Cole, J. and S. Shattuck-Hufnagel. 2018. Quantifying phonetic variation: Landmark labeling of imitated utterances. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, M. Zellers Eds. *Rethinking Reduction*, Mouton de Gruyter, pp. 164-204.

Jennifer Cole and Stefanie Shattuck-Hufnagel

6 Quantifying phonetic variation: Landmark labelling of imitated utterances

Abstract: Speech is known to be highly variable across speakers and situations, and listeners pay attention to some of this phonetic detail for the rich contextual information it carries. In this chapter we introduce a method for investigating phonetic variation from the dual perspectives of perception and production. We analyse serial imitations of a heard utterance, where the linguistic object to be produced is fixed syntactically, lexically and prosodically, and employ a novel method for quantifying phonetic variation using acoustic landmarks (LMs) (Stevens 2002) as correlates of phonologically contrastive manner features. Imitated utterances produced by ten native speakers of American English resulted in 3,500+ consonant and vowel LMs, which were labelled and compared both to the lexically specified LMs, and to the LMs as produced in the stimulus. We report five main observations from this exploratory study: (1) Phonetic reduction due to variation in LM realization occurs even in the highly constrained imitation task; (2) variation is asymmetric across classes of LMs: Vowel LMs seldom vary, while glide LMs are most vulnerable; (3) certain patterns of LM deletion were very frequent in our data, but no pattern of phonetic variation prevailed over all imitated instances across or within speakers; (4) phonetically reduced forms in the stimulus, identified in terms of LMs, are not reliably imitated; (5) about 20% of lexically predicted LMs are produced with variable outcomes, both within speakers (across repetitions) and across speakers. These findings demonstrate and quantify systematicity in phonetic reduction as measured in terms of LMs. They also reveal that speakers exercise choice in phonetic implementation, deviating both from lexical targets and from the phonetic detail of the heard stimulus. These results hold promise for the use of imitated speech in the study of phonetic variation, and for the use of LMs (and by extension other feature cues) as a phonologically grounded measure of variation in speech production.

Keywords: phonetic reduction, phonetic variation, landmarks, imitation.

Jennifer Cole, University of Illinois at Urbana-Champaign and Northwestern University
Stefanie Shattuck-Hufnagel, Massachusetts Institute of Technology

https://doi.org/10.1515/9783110524178-006

6.1 Introduction

Research over the past few decades provides mounting evidence of systematic and contextually governed phonetic variation in continuous speech. This variation, which is non-contrastive in the lexical sense, arises due to coarticulation with adjacent segments (Cole et al. 2010; Farnetani and Recasens 1997) or to prosodic structure (e.g. Cho 2005; Choi et al. 2007; Turk and Shattuck-Hufnagel 2007), and there is also variation in the cues to prosodic features themselves (Dilley, Shattuck-Hufnagel, and Ostendorf 1996; Mo 2011; see also Cole 2015). Phonetic variation can take the form of reduction, strengthening or other kinds of pronunciation change, and has long been seen as a driving force in sound change.

AU: 'Choi et al. 2007' is cited in the text, but missing in reference list. Kindly include in the list or remove it from text.

Recent studies have shown the ability of language users to hear, learn and make use of these systematic context-driven and speaker-specific patterns. Evidence for this is seen in phenomena such as the facilitation effect of a familiar talker's voice on word recognition (e.g. Goldinger 1998), phonetic convergence between interlocutors (e.g. Pardo 2006; Pardo et al. 2012), and the perceptual "retuning" of phoneme category boundaries based on auditory exposure to acoustically ambiguous stimuli (Norris, McQueen, and Cutler 2003). Studies of perceptual learning further demonstrate that listeners can learn to associate specific patterns of segmental phonetic variation, synthetically created, to the voice of an individual talker (e.g. Allen and Miller 2004; Eisner and McQueen 2005; Kraljic and Samuels 2005, 2007). Kraljic et al. (2008) argue that perceptual learning of this sort may be critical in enabling listeners to differentially accommodate variation that is idiosyncratic to an individual talker and also the more systematic patterns that characterize dialectal variation; Cutler (2008, 2010) argues that such accommodation is accomplished by reference to abstract phonemic segments in the lexicon, and contributes to the efficiency of lexical access.

In light of these findings, a comprehensive model of phonetic variation must not only provide an inventory of the types of variation that can occur and the contexts in which each type is licensed, but it must also take into account how individual speakers attend to, store and make use of variable phonetic patterns. In this study we explore a new methodology that addresses two challenges for developing such a model, the first concerning data collection and the second concerning measurement of phonetic variation.

6.1.1 Data collection through elicited imitation

Contextual factors play a significant role in conditioning phonetic variation, so in order to discover systematic patterns in variation it is desirable to have multiple

instances of the same lexical items (defining the production targets) in the same contexts, and produced by numerous speakers. For this we use an elicitation task of repeated imitation that offers substantial control over the linguistic object produced by the speaker. Prior studies using imitation and the related task of speech shadowing show that speakers converge to the sub-phonemic detail of heard speech, (Goldinger 1998; Pardo 2013), with measurable effects on, e.g. VOT and vowel formants (Babel 2012; Neilson 2011; Shockley, Sabadini, and Fowler 2004). There is some evidence to suggest that imitation is limited to phonetic detail that cues phonological contrast (Mitterer and Ernestus 2008). The findings from these studies predict that imitators will reproduce phonetically reduced forms that they hear, especially when the reduction affects cues to phonologically contrastive features.

AU: Please define VOT.

Alternative methods to elicited imitation are not as suitable for investigating phonetic reduction. In studies that rely on the production of written stimulus materials, it is not easy to control the prosodic structure (which is known to influence surface phonetics); in studies that rely on corpora of spontaneous speech, it is not easy to control the lexical and syntactic content to make direct comparisons across speakers possible. In contrast, the imitation task severely constrains the syntactic, lexical and phonological (i.e. segmental and prosodic) shape of the utterance, and this means that the effects of these factors are relatively consistent across speakers and across imitations by a single speaker (for evidence of consistency in prosodic imitation see Section 6.2.2). Thus, any consistent patterns of phonetic reduction/variation produced in this task can be understood as patterns that are favoured by the language in those contexts. Although a complete inventory of such processes is well beyond the scope of this exploratory study, if results are promising, it will motivate future expansion of the method, to provide a comprehensive inventory of the nature and scope of surface phonetic variation.

6.1.2 Measuring phonetic variation with landmarks

To model systematic patterns of phonetic variation, including reduction, we need a measure that captures variation in the mapping between discrete lexically contrastive units (e.g. phonemes) and continuous-valued acoustic parameters. As observed by Pardo (2013) in her study of phonetic entrainment, it's not an easy task to identify raw acoustic measures that capture both what is heard by listeners and the phonetic adjustments controlled by speakers. We think the solution lies not in raw acoustic measures, but in a measure that is more directly related to units involved in speech processing, for example, phonological units. Here, we explore the use of landmarks (LMs), as proposed by Stevens (2002), as a

quantifiable, acoustic-phonetic metric that captures variation in the realization of cues to phonological manner features at a finer level of detail than the symbolic allophone, however narrowly defined, permits. LMs provide a way to discretize information from acoustic measures as cues to phonologically contrastive manner features.

In the rest of this section we introduce LMs and expand on the reasons why we expect LMs to be appropriate measure of phonetic variation. Definitions of the LMs used in this study, with examples from our speech data, and the methods for LM labelling are presented in Section 6.2.

We use LMs as the units for measuring pronunciation variation, rather than the symbolic allophone, based on the proposal of Stevens (2002) that individual acoustic cues to contrastive features, rather than symbolic allophones, are significant units of representation in human speech processing. Stevens proposed that the first step in the processing of a perceived utterance by a human listener is the detection and identification of LMs, i.e. the abrupt acoustic discontinuities associated with consonant closures and releases, as well as intensity minima and maxima in glides and vowels, respectively. LMs are a particular class of feature cues which signal information about one class of contrastive phonological features (i.e. the articulator-free features, after Halle 1992, which roughly correspond to the manner features). In this framework, LMs (like other feature cues) are not raw acoustic measures, but are derived from acoustic measures; they are acoustic edges or inflection points, i.e. events which require comparison across multiple measurement values. The LMs used in this study, adopted without modification from Stevens' proposal, mark the acoustic expression of the closure and release of consonantal constrictions for plosives, affricates, fricatives and nasals (e.g. stop-closure, stop-release), the energy valley for glides, and the energy peak for vowels. These LMs are further described, with illustrative examples, in Section 6.2.3.

As an illustration, consider the LM representation for the word peak (from our database) in its unreduced (full) form. The lexical specification of this word identifies the phoneme sequence of the unreduced form as /piːk/. The initial and final consonants are plosives, which have two LMs, one marking the abrupt intensity drop across a range of frequencies corresponding to the onset of the closure interval and the other marking the abrupt intensity spike marking the onset of stop release noise. The vowel has a single LM marking an intensity maximum. Thus, the LM sequence for this word consists of five LMs: stop-closure, stop-release, V, stop-closure, stop-release, which specifies an unreduced CVC structure. LMs are particularly informative acoustic events for listeners, since they not only signal the identity of or changes in manner (providing an initial estimate of the CV structure of an utterance), but also identify regions that are rich in cues to the voicing and place features, such as formant transitions and release-burst spectra.

AU: Please define CVC and CV.

Stevens' proposal was originally concerned with individual feature cues in perceptual processing; here we begin to explore the possibility of extending it to the task of speech transcription (and by implication to speech production). By annotating imitations of the target utterances in terms of LMs, we lay the groundwork for testing the hypothesis that individual feature cues are an appropriate vocabulary for capturing patterns of context-driven surface phonetic variation. That is, LMs may constitute a level of description that links the abstract symbolic specification of lexical items (i.e. in terms of features that define phonemic manner categories) to the continuous-valued variation in the speech signal (i.e. in terms of quantitative parameter values for the cues to manner features). We emphasize here that LMs by themselves will not capture all information about phonetic variation, nor are they the only acoustic cues to inform speech processing. Acoustic cues to voicing and place features, and other spectral information not captured by LMs will also be informative, as will the specific parameter values for the cues, but here we restrict our focus to the presence vs. absence of LMs as cues to manner features.

Labelling LMs (and, eventually, acoustic cues to other kinds of phonological features) offers several advantages over positional allophones for capturing the type of systematic context-driven phonetic variation that has become evident in detailed acoustic-phonetic studies of large corpora, and that experimental studies have indicated are under speaker control and attended to by listeners. For example, individual cues to a given feature are sometimes omitted or added independently, leaving other predicted cues to the features of a target sound segment intact, as when a sequence of two stop consonants is produced without the release burst for C1 and without a closure LM for C2, or when a final /t/ is produced with both glottalization and a release burst. Similarly, a speaker may omit the LM cues to a stop coda, but retain the duration cues to its voicing in the duration of the preceding vowel. Segmental transcription requires a binary decision as to whether a segment was included in the surface form of the utterance or not; cue-based labelling permits a more fine-grained annotation which can capture the fact that some cues may remain to the features of an apparently "deleted" segment. Niebuhr and Kohler (2011) have described such phenomena as the "phonetic residue" of apparent segment deletion processes.) LMs (and other feature cues) can also capture detailed (and potentially significant) differences among tokens within an allophonic category. For example, the allophonic category "flap" is applied in American English to a wide range of tokens, from a very short-closure /t/ with clear acoustic evidence for a closure and release burst, to a small glide-like dip in amplitude in a voiced region, with or without a small release burst (due to some build-up of pressure behind the incomplete constriction). If we want to determine whether these variations within an

AU: 'Neibuhr and Kohler' is cited in the text, but missing in reference list. Kindly include in the list or remove it from text.

