.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review, 13*, 262-268.

Ladefoged, P. (1993). *A course in phonetics* (3rd ed.). Fort Worth: Harcourt Brace Jovanovich.

Llamas, C. (2010). Convergence and divergence across a national border. *Language and Identities*, 227-236.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lawson, E., Stuart-Smith, J., & Scobbie, J. (2008). Articulatory insights into language variation and change: Preliminary findings from an ultrasound study of derhoticization in Scottish English. *University of Pennsylvania Working Papers in Linguistics: Selected Papers from NWAV 36, 14*, 100-110.

Levelt, W. J. M. (1980). *Speaking*. Cambridge, MA: MIT Press.

Lisker, L., & Abramson, A.S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech, 10,* 1-28.
Logan, J., Lively, S. & Pisoni, D. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89(2),* 874-886.
Markham, D. (1999). Listeners and disguised voices: The imitation and perception of dialectal accent. *Forensic Linguistics, 6(2),* 1350-1771.
Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science, 32, 543–562.*
Mendoza-Denton, N., Hay, J., & Jannedy, S. (2003). Probabilistic sociolinguistics: Beyond variable rules. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 97-138). Cambridge, MA: MIT Press.
Mitterer, H. & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109,* 168-173.
Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics, 9,* 283-303.
Munro, M. J., Derwing, T. M., & Flege, J. E. (1999). Canadians in Alabama: A perceptual study of dialect acquisition in adults. *Journal of Phonetics, 27,* 385-403.
Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research, 44,* 778-792.
Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39,* 132-142.
Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review, 115*(2), 357.
Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5,* 42-46.
Palmeri, T. J., Goldinger, S. D. & Pisoni, D. B. (1993). Episodic encoding of speaker's voice and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19,* 309-328.
Patterson, D. & Connine, C. M. (2001). Variant frequency in flap production. *Phonetica, 58,* 254-275.
Peperkamp, S. & Dupoux, E. (2007). Learning the mapping from surface to underlying representations in an artificial language. In J. Cole & J. Hualde (Eds.) *Laboratory Phonology 9* (pp. 315-338). Berlin: Mouton de Gruyter.
Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137-157). Amsterdam: John Benjamins.
Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 101-139). Berlin: Walter de Gruyter.
Pierrehumbert, J. B. (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 177-228). Cambridge, MA: MIT Press.
Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. Doherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, segment, prosody* (pp. 90-117). Cambridge: Cambridge University Press.
Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29,* 191-215.

Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *Journal of the Acoustical Society of America, 89(6),* 2961-2977.
Purcell, E. T., & Suter, R. W. (1980). Predictors of pronunciation accuracy: A reexamination. *Language Learning, 30,* 271-287.
Sankoff, G. (2004). Adolescents, young adults and the critical period: Two case studies from *Seven Up*. In C. Fought (Ed.), *Sociolinguistic variation: Critical reflections* (pp. 121-139). New York: Oxford University Press.
Schockley, K., Sabadini, L. & Fowler, C. A. (2004). Imitation in shadowing words. *Perception and Psychophysics, 66(3),* 422-429.
Scobbie, J. M., Gordeeva, O. B., & Matthews, B. (2006). Acquisition of Scottish English phonology: An overview. *QMUC Speech Science Research Centre Working Paper WP-7,* 3-30.
Steinberg, D. D., & Krohn, R. K. (1975). The psychological validity of Chomsky and Halle's Vowel Shift Rule. In E. F. K. Koerner (Ed.), *The transformational-generative paradigm and modern linguistic theory* (pp. 233-259). Amsterdam: John Benjamins.
Stevens, K. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.
Steriade, D. (2000). Paradigm uniformity and the phonetics-phonology boundary. *Papers in laboratory phonology, 5,* 313-334.
Stone, M. & Hamlet, S. (1982). Variation in jaw and tongue gestures observed during the production of unstressed /d/s and flaps. *Journal of Phonetics, 10,* 401–415.
Strange, W. (1995). Cross-linguistic studies of speech perception: A historical review. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-linguistic research* (pp. 3-48). Timonium, MD: York Press.
Stuart-Smith, J. (1999). Glasgow: Accent and voice quality. In P. Folkes & G. Doherty (Eds.), *Urban voices: Accent studies in the British Isles* (pp. 203-222). London: Arnold.
Stuart-Smith, J. (2007). A sociophonetic investigation of postvocalic /r/ in Glaswegian adolescents. *Proceedings of the 16th International Congress of Phonetic Sciences,* 1449-1452.
Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language, 60,* 487-501.
Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning, 41,* 177-204.
Tilsen, S. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics, 37,* 276-296.
Van Dommelen, W. A., Holm, S. & Koreman, J. (2011). Dialectal feature imitation in Norwegian. *Proceedings of the 17th International Congress of Phonetic Sciences,* 599-602.
Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language, 50,* 1-25.
Whalen, D. H., Best, C. T., & Irwin, J. R. (1997). Lexical effects in the perception and production of American English /p/ allophones. *Journal of Phonetics, 25,* 501-528.
Withgott, M. M. (1982). *Segmental evidence for phonological constituents* (Doctoral dissertation, University of Texas at Austin).
Yarnold, P. R., & Soltysik, R. C. (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, D.C.: American Psychological Association.
Zue, V. W., & Laferriere, M. (1979). Acoustic study of medial /t,d/ in American English. *Journal of the Acoustical Society of America, 66(4),* 1039-1050.