AU: An opening parenthesis preceding the text "apparent segment deletion processes.)" seems to be missing. Please check, and correct if necessary.

allophonic category are perceptible, learnable and reproducible by language users, it is useful to have a labelling system which captures them. Individual feature cue labelling also permits the capture of temporal asynchronies among feature cues in the signal, as when frication noise for a voiceless fricative begins before the voicing for a preceding vowel ends, or when the velum opens to create a nasal formant for a coda nasal, somewhere in the preceding vowel. Transcription using sequences of symbols, no matter how detailed and narrowly defined, require the annotator to determine where in the signal the acoustic implementation of one symbol ends and the implementation of the following symbol begins; as practitioners of phonetic labelling are only too well aware, this requirement is often impracticable. That is, in many cases the various cues to a feature (or to the segment that the features define) are spread in time, so that they overlap with cues to adjacent segments (as when the duration of a vowel correlates with the [voice] feature of a following coda consonant) or they are limited in time, so that they do not extend throughout the region that a labeller must designate as corresponding to the relevant phonetic symbol (as when vocal fold vibration is limited to just a few pulses at the beginning of the frication noise associated with a voiced fricative). By labelling individual cues, such asynchronies can be captured and studied for their systematicity, with potentially profound implications for the types of acoustic-phonetic information that are represented and controlled by language users.

LMs and other feature cues are also more amenable to fine-grained quantification than allophonic categories are. For example, it may be difficult to use allophonic symbols to specify the sense in which two speakers' voice onset times become more similar during a conversation, since these changes are typically sub-phonemic (Neilson 2011). But in a transcription system based on individual feature cues, quantitative specification of cue values will be natural and precise; to the degree that such transcriptional analyses reveal systematic control by speakers, it will open the door to the development and testing of speech processing models that incorporate representations of individual cues to contrastive features. Finally, LM transcription provides a simple way of quantifying certain aspects of variation: counting the number and type of LMs that are modified from the lexically predicted pattern (or in the case of imitation studies, from the heard stimulus) allows a straightforward comparison between utterances. In this study, we restrict ourselves to labelling LM cues, because this class of feature cues is particularly robust. But if LM-based transcription emerges as a useful tool for capturing some of the systematic phonetic variation produced by speakers in an imitation task, it will serve as the basis for developing an approach to speech analysis that is more robustly and extensively based on individual acoustic cues to phonological features, and their parameter values.

6.1.3 Research questions

In this study we pose a number of specific questions about the LM behaviour of the speakers in our small sample:

Q1: *Variability in LM outcomes*: Does phonetic variation, as measured by LM modification, occur even in the highly constrained imitation task? If so, what type of modification is most common (e.g. deletion, substitution or insertion)?

Q2: *Variability by LM class*: Are different types of LMs, representing different manner classes, differentially likely to be modified?

Q3: *Between-speaker variation*: Are some patterns of LM modification (e.g. in specific words) consistently produced across speakers? If so, what are the phonological environments that most frequently condition variable outcomes?

Q4: *Accuracy in imitation vs. realization of lexical target*: Do speakers differ in the accuracy with which they realize lexically specified LMs? Do they differ in the accuracy of imitation? What happens when the target of imitation differs from the lexically specified target?

Q5: *Within-speaker variation*: Are speakers internally consistent in the way they realize a LM in a given phonological context, producing an individual "phonetic signature" in terms of preferred patterns of phonetic reduction?

We emphasize that this is an initial exploration, undertaken to evaluate the viability of combining imitated elicitation and LM analysis as a measure of certain aspects of variability and reduction. The domain of the study is restricted to 60 utterances by 10 speakers, i.e. 3 imitations per speaker of 2 target utterances, but, as shown below, the resulting 3,502 LM annotations provide a window into the contexts in which phonetic variation occurs, the nature of that variation and the insights that LM annotation can provide into the processes of that underlie it. Thus, the results serve as an initial demonstration of the usefulness of LM labelling as a tool for the quantitative comparison of the phonetic similarities and differences between utterances of the same phonemically specified sentences.

A final comment on terminology is in order here. While we are broadly interested in patterns of phonetic variation as measured by LMs, the findings presented below reveal that the most common patterns of variation involve the loss of a lexically predicted LM, i.e. LM reduction. Other patterns show substitutions of lexically predicted LMs, which, like the examples of LM loss, often result in the partial or complete loss of information about the manner class of phonemes specified in the unreduced lexical form of a word. In what follows we use the terms

"reduction" and "variation" interchangeably in referring to variable outcomes in the imitation data. Distinguishing between these terms will necessitate further work measuring the degree to which lexically specified phonological information is recoverable for the listener.

6.2 Methods

6.2.1 Imitation experiment

Stimuli: Target utterances for the imitation task were drawn from the American English Map Task Corpus of task-driven spontaneous speech (Shattuck-Hufnagel and Veilleux 2007). This corpus was collected using the Map Task elicitation method, described in Anderson et al. (1991). In this speech elicitation task, two speakers (one the instruction giver, the other the instruction receiver) are each furnished with a map; the instruction giver's map shows a path through the items pictured on the map, and the instruction giver is asked to guide the instruction receiver through the task of reproducing the path on the receivers map which shows no path. The two maps differ slightly in the geographical items shown, but this fact is initially unknown to the participants, since neither participant can see the other's map; this manipulation introduces just enough complexity into the task so that the two speakers soon become absorbed in solving the problem and begin speaking in a very natural manner. The resulting speech exhibits the kinds of surface phonetic modification of word forms that occurs widely in natural speaking situations, but is otherwise more difficult to elicit in controlled conditions of laboratory recording which afford the opportunity for the highest-quality acoustic recording and pre-specification of target lexical items.

Thirty-two utterances from 4 of the 16 dialogues in the AEMT were selected for the imitation task; all 4 of these dialogues concerned the same pair of maps. Eight utterances from the instruction giver were selected from the middle portion of each dialogue. The extracted utterances were 7–15 words long (average length 11.5 words), and were chosen to minimize disfluent intervals and laughter. Data from imitations of two of the target utterances are presented here:

AU: Please define AEMT.

Utterance 1
Um Kate d'you see the Canadian Paradise?

Utterance 2
Um you're gonna be standing at the peak of the mountain on the Canadian Paradise.

These two were selected to represent a short and long utterance, and had a minimum of lexical substitutions and disfluent imitations relative to some of the other utterances in the full data set. Both utterances begin with *um* ending in a mid-level pitch plateau that marks a fluent continuation into the following phrase. These *um*s were included in the stimulus utterances for the analysis of prosodic imitation, a part of the larger project for which these data were elicited, but which is not reported here. Note that the orthographic rendering of these utterances reflects three contracted elements: *d'you*, *you're*, and *gonna*. LMs for these items are discussed below (Section 6.2.3.1). These two utterances, unlike most of the others in the data set, have a common word sequence as well, *Canadian Paradise*, which allows us a small opportunity to look at variation for lexical items across sentence contexts.

Participants: The imitated speech analysed here was recorded from 10 female speakers (18–25 years old) recruited from the student body at the University of Illinois, and paid $10 for participating in this study. The restriction to young female participants was intentional, since the speech to be imitated was taken from dialogues between young female speakers of similar age range. All participants were speakers of the Midland dialect area of American English, and reported no history of speech or hearing deficits.

Procedure: Participants were seated in a quiet room where they received brief instructions from the experimenter and provided written consent prior to the start of the experiment. Participants were equipped with a head-mounted cardioid microphone (AKG C520) and headphones. Target utterances were presented to participants in auditory form through the headphones, with no accompanying text presentation. Participants were told they would be reproducing utterances recorded from a dialogue, and the nature of the Map Task was briefly described to provide context for the dialogue excerpts they would be imitating. The experimenter instructed participants to reproduce each utterance by "repeating the words and the way the utterance was said". Participants listened first to an example utterance to get familiarized with the speech materials, and then proceeded to the imitation task. The auditory stimulus was presented three times in succession with a 2-s pause between presentations. Participants were instructed to reproduce the utterance three times in succession immediately following the three auditory presentations, for a total of 96 imitated productions per subject (32 utterances × 3 repetitions).[1] The timing of the repetitions and the speech rate

1 The intention of the instructions was that participants would reproduce not only the lexical and syntactic content of the stimuli, but also the prosody and other pronunciation qualities representative of the speech style, such as speech rate. The word "imitation" was not used in the

were produced by the participant without instruction. Experiment sessions lasted about 30 minutes. Imitated productions were recorded through a head-mounted microphone (AKG-C520) onto a Marantz solid-state digital recorder, and later transferred to computer for processing and analysis. </IP>

6.2.2 Prosodic annotation

Impressionistically, the imitated utterances achieved the spontaneous speech style of the stimuli, and were in fact very hard to distinguish from the set of original productions of the Map Task speakers. To evaluate the extent to which the prosody of the imitated utterances was a match to the prosody of the stimulus utterances, an agreement analysis was conducted on prosodic labels assigned to both stimulus and imitated utterances. The stimuli were prosodically labelled for pitch accents and prosodic boundaries using the full ToBI transcription system (Silverman, et al. 1990). Imitated utterances were prosodically labelled for the location of pitch accents and prosodic boundaries (using the labels "A" and "B"), but without annotation of tonal melodies, and treating intonational and intermediate phrase boundaries as alike. A comparison of prosodic labels between the stimulus utterance and the third imitated production was performed. This comparison using the third imitation rather than the first or second was considered to be a more conservative test of prosodic imitation, on the grounds that the auditory record of the stimulus utterance would be more remote in short-term auditory memory, or not present at all, so a match in prosodic features with the imitated utterance should reflect the cognitive representation of those features in the mind of the imitator.

AU: Please advise if the word "achieved" should read "resembled" in the text "imitated utterances achieved the spontaneous speech style of the stimuli" for readability.

Cohen's kappa scores for pitch accent and boundary were calculated as the agreement metric for a subset of six imitators. This statistic measures observed agreement against expected agreement, taking into account the frequency of each label. Kappa scores range from 0 (no agreement) to 1 (perfect agreement), and the scores for stimulus-imitation agreement are in the range of 0.61–0.71 for the location of pitch accent, and between 0.6 and 0.7 for the location of prosodic phrase boundaries. These represent substantial agreement according to the common interpretation of this statistic. The kappa scores are in the same range as has been reported for trained transcribers doing a ToBI-style "A" and "B" annotation of a

instruction, to avoid the suggestion that participants should attempt to reproduce pitch range or other aspects of the stimulus speaker's voice that reflect physical characteristics of the speaker rather than linguistic or communicative features of the speech.

similar genre of American English spontaneous speech, with kappa scores of 0.75 for accent and ~0.65 for boundaries (Yoon et al. 2004). Further details of the prosodic annotation, agreement analysis and phonetic measures of prosodic similarity are reported in Cole and Shattuck-Hufnagel (2011) and Mixdorff et al. (2012).

6.2.3 Landmark labelling

The acoustic-phonetic labelling scheme employed in this study was designed to capture the ways in which the predicted LMs, as well as the LMs produced by the speakers of the stimulus utterances, were implemented in the productions of the imitators. We define the predicted LMs to be those that derive from the lexically specified phonemes, i.e. the contrastive segmental units of the full, unreduced pronunciation of the word. For the data analysed here, the lexically predicted LMs were identified by the authors (native speakers of American English) based on their understanding of English phonology and familiarity with the words in this sample.

A further comment is in order here regarding the status of unreduced pronunciations. In using the unreduced form as the reference form against which variable, reduced pronunciations are measured, we do not claim that full, unreduced forms are the *only* kinds of representation encoded by language users, or even that they are the forms that are the most likely to be produced in a given context. Frequent patterns of reduction may be encoded, for example the intervocalic flapped /t/ in *butter*, or deletion of the medial unstressed vowel deletion in *fam(i)ly*. But to the extent that the unreduced pronunciation is possible, perhaps associated with certain conditions of speech style or rate (e.g. extremely clear speech), we hold that it has a privileged status as the form which links all potential productions of a word, including both reduced and strengthened forms. Exemplars on their own do not capture the systematic relationships between surface forms, nor do they capture relationships between exemplars that generalize across lexemes. We maintain that the unreduced form must be available and identifiable as such, even in theories that propose a lexicon defined over clusters of phonetically detailed exemplars. As discussed below, our findings lend some support to this view, as a reduced word in the stimulus is sometimes restored to its full, unreduced form in imitation.