Appendix 1: Target items

Block A

1. The class does yoga on the matting.
2. Some day, he will find some courage.
3. The hill-dwelling monks can be seen building a temple.
4. The baby consumed a bowl of rice.
5. I suppose the illness caused his delirium.
6. He does the job, though he sounds funny when he talks.
7. We will need the long rope.
8. The family's chubbiness was mainly genetic.
9. Chuck always goes on the ferry.
10. Place the hassock inside the room.
11. The slimy animal by the pool is a toad.
12. The damp wind made him all sweaty.
13. The castle was held by a rebel.
14. My niece likes playing with Tonkas.
15. Leah was planning a vacation in Florence.
16. The yelling of the fans was muted.
17. The memo was funny because of a typo.
18. By the end of the movie, love was found by the heiress.
19. Deep in the woods was an old cottage.
20. He smoothed the edges with a rasp.
21. Chicago has a famous marathon.
22. This essay will be done on time.
23. The peace negotiation was plagued with racism.
24. The ball had seemed unhittable.
25. The candy bin is full of toffee.
26. The kids sang a silly rhyme.
27. How many books can you carry?
28. No one in the family believed Uncle Bob was batty.
29. An unlucky buck has a wide rack.
30. Jonathon likes milk in his porridge.
31. Good things come in twos.
32. He's thinking of the beans he's been eating.
33. The pancakes could be good with syrup.
34. The chalice was classified as a relic.
35. The swamp was the location of a big battle.
36. Though a snake, the python is tame.
37. Jack landed the salmon in a riffle.
38. The window glass was held in with putty.
39. In college, you buy books by the ton.
40. Bad shampoo can be made with oranges.
41. In the zoo, he saw a lonely rhino.
42. By one a.m., they deployed the shuttle.
43. The boy swallowed mud because he was curious.
44. Lyme disease is often blamed on ticks.
45. Ned's love of walking could be called fanatic.
46. Civil though she may be, his feelings could be ruffled.
47. The shah was in a fury.
48. The maid needs help with this task.

Block B

1. With heavy use, the cloth became a rag.
2. I believe the fish of the day is whiting.
3. The valley is unlivably arid.
4. I seldom see Melanie in town.
5. The log was the home of a raven.
6. A nice pie will be made with the berries.
7. The cook slowly made the beef patties.
8. He gave away his only token.
9. Then he skillfully sings an aria.
10. The village was enslaved by the Romans.
11. Selma's clothing was always fashionable and exotic.
12. Bush has an obsessive love of low taxes.
13. The couple enjoyed choosing a ring.
14. The passage of this bill is vital.
15. The new Nanolab will have unique tools.
16. In the old days, you soothed a baby with marrow.
17. The chess club held one final meeting.
18. Fax me a copy of his resume.
19. The boss came in a toga.
20. The message lacks an obvious moral.
21. The young couple should speak with a rabbi.
22. Someone should clean the tiles.
23. The flu is caused by a virus.
24. The gossip in the school was awfully petty.
25. The small flying thing is a wren.
26. In the evening, he munches cereal.
27. The navy loaned him a tank.
28. The news channel mentioned a UFO sighting.
29. Jacques lives cheaply in Paris.
30. This wood will be used in making a table.
31. He was deafened by the rifle.
32. The bed was below the folds of netting.
33. The essay should have a specific topic.
34. The sickly youth has no endurance.
35. Leo saw something askew in the rhombus.
36. Len's business office was inside the city.
37. He did fax them one query.
38. These bananas look ripe.
39. In the field I found a Mayan fetish.
40. You spoke slowly on the tape.
41. His whole life, Jack had been in a hurry.
42. Excess cleavage in an office is unsuitable.
43. I saw Andy in the hall with his twin.
44. The small child won the race.
45. The lamb seemed happy, though amazingly little.
46. The infection began in his tonsils.
47. Picasso designed this epic mural.
48. A Chicago dog always comes with relish.