To carry out LM labelling, we used criteria that have been developed in a LM labelling project at the Speech Communication Group at MIT (Shattuck-Hufnagel and Veilleux 2007), based on the ideas of Stevens (2002). In this approach, LMs are initially defined in terms of the acoustic characteristics of a segment (consonant or vowel) in its canonical context. In this sense, “canonical” is defined as

the form that a LM takes when it occurs in its most definitive context. For consonants, the canonical context is between two full (in English, stressed) vowels, as for the closure and release LMs for /b/ in /aba/ or for /m/ in /imi/. (We use the terms "closure" and "release" to designate the acoustic outcome of the articulatory events which cause them; this close mapping between acoustic events and their articulatory causes is an important aspect of Stevens' (2002) proposal. Results of experiments perturbing the acoustic-perceptual consequences of a speaker's articulatory configurations (Villacorta, Perkell, and Guenther 2007) support the view that, despite this close mapping, the targets of speech production are acoustic in nature. Consonantal stops, fricatives and nasals are predicted to have two LMs, i.e. one created at the moment of formation of the oral constriction and one at the oral release, while affricates have three, i.e. one generated at the moment of constriction, one at the partial release of closure into a configuration that produces frication noise and one at the final release of that constriction. Examples are illustrated in Figure 6.1, panels a–c. In contrast, canonical vowel segments are produced with just one LM, which represents the acoustic consequences of the maximum opening of the vocal tract, i.e. when the vocal tract cross-sectional area is greatest (example in Figure 6.1, panel d). Canonical intervocalic glides are produced with a single minimum opening occurring when the vocal tract is the most constricted, i.e. has the smallest cross-sectional area, and the glide LM marks the valley of the corresponding dip in acoustic energy (example in Figure 6.1, panel e). The string of predicted LMs for an utterance is derived from the string of phonemes that define each word in the lexicon.

AU: A closing parenthesis following the text '(We use the terms "closure" and "release" to designate' seems to be missing. Please check, and correct if necessary.

6.2.3.1 Predicted and observed LMs for the stimulus utterances

Figures A1 and A2 in the appendix display the phonemes, the predicted LMs for the full, unreduced form for each word in the two target utterances, and the LMs and prosodic features that were realized in the utterances as they were produced by the Map Task speaker and labelled by the authors. As already noted, the orthographic rendering of these utterances reflects three contracted elements: *d'you*, *you're* and *gonna*. These contractions exist in the language as reductions from full forms (*do you*, *you are* and *going to*), but we allow the possibility that the reduced forms are the lexical targets for contractions such as these that have a conventionalized spelling. Thus, we establish the lexically predicted LMs for these items based on the contracted forms, not the corresponding full forms.

There are 40 predicted LMs in Utterance 1, and 82 in Utterance 2, making a total of 122 predicted LMs (see Figures A1 and A2 in appendix). These LMs comprise the lexically specified targets for the imitation task, and are predicted to

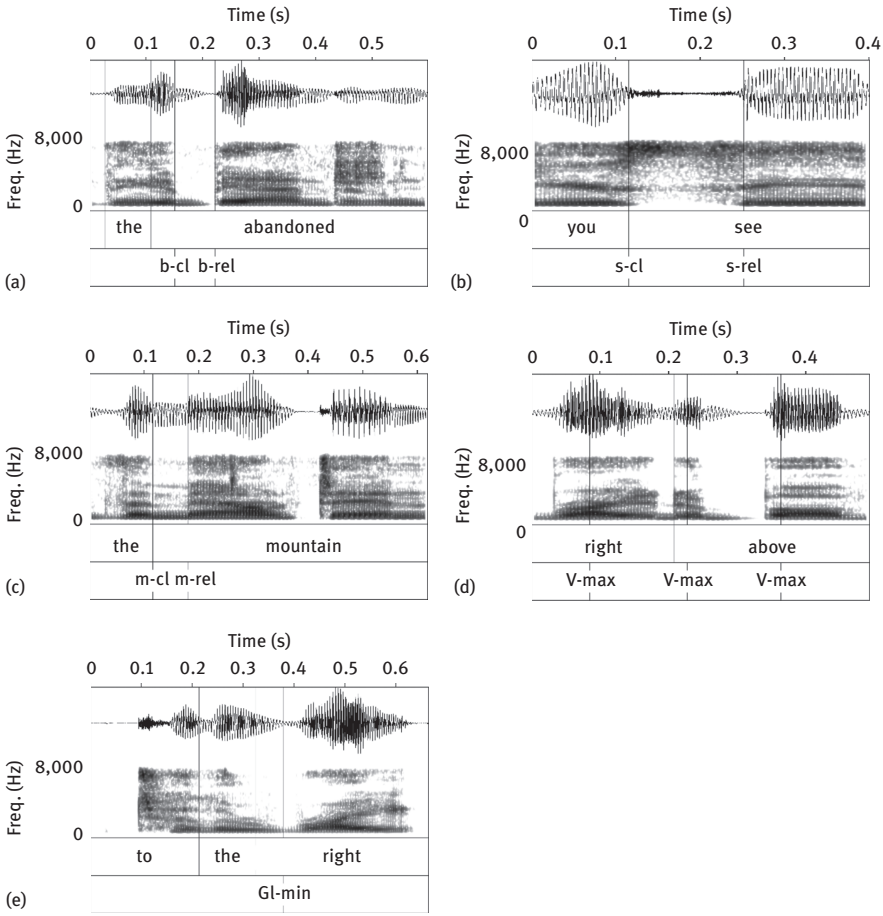

Figure 6.1:

AU: Figure captions were missing for all the figures within this chapter. Kindly provide the same if required.

occur in any clearly produced instance of the words in these utterances. Of the 122 target LMs, six LMs in Utterance 2 were excluded from the analysis of imitated productions reported below. The excluded LMs are from the prepositions *at* and *on*, which were frequently subject to lexical substitution in the imitated productions. Thus, where the stimulus contained "*... at the peak ... on the Canadian...*," imitators frequently swapped the prepositions or used the same preposition twice, e.g. "*... on the peak ... at the Canadian ...*". These lexical errors were frequent and variable across speakers, but occurred in otherwise fluent imitated productions, suggesting that the lexical target for the imitator may have been different than the word produced by the Map Task speaker. LMs for these variably produced prepositions were removed from all imitated utterances and from the

AU: Please advise if italics within double quotes, in the text 'contained "... at the peak ... on the Canadian...," imitators' and in subsequent occurrences, should be removed to avoid double emphasis.

Table 6.1: Number of target LMs in each manner class for Utterances and 1 and 2, combined.

Plosive		Fricative		Nasal		Glide	Vowel	Total
Closure	Release	Closure	Release	Closure	Release			
18	18	9	9	12	12	5	33	116

stimulus prior to measuring agreement in LM production. A breakdown of the remaining 116 target LMs by manner class is shown in Table 6.1.

6.2.3.2 Categorizing LM outcomes as intact or deviant

The way each predicted LM was implemented in each utterance was labelled by hand, as follows: the LM was either (a) implemented in its canonical form (termed "no change"), (b) merged with the following LM (see below for further discussion of LM merges), (c) modified to a different type of LM, or (d) deleted. In addition, occasionally an unpredicted LM was produced, labelled as (e) inserted. Labelling was done on the basis of visual inspection of the speech waveform and spectrogram, in conjunction with listening. The two authors labelled about 10% of the data from both utterances together to achieve consistency in labelling, and then labelled the remaining data independently, with regular discussion to resolve ambiguous cases.

Figure 6.1 provides illustrative examples of the 8 canonical LM types for American English: stop closure, stop release, fricative closure, fricative release, nasal closure, nasal release, glide and V. LM locations are labelled in the textgrid for each panel. (Affricates, which combine stop closure with fricative closure and release LMs, did not occur in our data sample.) The examples shown here are drawn from the larger corpus of stimulus utterances, including utterances whose analysis is not included in this study, chosen to provide the clearest illustrations of canonical LM realization.

As noted above, four different codes were used to annotate the outcome of each predicted LM.

- **No Change:** When the acoustic characteristics of a predicted LM matched those of the canonical definition described above. No Change is also described as a Match to the prediction.[2]

2 Note that the label No Change refers only to the acoustic properties that define the LM, and does not imply that other acoustic properties predicted by lexically specified features, or other acoustic properties present in the imitation stimulus, are realized intact.

- **Merge:** When two target consonants occurred in sequence with the release of the first C occurring simultaneously with the closure for the next C. For example, in an /st/ cluster, the LM associated with the release or end of the frication noise for the /s/ often coincides with the LM at the closure for the /t/.[3] In this case a single abrupt spectral change is simultaneously signalling the release of one constriction and the formation of another.
- **Substitution:** When the predicted LM was replaced by a different LM, i.e. when the cues in the signal matched those predicted for a different manner category.
- **Deletion:** When the predicted LM was missing altogether, and no substituted LM occurred between the preceding and following predicted LMs.
- **Insertion:** When a non-predicted LM was produced.

No Change and Merged LM outcomes are considered to be *intact* – the LM is produced as expected, given the lexical specification of the unreduced form and taking into account the adjacent context (for Merge). In Merge contexts, such as sequences of stops consonants and/or fricatives, merged LMs are expected to occur even in clear speech. Substitutions, deletions and insertions are considered as *deviant* LM outcomes, where the expected LMs are not realized. Perceptually salient reduction that relates to manner features, or C/V structure more generally, is expected to occur in contexts with deviant LM outcomes, though there may also be deviant outcomes that are transcribed based on evidence from the acoustic signal but which are not perceived.

Examples of Merges, Substitutions and Deletions – the three most common outcomes other than No Change – are illustrated in Figure 6.2. In the **Merge** example of Figure 6.2a, notice the abrupt end of frication noise for /s/ that is simultaneous with the abrupt beginning of silence for /t/ closure. In the **Substitution** example in Figure 6.2b, the abrupt spectral changes of the predicted LMs marking /d/ closure and release are not present and instead there is a gradual valley in intensity resembling a glide, with voicing continuing throughout. But note that not all alveolar stops that would be transcribed as flapped show this pattern of LM substitution, as shown in Figure 6.2c, where closure and release

3 The spectral characteristics of an acoustic LM associated with a change in manner can differ substantially, depending on the manner feature of the adjacent phoneme. For example, the spectral characteristics of the release LM for /s/ are quite different if /s/ is followed by a target stop vs. by a target nasal vs. by a target vowel. In fact, it would be difficult to imagine the same LM outcomes for /s/-release across these contexts. In these cases, the existence of a robust acoustic edge can serve as a perceptual cue to both the occurrence and the nature of the change in manner features.

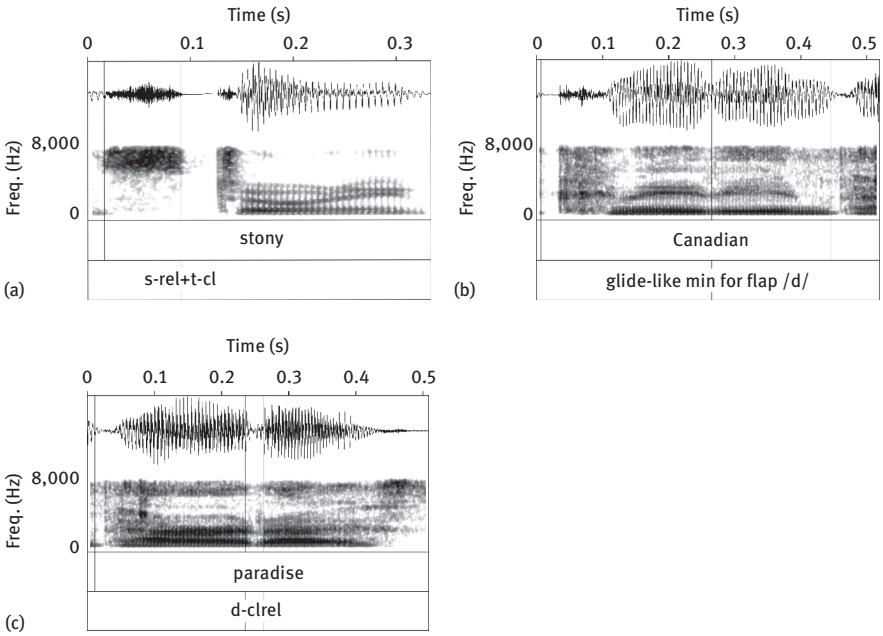

Figure 6.2:

LMs are observed in a flapped /d/. All LM substitutions in our sample involve realizations of the phoneme /d/ as lenited (i.e. an approximant realization) or flapped. Substitution occurs when the realization involves a change in manner – the stop closure/or release are not realized. When the /d/ is fully lenited and manifests as a glide the substitution results in the loss of an LM: this is labelled as substitution of a Glide-Min LM for the expected d-cl LM and deletion of the d-rel LM. In other cases, the /d/ may be partially lenited, manifesting with a glide-like transition at the left or right edge, but with an intact stop closure or release at the opposite edge. Such tokens would be labelled as having one substitution and one unchanged LM, and represent hybrid realizations – part stop and part flap or glide, which would be difficult to capture with a segmental transcription.

Examples of **Deletion** can be seen in these figures as well. In Figure 6.2a the predicted abrupt spectral transitions marking the closure and release LMs for /n/ in *stony* are absent; there are no LMs between the /o/ and /i/ vowels. In Figure 6.2b, the word-initial /k/ of *Canadian* is released directly into the nasalized voiced region for the /n/, with deletion of the predicted V LM in the initial syllable. Finally, in Figure 6.2c, we expect a glide LM for /ɹ/ in *paradise*, but there is no amplitude minimum in the vocalic interval spanning the first two syllables. Instead, the /ɹ/ is realized in rhoticization that extends over a long portion of the vocalic interval.

6.3 Results

6.3.1 LM outcomes: Comparing stimulus to lexically predicted LMs

The first objective of this study is to measure variability in the realization of the target LMs that are predicted from the lexical specification of a word in its unreduced, full form. We begin by evaluating the LM outcomes of the stimulus utterances. There is evidence of phonetic reduction in the stimulus utterances as they were originally produced by the Map Task speaker (see charts of LMs for Utterances 1 and 2 in appendix Figures A1 and A2). In Utterance 1, the speaker produces *Kate* as [kejtˀ] (in familiar allophonic terms, i.e. with an unreleased /t/), which is represented as deletion of the t-closure LM. She also produces *Canadian* as [kʰnejɹɪən], with no vowel LM for the unstressed schwa in the first syllable and with an approximant realization of /d/, which is labelled as deletion of the vowel LM and with a glide LM that substitutes for the predicted d-closure and d-release LMs.

Utterance 2 displays many more variable LM outcomes. One surprising feature of this utterance is the relatively oral-sounding production of the medial /n/ in gonna, transcribed as [gʊdə], which appears on the spectrogram as an oral [d]. This word is produced in the rapid, phrase-initial, unaccented sequence you're gonna be ..... We have no way of knowing if the oralization of /n/ reflects a speech error from an intrusive /d/ target, or if the intended target was a nasal that was ineffectively implemented. Nasal and oral stops share the same LM specification, with closure and release LMs, so the oral realization of /n/ in this utterance is considered to have intact LMs. Looking further into Utterance 2 we observe reduction of the medial /d/ and final /ŋ/ of standing, produced as [stæñĩ]. The initial /ð/ of *the* is produced after an interval of irregular pitch periods (ipp), which effectively masks the cue to ð-closure LM, although the ð-rel is intact. The /v- ð/ sequence in of the exhibits merger of the v-release and the ð-closure, which is an expected realization for a sequence of two fricatives (or stops). More reduction follows, with a deleted vowel LM for the, frication of the beginning of the /k/ in Canadian that is marked by substitution of the k-cl LM with x-cl, and subsequent deletions that yield the reduced form [x͡kẽĩɹɪm]. The assimilation of the final /n/ of *Canadian* to the labial place of the following /p/ in Paradise is not an LM effect, but the expected merger of the n-release and the p-closure is noted. Paradise exhibits one more reduction, with a lenited realization of the medial /d/ that it has an intact d-closure but deletion of the d-release.

We turn next to consider the patterns of reduction in imitated productions of these two utterances, where we are especially interested to see if the specific reductions that are present in the stimulus utterances are imitated in the same way.

6.3.2 LM outcomes: Comparing imitations to lexically predicted LMs

6.3.2.1 Frequency of intact and deviant LM outcomes

We turn now to examine the realization of lexically predicted LMs in the imitated productions of Utterances 1 and 2. Recall that we examine all three imitations from each participant, for both of the stimulus utterances. The reader should also bear in mind that the target LMs refer to the lexically derived LMs of the unreduced form of the words in the utterance, which are not always realized as intact in the stimulus utterances themselves, as shown in the preceding subsection.

There are a total of 116 target LMs from the combined stimulus Utterances 1 and 2 (Table 6.1). Each LM was produced 30 times in the imitations (10 speakers × 3 repetitions), and, including 22 inserted LMs (not predicted from lexical specification), there was a total of 3,502 LM *outcomes* in our data. As shown in Table 6.2, the large majority (79%) of these target LMs are realized in their predicted form with no change (NC), or in the form that is predicted from the immediately adjacent phones (Merged). We refer to these as *intact* outcomes. There are also instances of LM substitution and deletion, and a few additional

Table 6.2: Classification of produced LMs relative to their predicted form: Intact (No Change or Merged) and Deviant (Deletions, Insertions, Substitutions). Each cell reports the number of LMs produced (outcomes), and in parentheses that number as the proportion of outcomes from 3 repetitions of each utterance by 10 speakers.

	Utterance 1	**Utterance 2**	**Total (Utts. 1–2)**
No change	794 (0.66)	1,477 (0.64)	2,271 (0.65)
Merged	188 (0.16)	306 (0.13)	494 (0.14)
Intact (NC + Merg)	**982 (0.81)**	**1,783 (0.78)**	**2,765 (0.79)**
Substitution	56 (0.05)	95 (0.04)	151 (0.04)
Deletion	164 (0.14)	402 (0.18)	566 (0.16)
Insertion	11 (0.01)	11 (0.00)	20 (0.01)
Deviant (S+D+I)	**229 (0.19)**	**508 (0.22)**	**737 (0.21)**
Total (Intact + Dev.)	1,211	2,291	3,502

cases of inserted LMs associated with segments not included in the lexical specification of a word, such as sporadic appearance of a full glide LM for a [j] inserted between the last two vowels in *Canadian* [kʰənɛɪdijən]. We refer to LM substitutions, deletions and insertions as *deviant* outcomes, which collectively represent 21% of the total number of LM outcomes produced by our participants.

Among the deviant LMs, the most common outcome is deletion, representing 16% of total outcomes and 77% of the deviations. In comparison, insertion and substitution account for 1% and 4% of LM outcomes, respectively. This finding indicates that LMs are capturing some aspects of the patterns of phonetic reduction in the sense of a production that is reduced with reference to its full form, by virtue of providing fewer cues to signal the presence of a phoneme (cues to syntagmatic structure), or by providing fewer cues to signal contrastive manner features (cues to paradigmatic contrast). In this sense, LMs provide a means to measure certain of the missing components from speech. This merits further analysis of the patterns of deviant LMs in our data.

6.3.2.2 LM outcomes by manner class

Having established that LMs index some patterns of phonetic reduction, we turn to the second objective of this study, which is to determine if all LMs are equally susceptible to variation in production outcome, or if deviant outcomes occur more often for some LMs than for others. In this analysis we compare outcomes based on the manner class of each LM, irrespective of its local (left and right) context, for these manner classes: Plosive, Fricative, Nasal stop, Glide (/r, l, j, w/) and Vowel. As described in Section 6.2, there are distinct LMs marking the closure and release of Plosives, Fricatives and Nasal stops.

Figure 6.3 shows the percentage of LM outcomes that are deviant, for each manner class. These figures reveal a number of interesting asymmetries. The most frequent types of deviant LMs that occur in our small data set are the plosive closure and release LMs, though this is primarily due to the greater number of plosives in this small sample, compared to the other consonantal LMs. When deviant plosive LMs are counted as a proportion of the total number of plosive LMs (Figure 6.3), the deviant outcomes are only slightly more frequent for plosives than for fricatives, with the biggest difference between these two manner classes found for the closure LM. Furthermore, despite occurring with medium frequency relative to plosives and fricatives, nasal LMs are overall less likely to have deviant outcomes, and are especially infrequent in the case of nasal closure LMs. Glide LMs, though less frequent in our sample than other

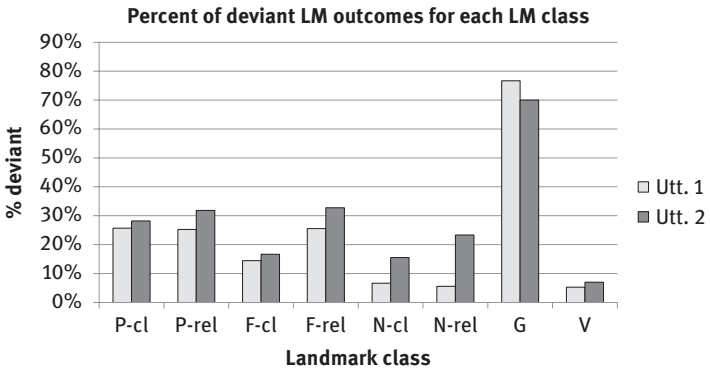

Figure 6.3:

consonant types, are realized as deviant in 70% and 77% of outcomes in Utterances 1 and 2, respectively. The opposite pattern is found for vowel LMs, the most frequent LM type by a large margin, which are almost always realized as intact, with only 5–7% deviant outcomes. A final observation from these findings is that LM outcomes are different for the closure and release components of fricatives and to a lesser degree for nasals and plosives. For these "two-phase" consonants, the release LM is more likely to have a deviant outcome than the closure LM.