Block C

1. On the weekends, old men walk along the Thames.
2. The sheep dog was lying by the rock.
3. Mrs. Jackson came up with a new theory.
4. The cheese by the olives is feta.
5. In the necklace was a humongous ruby.
6. The judge scolded the jury.
7. The ocean has both low and high tides.
8. The thief thinks she can escape all notice.
9. The policeman sounded his siren.
10. The ad was awfully racy.
11. Happiness is only fleeting.
12. The small dog was wagging his tail.
13. Climbing in the Himalayas involves many risks.
14. Insulation is made with batting.
15. Good fishing begins with good tackle.
16. She's in Mexico, climbing a Mayan pyramid.
17. A guinea pig is amazingly pettable.
18. The house had become an old ruin.
19. Galahad was anxious when he was in peril.
20. With his pencil, he keeps an ongoing tally.
21. Clownfish and sea anemones live on the reef.
22. Well, the man has had some experience.
23. Life was no fun among the Ottomans.
24. Five people play on the team.
25. The milk was sold in the dairy.
26. This salad needs six kinds of lettuce.
27. Along the lake, the couple cycled in tandem.
28. The slugs will avoid the roses.
29. The company sells useless insurance.
30. The flaw is on the tip.
31. The leak in the hull was sealed with a special resin.
32. The old donkey was given a heavy beating.
33. The diva was accompanied by a full chorus.
34. This season, the high-heeled shoe is all the rage.
35. He was killed by an unknown toxin.
36. Old wigs belong in an attic.
37. The king has no loyal men in the realm.
38. The coffee will keep Jan cozy inside the tollbooth.
39. This cloud looks like a cirrus.
40. In five days, the blooms will lose some petals.
41. The nebula is visible with a telescope.
42. The milkshake was done up with a cherry.
43. The consul said they’ve been invited.
44. How do you like the new rug?
45. In Vilnius, you can buy amazingly spicy curry.
46. A missing copy was shown by all the tags.
47. The dog could become rabid.
48. The canoes should be inflatable.

Block D

1. His language was so foul, only one line was quotable.
2. My dad has many worries.
3. He lived his life by the wisdom of the Talmud.
4. Seafood gives me a rash.
5. Ms. Jones gave examples of Eskimo-Viking borrowings.
6. When you sneeze, please use a tissue.
7. The clouds opened up and she saw a rainbow.
8. The usefulness of this device is debatable.
9. A sunny vacation leaves you looking tan.
10. Will they have a good marriage?
11. All the family's belongings lay beneath the rubble.
12. With a knife and some wood, you could whittle.
13. Do you think Sheila saw the heron?
14. The possum climbed up on the roof.
15. People in confined spaces can become catty.
16. As a hobby, Kim plays the timpani.
17. The police chief made the mob leave the area.
18. How many novels have you completed?
19. His cap held a long, golden tassel.
20. Sue could think of a good reason.
21. The special comes with pita.
22. The young amphibian became a tadpole.
23. Mike was amazed when he won the raffle.
24. Somehow she can deal with his snoring.
25. Sam found a casino and began betting.
26. His only companion is an unspeaking wrasse.
27. Even with my glasses, my vision is blurry.
28. Melvin hailed a taxi.
29. In the oven, she is baking some rolls.
30. How do hyenas find carrion?
31. The woman smiled with pity.
32. The café gave him a choice of teas.
33. I think Kim and Mike sound serious.
34. You always pick wrong.
35. The biology class was discussing a beetle.
36. Did one of the halfbacks pull a tendon?
37. The chess game was won with the rook.
38. The whale is a mammal and aquatic.
39. Business is done on the telephone.
40. This ice cube has a funny appearance.
41. We will film the movie in a jungle setting.
42. He sold me a ribbon.
43. Becky enjoys chewy candy like taffy.
44. The lilac bush is beside the sorrel.
45. We'll need his decision on the new road.
46. He yelled when he chewed his tongue.
47. The guinea pig sleeps in a nice burrow.
48. The ocean waves pounded the jetty.

Appendix 2: Non-target items

Block 1

1. A display of the dig can be seen in the lobby.
2. Dolphins swim and play alongside the ship.
3. May I buy some chicken feed?
4. In the piano division, the champion was Michael Hawley.
5. The clock face glows dimly in the evening.
6. The second copy of the code has many bugs.
7. If the ball bounces on this wall, then the game ends.
8. Simply place the apple on the napkin with a bow.
9. Food supply in the developing nations should be closely followed.
10. Some books will always be appealing.
11. Some music could soothe the savage babies.
12. Olga was hoping the food would be well-done.

Block 2

1. Why should he push himself, when he has all the money he can use?
2. The only way one could fail his class is by sleeping when he gives the quizzes.
3. The sheep in the field fell down in the wind.
4. Seven men will be assigned these five offices.
5. Melanie's solution is a classic in the field.
6. Physics labs will be open by noon in the fall.
7. The Buck company will spend a million and fix the building.
8. The koala seems so lovable and sleepy.
9. She finally gave up on being queen of all the lands.
10. Of this class, only one will be successful.
11. His mind shows signs of senile decay.
12. When five decades have gone by, you'll need a new guide.

Block 3

1. The glass was oozing a luminous fluid.
2. The gel was molding in the shape of a buffalo.
3. She found a bag of cash in the subway.
4. Globs of muck fell along the sides.
5. Louis loves the sound of moaning voices.
6. Pam only knows osmosis, so she failed the biology exam.
7. The evil demon came in a puff of smoke.
8. The hail in Spain falls mainly on the mesas.
9. The invoice displayed the shipping and handling fees.
10. The film was shown in the evening.
11. The young musk oxen can be pleasingly affable.
12. Winning the game would be a bonus.