6.3.2.3 Contexts for frequent deletion of target LMs

Our LM observations are based on repeated productions of only two utterances, with unequal representation of LMs from the different manner classes, and represent only a fraction of the local contexts in which each LM type may occur, given the phonotactics of English. We are therefore cautious in drawing generalizations from these data, especially concerning the relative likelihood of deviant LM outcomes for different LM classes. It is useful, though, to examine the specific lexical, prosodic and segmental context for the most common types of deviant LM outcomes in this small corpus as an indication of the contextual factors that condition phonetic reduction. Recall that deletion is the most common type of deviant outcome, with substitutions and insertions occurring much more sporadically. Towards this end, we examined our data to identify the individual consonant LMs in the stimuli that exhibit the most frequent occurrence of deletion outcomes. We qualitatively characterize the contexts with the highest incidence of LM deletion (identified to be 10 or more deleted outcomes out of 30 possible):

a. Consonants in intervocalic position, preceding an unstressed vowel: Here we find frequent deletion of the closure and release LMs for the nasal in /… VŋV …/ in *standing* at, and for the /r/ in /… VrV …/ in *Paradise*. This is also the context for optional flapping of /d/, which occurs often but not always in *Canadian* and *Paradise*. An intervocalic context is also the frequent context for deleted word-initial /g/ LMs in *gonna* following a deleted /r/ LM in *you're* in the phrase *you're gonna*. Here, speakers often produce a very weakly constricted velar approximant with no evident closure or release LM.
b. Consonant clusters: In CC clusters like the /st/ in *standing*, the /nt/ in *mountain*, and the /td/ across a word boundary in *Kate dyou*, there is frequent but not consistent deletion of the release LM of the first consonant with or without deletion of the closure LM of the second consonant. When both LMs are deleted in the same production, the result is a noticeably reduced pronunciation, e.g. when the /st/ cluster is realized as an [s] that releases into a burst characteristic of /t/, but without the prior /t/-closure interval. It's interesting to note that deletion in CC clusters often leaves one LM for each consonant intact, which may support the perceptual identification of both consonants. In our corpus such deletion is most extensive in the word *mountain*, with fully 10 instances out of 30 having deletion of all four LMs for the /nt/ cluster, leaving nasalization of the preceding vowel as the only clear clue to the /n/, and no more than a miniscule burst of irregular glottal pulsing as a cue to the /t/.
c. Schwa vowels in syllables preceding the stressed syllable: This is one of the very few contexts in which a vowel LM is deleted in our sample, and it is the most frequent outcome for the schwa LM in the initial syllable of *Canadian* in both utterances. A similar context is found for the schwa in the final syllable of *mountain*, which is deleted in the 10 (out of 30) productions that have a syllabic /n/ instead.
d. Coda /t/: The final /t/ of *Kate*, which also often occurs before a pausal juncture, is very often entirely deleted in our data, leaving at most a trace of ipp marking the characteristic glottal constriction that accompanies coda /t/.
e. Onset /j/: In 16 of the 30 tokens, the initial glide in *you're* is manifest primarily in the formant transitions into the following vowel, but without evidence of the diminished amplitude that defines the glide LM. In this case, we might consider whether the glide has migrated from the onset to the nuclear position, creating a [ɪɔ] diphthong. Similarly, the glide in the onset cluster /dj/ of *d'you* is present in only one token, with other tokens showing a fricated release of /d/ and a heavily fronted /u/ vowel. The frequency of the deviant LM pattern in this word suggests a stored specification of the reduced variant, though we do note the occurrence of one production that preserves a clear glide LM.

6.3.2.4 Between-speaker variability in LM realization

It is clear even from this limited data set that variation in the production of target LMs is fairly systematic. Across speakers, the frequency of deviant LMs is relatively low, with deletion as the favoured deviant outcome. In addition, deviant LM outcomes are more likely for glide and plosive LMs than for vowel or nasal LMs, and the most common deleted variants tend to occur in specific phonological contexts. Considering these systematicities, we ask if they hold uniformly across speakers.

The distribution of LM outcomes across the Intact and Deviant classes shown in Table 6.2 is representative of the distributions for each speaker, as shown in Figure 6.4, which plots the distribution of LM outcomes over the total number of LMs produced by each speaker. Overall, there are very consistent patterns of LM realization across speakers, in terms of the frequency of intact LMs and the relative frequency of deletion compared to substitutions or insertions in deviant LM outcomes.

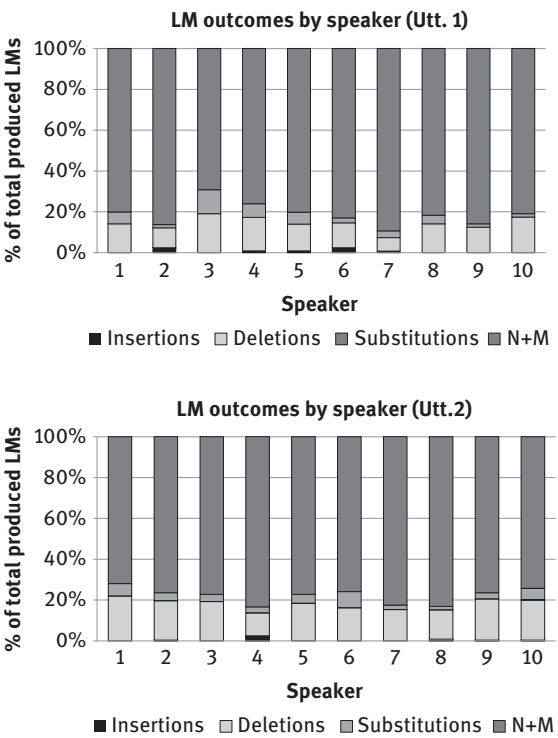

Figure 6.4:

The fact that different speakers produce deviant LMs with similar frequency raises the question of whether different speakers are producing deviant LM outcomes for the *same* target LMs in the stimulus. This would be the case if deviant LM outcomes are systematically produced in certain segmental or prosodic contexts. Identifying the contexts where target LMs are systematically produced as deviant is important because it might yield insight into the mechanisms of articulation and speech planning that give rise to phonetic reduction. We examined each target LM in the stimulus for the consistency of outcomes across speakers, counting the number of speakers who produced intact outcome for that LM in all three repetition of the stimulus (the "high-intact" LMs), as well as the number of speakers who produced no intact outcomes – i.e. for whom every outcome was deviant with reference to the target (the "high-deviant" LMs). These two categories were taken as the endpoints of a deviance scale with values from 0 to 10, with high-intact LMs assigned a value of zero and high-deviant LMs assigned a value of 10. Intermediate values were assigned to each LM based on the number of speakers (out of 10) who produced one or more deviant outcomes for that LM. If there is consistency among speakers in producing deviant LM outcomes for certain targets, e.g. as determined by LM type and the phonological context of the LM, then we expect to find a lot of LMs in both the high-intact and high-deviant groups. On the other hand, if there are strong individual differences or token-by-token differences in LM realization, we expect to find more variable outcomes, and a higher number of LMs with intermediate values on the deviance scale. This prediction is only partially confirmed.

Figure 6.5 shows the distribution of the LMs (Utterances 1 and 2 pooled) along the deviance scale. While we observe a concentration of LMs in the

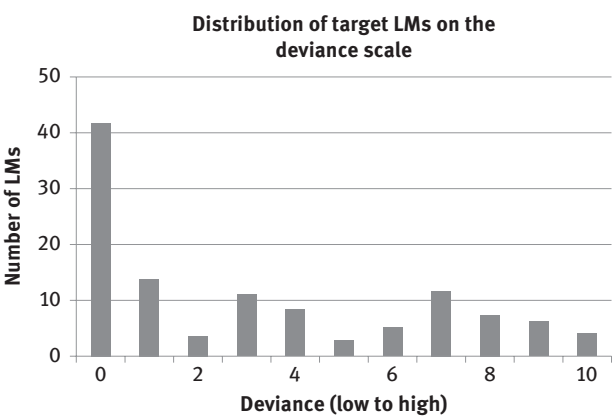

Figure 6.5:

high-intact category (42 out of 116, or 36%), there are very few LMs in the high-deviant category (4 out of 116, or 3%), and relatively few at intermediate values. This finding indicates that while many LMs (36%) are consistently produced as intact across speakers, there is less consistency across speakers in the production of deviant LMs.

Extending the "high-deviant" category to include the target LMs with values from 0 to 2 on the deviance scale (i.e. those for which 8 out of 10 speakers produce one or more deviant outcomes), there are 17 out of 116 LMs, or 15%. The deviant outcomes of these LMs represent several well-known types of phonetic reduction in American English: schwa deletion (*Canadian*), lenition of nasal stop closure in NC clusters following a nasalized vowel (*standing, mountain*), deletion of /r/ following an r-coloured vowel (*you're, paradise*), lenition of stop or fricative closure intervocalically (*see the, you(re) gonna, gonna*), and loss of oral closure for post-vocalic coda /t/ in the presence of glottal constriction (*Kate*). These LMs are also among those listed earlier as having the highest occurrence of deletion outcomes over the entire data set (considering all repetitions from all speakers).

6.3.3 LM outcomes: Comparing imitations to stimulus

In the preceding paragraphs we examined variability in the consistency with which a target LM was produced as intact in all three outcomes by the speakers in our study, and we looked into the identities of the LMs with consistently deviant outcomes for all or most speakers. A question arises here as to whether deviant outcomes of the target LMs are in fact produced as a faithful imitation of the stimulus. After all, the original speakers of these utterances from the Map Task corpus themselves produce deviant LMs for some of the LM targets in their speech. Figure 6.6 illustrates, for each speaker in our study (the imitators), the number of LM outcomes that are deviant with reference to the target LM, and those that are deviant with reference to the stimulus (i.e. where the imitation fails to match the LM outcome in the stimulus). In Utterance 1, LMs differ from the target at about the same frequency as they differ from the stimulus, but in Utterance 2 there are somewhat more LM outcomes that differ from the stimulus compared to those that differ from the target. That means that in Utterance 2, which is the longer utterance, speakers are relatively more reliable in producing LMs as projected from the dictionary specification of each word than they are in accurately imitating the phonetic realization of LMs in the stimulus.

In considering the behaviour of individual speakers across the two utterances, we ask if some speakers are overall more accurate in producing intact outcomes for target LMs, or conversely, if some speakers are more accurate in imitating

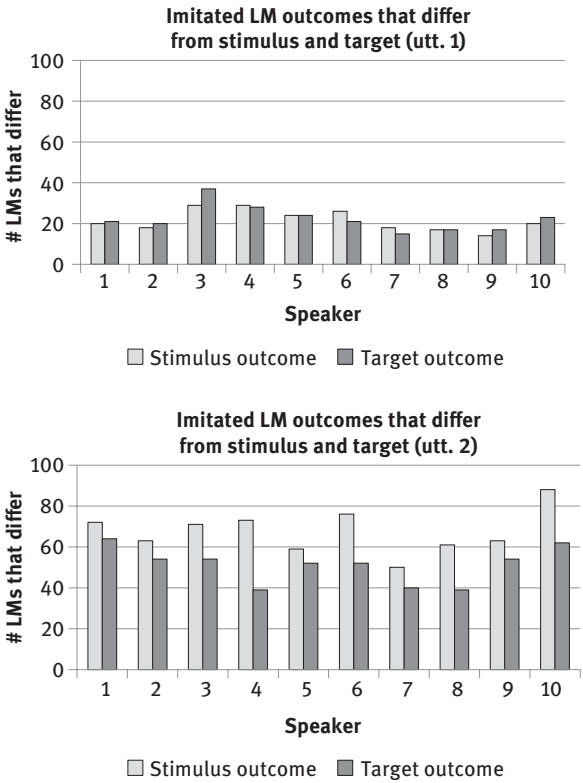

Figure 6.6:

LM-related phonetic detail from the stimulus. Comparing the data shown in Figure 6.6, we do not see a relationship between the accuracy in either dimension for these ten speakers. For instance, among all the speakers it is Speaker 3 who produces the greatest number of inaccurate (mismatched) outcomes of target LMs in Utterance 1, but this speaker's productions are not very different from the other speakers for Utterance 2. To be clear, there are individual differences in accuracy among the speakers, both in comparing outcomes to the target and to the stimulus, but it's not clear if the differences reflect individual speaker differences that would generalize across more utterances. An alternative scenario is that the observed differences in accuracy of LM production (or imitation) reflect a range of variation in speech production, with comparable variation within and between speakers.

To determine if speakers are producing deviant LM outcomes (i.e. deviant with respect to the dictionary-predicted LM) in order to more accurately imitate those LMs that are deviant in the stimulus (i.e. that the stimulus speaker

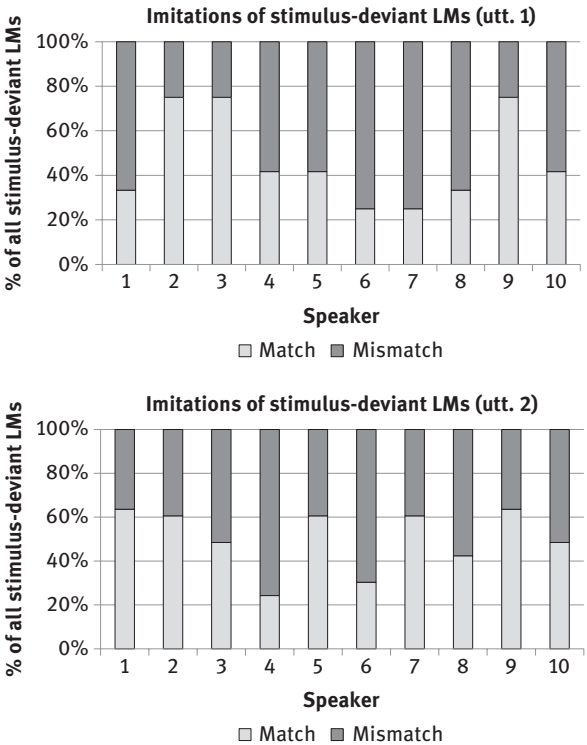

Figure 6.7:

produced as deviant), we examine the LM outcomes for the LMs that were deviant in the stimulus utterances. There are 4 deviant LMs in stimulus Utterance 1 and 11 in Utterance 2, making a total of 15 stimulus-deviant outcomes. Considering only the imitations of those 15 stimulus-deviants, we ask how often our speakers produced identical deviant outcomes. Figure 6.7 displays the number of matching vs. mismatching outcomes for these stimulus-deviant LMs as relative proportions, out of 12 for Utterance 1 (4 stimulus-deviant LMs × 3 repetitions) and 33 for Utterance 2 (11 stimulus-deviant LMs x 3 repetitions). The highest rates of matching to stimulus (Speakers 2 and 3, but only for Utterance 1) still fail to match the precise pattern of phonetic reduction in the stimulus about 20% of the time, which is the overall rate of variability in LM outcomes in our data set. It's possible to claim that for the shorter utterance these two speakers have achieved the maximum imitation precision possible for this task. But even these same speakers do not perform as well for the longer utterance, Utterance 2, and all other speakers show imitation match that is far lower than the overall rate of variation for LM outcomes. The relative proportion of matched (to stimulus) vs. mismatched LM

outcomes is quite variable across speakers. It appears from this small data set that speakers do not reliably or consistently imitate deviant LMs in the imitation stimulus. In other words, reduced forms are not imitated in the same way as they occur in the stimulus.

6.3.4 Within-speaker variability in LM realization

Although the overall frequency of deviant outcomes, among all 3,502 outcomes in the data set, is moderate, at 21% (see Table 6.2), when we consider the 116 target LMs individually, we observe that fully 74 (64%) are realized with variable outcomes by one or more speakers (these are the LMs with values greater than zero on the deviance continuum, see Figure 6.5). Together these facts point to within-speaker variability in the production of LMs. We assess within-speaker variability to determine the extent to which an individual speaker is consistent in her implementation of LMs. Consistency could result from the faithful realization of lexically predicted LMs, or from the careful imitation of the stimulus, or even from an individual speaker's idiosyncratic patterns of deviant LM production – a kind of phonetic "signature".

Figure 6.8 shows, for each speaker, the relative proportion of target LMs that are realized with the same outcome over all three repetitions, and the number that are realized with two outcomes, out of the 116 target LMs in Utterances 1 and 2 combined. It is notable that over the entire data set there is only one occurrence of a target LM that is produced in three distinct outcomes by a single speaker, and that was the /p/-closure LM for the /p/ in *paradise*, which Speaker 9 produced once as merged with the preceding word-final nasal of *Canadian*, once as a voiced

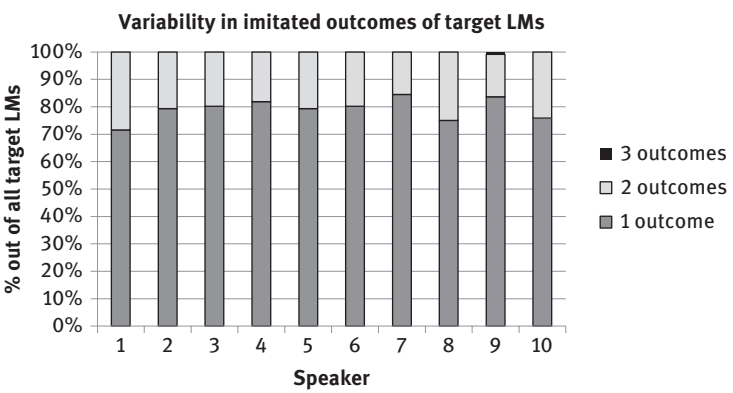

Figure 6.8:

/b/, and once "deleted" (with no interval of voiceless closure). All other target LMs were produced either with the same outcome, or with two distinct outcomes over three repetitions by a given speaker. Across speakers, the vast majority of target LMs (79%) were produced with a single outcome in all three repetitions by the same speaker. For the other 21% of target LMs, there was within-speaker variation in LM outcomes over three repetitions. Note that this level of within-speaker variation is similar to the overall level of variation in LM production, where we find 21% of the total LM outcomes to differ from the lexically predicted LM (see Table 6.2).

Figures 6.9–6.11 provide examples of how imitated tokens can either reproduce the LMs of the stimulus or modify them, even within a single speaker. Figure 6.9 shows the stimulus word *mountain*, where all of the LMs predicted for the *-ntain* sequence were produced. Figure 6.10 shows three consistent imitations from a single speaker, which are produced with a set of LMs that are different from those in the stimulus. Specifically, the /n/ and /t/ in the medial cluster lack closure and release LMs, and the V LM for the vowel in the second syllable is also absent, with an /n/ closure and release signalling the syllable nucleus. Note that ipp are present in each imitation, providing perceptual cues to the obstruent and voiceless features of the LM-less medial /t/, while nasalization of the vowel in the first syllable (not labelled) provides a cue to the medial /n/.

Figure 6.11 illustrates, in contrast, an example of intra-speaker variation. The first of these three imitations produced by a single speaker reproduces the closure and stop cues to the medial /t/ and the vowel LM for the reduced second-syllable vowel that are present in the stimulus shown in Figure 9. But for the second and third imitations, this speaker produces a different set of cues for the medial /t/, including ipp, and deletes the V LM for the second-syllable vowel (whose presence may be cued in part by the long duration of the oral closure for the final /n/).

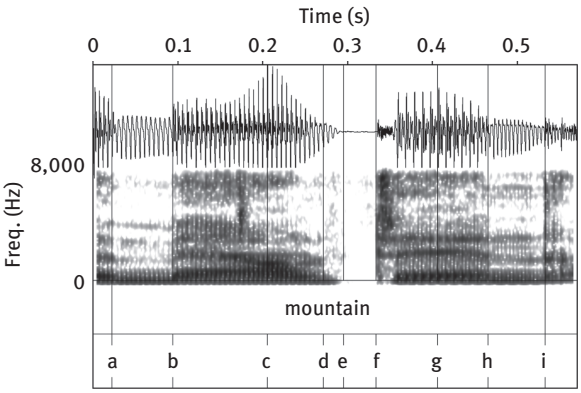

Figure 6.9:

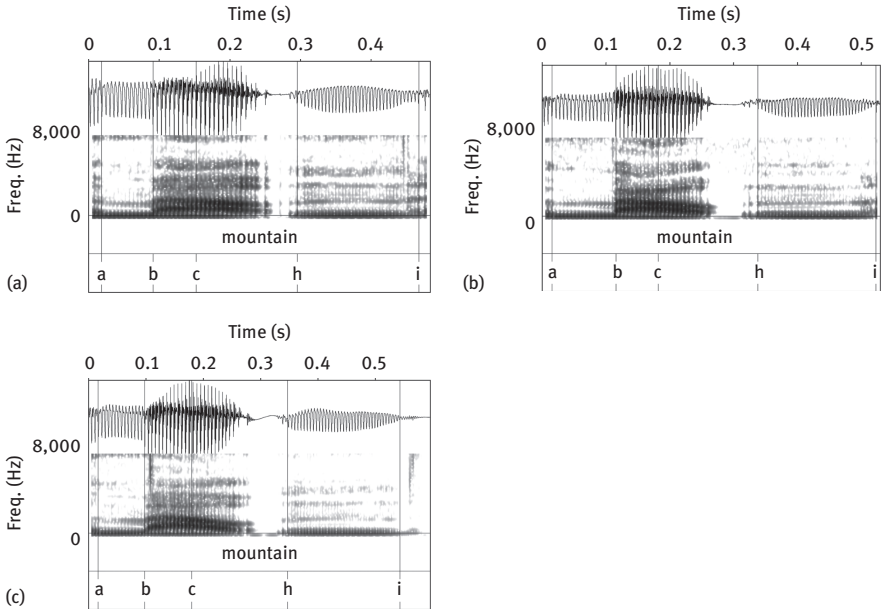

Figure 6.10:

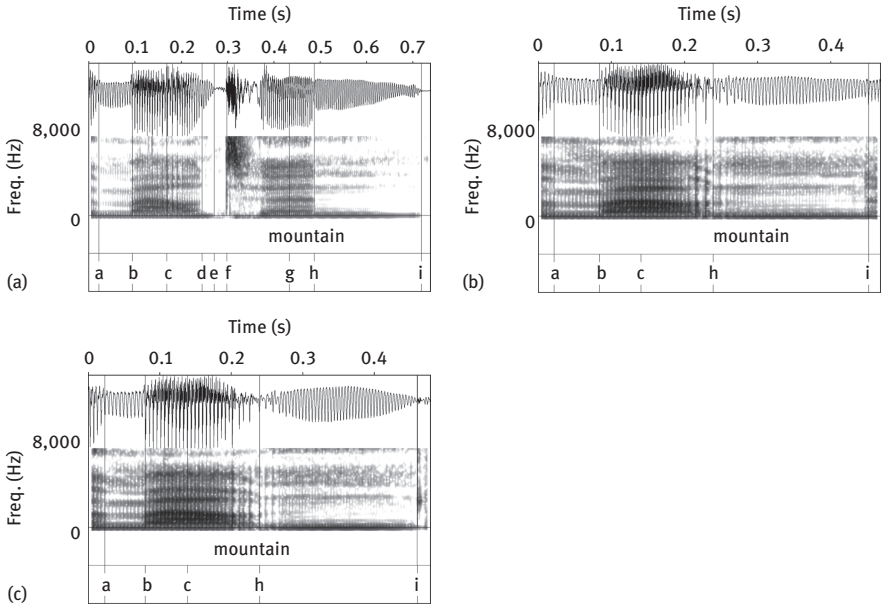

Figure 6.11:

The final observation we report for within-speaker variability is whether there were any effects of repetition order in the pattern of LM outcomes. We compared the distribution of intact and deviant LMs for each repetition. There were no distinct patterns of increase or decrease in deviant LMs across the three repetitions. Indeed, the number of deviant LM outcomes of each type was very similar across the repetitions by a single speaker, for each speaker.

These examples illustrate a significant aspect of surface phonetic variation which emerges even in this highly constrained imitation task: the degree to which speakers can choose among surface forms which differ in their acoustic details but nevertheless provide significant cues to many of the defining features of the phonemes of their intended words. As Gow (2003) has shown, even highly overlapping articulations (as for the target /t/-/b/ sequence in *right berries*) which result in explicit transcription of a different segment (e.g. a final /p/) often provide enough information in the signal to permit a listener to perceive the target /t/ when tested with a more sensitive online task such as lexical decision. The question of which feature cues are present in a given utterance is thus not always best answered by a transcription in terms of a string of discrete symbolic elements like allophones. The reduced and overlapping cue patterns suggest the need for an approach to transcription that can capture information about individual cues to contrastive features. LM labelling is a first step in that direction, providing a means of capturing the target segments whose presence is robustly signalled in the utterance, vs. those whose presence is less robustly signalled, and opening the way for a more comprehensive labelling system which can capture the additional feature cues that are richly represented in nearby regions of the signal, such as formant transitions that correlate with place features, and vocal fold vibration patterns that correlate with voicing.

6.4 Discussion

We have analysed three serial imitations of two stimulus utterances, from each of ten speakers of American English, to explore patterns of phonetic variation as indexed by LMs – spectral cues to the phonemically contrastive manner features of plosives, fricatives, nasal stops, liquids, glides and vowels. Our broad goal is to identify patterns of phonetic variation, including reduction, that are common to many or all speakers, and to look for effects of phonological context on phonetic variation. We are also interested in establishing the usefulness of imitated speech for investigating phonetic variation, and the adequacy of LMs for indexing and quantifying variable speech outcomes relative to a lexically specified target, or

relative to the phonetic details of the heard stimulus. This study reveals interesting findings related to each of these objectives, which are summarized here in relation to the five research questions posed in Section 6.1.

Q1: Variability in LM outcomes: The most general finding is that phonetic variation does occur across imitated productions of a target utterance, as measured by LM modification patterns, even when the lexical, syntactic and phonological contexts for each sound are held constant across imitated productions. This variation could be due to choices an individual speaker makes about other factors influencing variation, such as speech rate or speaking style (careful or casual), to the extent that the imitation does not match the stimulus utterance on these dimensions. Another possible explanation for variable phonetic outcomes is that to some degree speakers do not control non-contrastive phonetic detail – in other words, some degree of variability is inherent in the speech production process, perhaps reflecting bounds on the precision with which articulator movement is controlled. However, this account is hard to reconcile with other findings in the literature, such as conversational convergence (e.g. Babel 2012; Pardo 2006; Pardo et al. 2012), in which two speakers in a conversation appear to exercise exquisite degrees of control over aspects of acoustic-phonetic variation such as non-contrastive variation in vowel formant trajectories or VOT. The question about sources of variation in imitated speech cannot be fully resolved from analysis of the data presented here, but our findings confirm that imitated speech is useful for the study of phonetic variation. Imitated speech allows for the analysis of the same phonological content, holding constant other features of the linguistic context, so that phonetic outcomes can be quantitatively compared within and across speakers.

Phonetic variation was assessed by labelling LM outcomes as intact (either unchanged or modified in a way that is predicted by the adjacent segments), or as deviant (substituted, deleted or inserted). Out of 3,502 LM outcomes, 79% were intact realizations of target LMs in the stimulus. The most common type of deviant outcome was deletion, accounting for 16% of the total outcomes, and 76% of all deviant outcomes. This finding offers clear evidence that reduction – in the sense of fewer phonetic cues to contrastive phonological units – is the primary source of variation affecting LMs as cues to contrastive manner class features, in this imitation task.

Q2: Variability by LM class: Deviant LM outcomes are not equally probable for LMs from each manner class. Focusing on deleted LM outcomes, we find that glide LMs are the most susceptible to deletion, when deletions are counted in proportion to the total number of glide LMs in the stimulus utterances. Plosive-closure, plosive-release and fricative-release LMs have smaller and roughly equal proportions of deleted outcomes, while fricative-closure, nasal-closure, and nasal-release LMs are even less likely to be deleted. Vowel LMs are almost

always realized as intact, with far fewer deviant outcomes than any type of consonant LM. The fact that closure and release LMs differ in their frequency for some manner classes (fricative and nasal stop) suggests an advantage for LMs over segment-sized symbolic features in characterizing patterns of phonetic reduction, and motivate the further pursuit of this comparison in future work.

The very high likelihood of intact outcomes for vowel LMs points to an important difference between consonants and vowels: phonetic variation affecting manner class features, including variation resulting in the reduction of acoustic cues, is much more likely for consonants than for vowels. This C/V asymmetry is further heightened by the fact that glides, which among consonants are the most similar to vowels by acoustic criteria, show the opposite pattern, with deviant outcomes being the most likely: 74% of glide LMs are deleted in the imitated utterances, while the proportion of deleted vowel LMs is only 6%.

The finding that vowels are much more robust to variation in LM realization than consonants are may be understood in terms of syntagmatic structure and paradigmatic contrast. Consonants from all the manner classes can occupy many of the same positions in syllable and word structure, so the substitution of a consonant LM of one class for that of another class will often result in a phonotactically legal outcome – for example, when the /g/ in *gonna* is realized with a weakened approximant constriction rather than the predicted plosive closure and release, the resulting C/V structure is unchanged. Similarly, in most contexts the deletion of a consonant LM does not create a phonotactic violation, e.g. loss of the coda /t/ in *Kate* results in a legal [CV:] syllable. Turning to vowels, we might expect that a nasal or liquid consonant LM could substitute for a vowel LM, as syllabic consonants, since the substitution is not likely to induce a phonotactic violation. Yet such substitutions are not observed and would be surprising, e.g. if *Kate* were realized as [kr̩t].[4] Clearly, there are factors beyond phonotactic output constraints that must play a role in shaping LM outcomes. On the other hand, phonotactic considerations may help explain the rarity of vowel LM deletions, as there are many contexts in which the loss of a vowel LM would result in a phonotactically illegal consonant cluster. For example, the (hypothetical) loss of the vowel in *Kate* would result in the unsyllabifiable sequence [kt].[5]

4 Note that the occurrence of a syllabic nasal in a word like mountain (second syllable) is not an example of substitution of a vowel LM for that of a nasal stop, since in this case the nasal is present in the lexical representation, so the reduced pronunciation results from deletion of the vowel LM, leaving the predicted nasal LMs intact.

5 We do observe frequent deletion of an unstressed vowel LM in contexts where the flanking consonants do not form legal onset or coda clusters, e.g. in productions of *Canadian* where the LM for schwa in the first syllable is deleted. The resulting [kn-] sequence is not a legal syllable

Another observation that may relate to the robustness of vowels is that there is an apparently parallel finding from studies of elicited speech errors, which have reported that errors on vowels alone, and not in combination with consonants, are very rare relative to errors on consonants (Rusaw and Cole 2011). It is possible that the relative stability of vowels in speech production, compared to consonants, reflects their status as the locus of C/V coordination in speech production (Browman and Goldstein 1988). Clearly, more research is needed to fully understand the source of vowel stability in speech production.

Q3: Between-speaker variation: This study was designed to elicit many instances of the same LMs from individual speakers, and across speakers, but with the trade-off that we do not have equally representative data from all LM types in all the contexts where they may occur in English. This limits our ability to identify contexts that condition reduction and other variable LM outcomes. Nonetheless, we observe that some LMs are very often realized with deviant outcomes across speakers in this corpus, and without exception these phenomena represent familiar patterns of phonetic variation. Specifically, we observe deletion of consonant LMs in intervocalic position before an unstressed vowel (e.g. *pa<u>r</u>adise*), deletion of the release LM for the initial consonant in a CC cluster (e.g. *mou<u>n</u>tain*), deletion of the vowel LM for schwa (e.g. *C<u>a</u>nadian*), deletion of coda /t/ (e.g. *Ka<u>t</u>e*) and deletion of post-consonantal onset /j/ (e.g. *d'<u>y</u>ou*).

What strikes us as most remarkable about LM deletion in the contexts described above is not that deletion is consistent across speakers, but that it is not universal. Of the 116 target LMs in this study, only one is never realized intact, and that is the vowel LM for the schwa in the first syllable of *C<u>a</u>nadian*. All other target LMs that undergo frequent deletion are realized as intact in one or more imitated productions. A related observation is that the production of a deviant LM outcome is not strongly predicted by the LM pattern in the stimulus that the imitator heard. Even if the stimulus is deviant (with respect to the lexically specified LMs), the imitation does not reliably reproduce the same deviant LM outcome, nor are all the deviant outcomes in the imitation matched to deviant outcomes of lexically specified LMs in the stimulus. The non-uniformity in deviant outcomes tells us that although a particular phonological context may license modification of a lexically specified LM, such a modification is not obligatory, at least not in most cases. Rather, the findings from this study suggest that speakers have a

(or word) onset, but in this case speakers seem to be recruiting the /n/ as a syllable nucleus, filling the position that the deleted V LM would otherwise occupy in syllable structure. In light of examples like this, it may be more appropriate to talk of syllable constraints on LM outcomes, more specifically.

range of options for the phonetic implementation of a lexical item, and the choice among them is not fully determined by the linguistic context.

Q4: Accuracy in imitation vs. realization of lexical target: LM outcomes were compared across speakers to test for differences in the frequency of intact realization of lexically specified (target) LMs, or in the accurate imitation of deviant LMs in the stimulus. We found no evidence for either. In other words, among the ten subjects in this study there are no exceptionally clear speakers, and no superior imitators. Rather, all speakers exhibit very similar overall patterns in the distribution of intact and deviant LM outcomes, along with some variation in the choice of deviant productions.

Q5: Within-speaker variation: Related to the question of whether some speakers are more accurate in producing target LMs, or in imitation, is the question of whether speakers are internally consistent in speech production, favouring one particular outcome for a given target LM across all repetitions. We reason that internal consistency in the production of phonetic variants would help to establish individual speech patterns that could function to index social identity. The data provide scant evidence that speakers are using LM variation in this manner. Target LMs are produced by the same speaker with a unique outcome and with divergent outcomes in nearly equal proportions.

The finding of moderate within-speaker variation contributes to the overall picture of phonetic variation as an inherent property of speech production, with very similar frequencies of variable outcomes across and within speakers. In the speech sample analysed here, variable LM outcomes occur in about 20% of the outcomes overall (counting all speakers and all repetitions), and in about 20% of an individual speaker's productions (counting all repetitions). The same proportion of target LMs, about 20%, are produced with variable outcomes in one or more imitations. Furthermore, this finding is an exact replication of the finding from Shattuck-Hufnagel and Veilleux (2007), who report 20% deviant outcomes for LM realization on a larger sample of the Map Task corpus, the same corpus from which the stimuli in our study were taken. The remarkable consistency among these measures of variability lends further support to the idea that some degree of variability is inherent to the speech production process, which we can estimate at 20% on the basis of the present speech sample.

6.5 Conclusion

In this exploratory study of phonetic variation in an imitation task that highly constrains the syntactic, prosodic and lexical aspects of the utterance, we have demonstrated that speakers do not always reliably reproduce the target

utterance at the level of detail measured by the cues to manner features known as acoustic LMs. The results provide support for the view that imitation is a useful method for eliciting phonetic variation under controlled conditions, and that LM labelling is a useful tool for quantifying certain aspects of the degree and type of phonetic variation. Because the range of types of phonetic variation in a spoken language is very large, affecting many aspects of an utterance beyond the LM cues to manner features, the hand-labelling method we have used in this preliminary study may not be practical for more comprehensive studies, aimed at inventorying the full set of variation patterns and the way these patterns are used in different contexts and by different speakers. However, the results of this study have provided some initial observations of this type, and have demonstrated the usefulness of this approach to the study of phonetic variation.

Future work can build on the patterns of LM variation reported here through controlled studies with specific phonological targets, or by manipulating prosodic context, speech rate, or other features of the elicitation stimuli. Developments in the direction of automatic detection of LMs and other cues to feature contrasts, applied to existing large corpora of spontaneous speech, will enable the study of the broader principles that govern patterns phonetic variation in spoken language. Future research along these lines would provide a more comprehensive test of the hypothesis that individual feature cues provide a useful vocabulary for annotating phonetic variation. In linking the perception of phonetic reduction with the listener's subsequent speech production behaviour, this line of research may also shed light on the mechanisms of sound change. Towards these goals, it will be instructive to compare the appropriateness of narrow symbolic transcription, raw acoustic measures and listeners' perceptual judgements for quantifying systematic phonetic variation.

AU: 'Choi and Shattuck-Hufnagel 2014' and 'Cole et al. 2007' were listed in the reference list but missing in the text. Kindly include in the text, or remove it from list.

References

Allen, J. Sean & Joanne L. Miller 2004. Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America* 115 (6). 3171–3183.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson & Regina Weinert 1991. The HCRC map task corpus. *Language and Speech* 34 (4). 351–366.

Babel, Molly. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics* 40 (1). 177–189.

Browman, Catherine P. & Louis Goldstein 1988. Some notes on syllable structure in articulatory phonology. *Phonetica* 45.2–4: 140–155.

Choi, Jeung-Yoon & Stefanie Shattuck-Hufnagel 2014. Quantifying surface phonetic variation using acoustic landmarks as feature cues. *The Journal of the Acoustical Society of America* 136 (4). 2174.

Cho, Tae-Hong 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/ in English. *The Journal of the Acoustical Society of America* 117 (6). 3867–3878.

Cole, Jennifer 2015. Prosody in context: A review. *Language, Cognition and Neuroscience* 30 (1–2): 1–31.

Cole, Jennifer., Kim, H., Choi, H. and Mark Hasegawa-Johnson 2007. Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics* 35. 180–209.

Cole, Jennifer, Gary Linebaugh, Cheyenne Munson & Bob McMurray 2010. Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics* 38 (2). 167–184.

Cole, Jennifer & Stefanie Shattuck-Hufnagel 2011. The phonology and phonetics of perceived prosody: What do listeners imitate? *Proceedings of INTERSPEECH-2011*, 969–972.

Cutler, Anne 2008. The abstract representations in speech processing. *Quarterly Journal of Experimental Psychology* 61 (11). 1601–1619. [omit? doi:10.1080/13803390802218542.]

Cutler, Anne 2010. Abstraction-based efficiency in the lexicon. *Laboratory Phonology* 1 (2). 301–318.

Dilley, Laura, Stefanie Shattuck-Hufnagel & Mari Ostendorf 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24 (4). 423–444.

Eisner, Frank, & James M. McQueen 2005. The specificity of perceptual learning in speech processing. *Perception & Psychophysics* 67 (2). 224–238.

Farnetani, Edda, & Daniel Recasens 1997. Coarticulation and connected speech processes. In William J. Hardcastle, John Laver and Fiona E. Gibbon (eds.), *The handbook of phonetic sciences*, 371–404. Wiley-Blackwell.

Goldinger, Steven D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105 (2). 251–279.

Gow, David W. 2003. Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics* 65 (4). 575–590.

Halle, Morris. 1992. Phonological features. In W. Bright (Ed.), *International encyclopedia of linguistics*, Vol. 3, 207–212. Oxford: Oxford University Press.

Kraljic, Tanya & Arthur G. Samuel 2005. Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology* 51 (2). 141–178.

Kraljic, Tanya & Arthur G. Samuel 2007. Perceptual adjustments to multiple speakers. *Journal of Memory and Language* 56 (1). 1–15.

Kraljic, Tanya, Arthur G. Samuel & Susan E. Brennan. 2008. First impressions and last resorts how listeners adjust to speaker variability. *Psychological Science* 19 (4). 332–338.

Mitterer, Holger & Miriam Ernestus. 2008. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition* 109 (1). 168–173.

Mixdorff, Hansjoerg, Jennifer Cole & Stefanie Shattuck-Hufnagel 2012. Prosodic similarity – evidence from an imitation study. *Proceedings of Speech Prosody 6* (Shanghai). 571–574.

Mo, Yoonsook 2011. *Prosody production and perception with conversational speech*. Urbana-Champagne: University of Illinois dissertation.

AU: Kindly provide the publisher location for 'Farnetani and Recasens 1997' in reference list entry.

Nielsen, Kuniko 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39 (2). 132–142.

Norris, Dennis, James M. McQueen & Anne Cutler 2003. Perceptual learning in speech. *Cognitive Psychology* 47 (2). 204–238.

Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America* 119 (4). 2382–2393.

Pardo, Jennifer S., Rachel Gibbons, Alexandra Suppes & Robert M. Krauss 2012. Phonetic convergence in college roommates. *Journal of Phonetics* 40 (1). 190–197.

Pardo, Jennifer S., Kelly Jordan, Rolliene Mallari, Caitlin Scanlon & Eva Lewandowski 2013. Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language* 69 (3). 183–195.

AU: Kindly provide the missing location for 'Rusaw and Cole 2011' in reference list entry.

Rusaw, Erin & Jennifer Cole 2011. Speech error evidence on the role of the vowel in syllable structure. *Proceedings of the International Congress on Phonetic Sciences*, 1734–1737.

Shattuck-Hufnagel, Stefanie & Nanette Veilleux 2007. Robustness of acoustic landmarks in spontaneously-spoken American English. *Proceedings of the 16th meeting of the International Congress of Phonetic Sciences, Saarbrueken*, 925–928.

Shockley, Kevin, Laura Sabadini & Carol A. Fowler 2004. Imitation in shadowing words. *Attention, Perception, & Psychophysics* 66 (3). 422–429.

Silverman, Kim E.A. & Janet B. Pierrehumbert 1990. The timing of prenuclear high accents in English. In John Kingston & Mary E. Beckman (eds.), *Papers in laboratory phonology 1: Between the grammar and the physics of speech*, 72–106. Cambridge: Cambridge University Press.

Stevens, Kenneth N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America* 111 (4). 1872–1891.

Turk, Alice E. & Stefanie Shattuck-Hufnagel 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics* 3 (4). 445–472.

AU: Kindly provide the missing location for 'Yoon et al. 2004' in reference list entry

Villacorta, Vergilio M., Joseph S. Perkell & Frank H. Guenther 2007. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America* 122 (4). 2306–2319.

Yoon, Taejin, Sandra Chavarría, Jennifer Cole & Mark Hasegawa-Johnson 2004. Intertranscriber reliability of prosodic labelling on telephone conversation using ToBI. *Proceedings of INTERSPEECH*-2004, 2729–2732.

AU: We have processed Appendix after References as per Source MS. Please confirm the placement for correctness.

APPENDIX

A1. Utterance 1: *Um, Kate, d'you see the Canadian Paradise?*

Words	**um**			**Kate**					**d'you**				**see**			**the**		
Prosody	H*	H-L%		H*			L-L%		H*				H*					
Phonemes	ʌ	m		k		ej	t		d		j	uw	s		ij	ð		ə
LMs-pred.	V-max	m-cl	m-rel	k-cl	k-rel	V	t-cl	t-rel	d-cl	d-rel	G	V-max	s-cl	s-rel	V-max	ð-cl	ð-rel	V-max
LMs-real.	V-max	m-cl	m-rel	k-cl	k-rel	V	t-cl	DEL	d-cl	d-rel	G	V-max	s-cl	s-rel	V-max	ð-cl	ð-rel	V-max

Words	**Canadian**											
Prosody												
Phonemes	k		ə	n		ej	d		ij	ə	n	
LMs-pred.	k-cl	k-rel	V-max	n-cl	n-rel	V-max	d-cl	d-rel	V-max	V-max	n-cl	n-rel
LMs-real.	k-cl	k-rel	DEL	n-cl	n-rel	V-max	G-min	--	V-max	V-max	n-cl	n-rel

Words	**Paradise**									
Prosody	L+H%									H-
Phonemes	p		æ	ɹ	ə	d		aj	s	
LMs-pred.	p-cl	p-rel	V	G-min	V-max	d-cl	d-rel	V-max	s-cl	s-rel
LMs-real.	p-cl	p-rel	V	G-min	V-max	d-cl	d-rel	V-max	s-cl	s-rel

A2. Utterance 2: *Um you're gonna be standing at the peak of the mountain on the Canadian Paradise.*

Words	**Um**			**you're**			**gonna**					**be**			
Prosody	H*		H-L%												
Phonemes	ʌ	m		j	ow	ɹ	g		ə	n		ə	b		ij
LMs-pred.	V-max	m-cl	m-rel	G-min	V-max	G-min	g-cl	g-rel	V-max	n-cl	n-rel	V-max	b-cl	b-rel	V-max
LMs-real.	V-max	m-cl	m-rel	G-min	V-max	G-min	g-cl	g-rel	V-max	d-cl	d-rel	V-max	b-cl	b-rel	V-max

Words	**Standing**												**at**		
Prosody	!H*										H-				
Phonemes	s		t		æ	n		d		ɪ	ŋ		æ	t	
LMs-pred.	s-cl	s-rel	t-cl	t-rel	V-max	n-cl	n-rel	d-cl	d-rel	V-max	ŋ-cl	ŋ-rel	V-max	t-cl	t-rel
LMs-real.	s-cl	s-rel	t-cl	t-rel	V-max	n-cl	n-rel	DEL	DEL	V-max	DEL	DEL	V-max	t-cl	t-rel

Words	**The**			**peak**					**of**			**the**		
Prosody				L+H*		L-								
Phonemes	ð		ə	p		ij	k		ə	v		ð		ə
LMs-pred.	ð-cl	ð-rel	V-max	p-cl	p-rel	V	k-cl	k-rel	V-max	v-cl	v-rel	ð-cl	ð-rel	V-max
LMs-real.	DEL	ð-rel	V-max	p-cl	p-rel	V	k-cl	k-rel	V-max	v-cl	MERGED		ð-rel	V-max

Words	**of**			**The**			**mountain**									**on**		
Prosody							L*									H-L%		
Phonemes	ə	v		ð		ə	m		aw	n		t		ə	n		a	n
LMs-pred.	V	v-cl	v	ð-cl	ð-rel	V	m-cl	m-rel	V	n-cl	n	t-cl	t-rel	V	n-cl	n-rel	V-max	n-cl	n-rel
LMs-real.	V	v-cl	MERGED		ð-rel	V	m-cl	m-rel	V	DEL	DEL	t-cl	t-tel	V	n-cl	DEL	V-max	DEL	DEL

Words	**The**			**Canadian**											**P...**
Prosody															
Phonemes	ð		Ax	k		ax	n		ej	d		ij	ax	n		p
LMs-pred.	ð-cl	ð-rel	V-max	k-cl	k-rel	V-max	n-cl	n-rel	V-max	d-cl	d-rel	V-max	V-max	n-cl	n-rel	p-cl
LMs-real.	ð-cl	ð-rel	DEL	SUBS x-cl	k-rel	DEL	DEL	DEL	V-max	G-min	DEL	V-max	DEL	n-cl	MERGED	

Words	**...n**	**Paradise**									
Prosody		L*									H-H%
Phonemes		p		æ	ɹ	ə	d		ay	s	
LMs-pred.	n-rel	p-cl	p-rel	V-max	G-min	V-max	d-cl	d-rel	V-max	s-cl	s-rel
LMs-real.	MERGED		p-rel	V-max	G-min	V-max	d-cl	DEL	V-max	s-cl	s-rel

Figures A1 and A2: Stimulus Utterances 1 (A1) and 2 (A2), used for elicited imitations. The second row displays the prosodic feature (pitch accent and boundary tone) as ToBI-labelled by the authors, and left blank if no accent or boundary label was assigned. The third row displays the phonemes specified for the unreduced form of the word. The fourth row displays the predicted LMs for each phoneme, and the bottom row displays the LMs as realized by the speaker (LM outcomes) and labelled by the authors. Cells for realized LMs (bottom row) that are deleted, substituted or merged relative to the predicted LMs are lightly shaded. Dark shading is used for cells corresponding to the words, at and on, in Utterance 2 that are excluded from analysis are the imitated utterance (see text for further details).

AU: Please confirm whether Figures A1 and A2 can be changed to Tables A1 and A2.

AU: Please rephrase the sentence "Dark shading is used for cells corresponding to the words, at ...", in the caption of Figures A1 and A2, for clarity.