A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition

Laura C. Dilley[a] and Mark A. Pitt
Department of Psychology, Ohio State University, 1835 Neil Ave., Columbus, OH 43220

(Received 27 June 2006; revised 14 June 2007; accepted 23 July 2007)

Regressive place assimilation is a form of pronunciation variation in which a word-final alveolar sound takes the place of articulation of a following labial or velar sound, as when *green boat* is pronounced *greem boat*. How listeners recover the intended word (e.g., *green*, given *greem*) has been a major focus of spoken word recognition theories. However, the extent to which this variation occurs in casual, unscripted speech has previously not been reported. Two studies of pronunciation variation were conducted using a spontaneous speech corpus. First, phonetic labeling data were used to identify contexts in which assimilation could occur, namely, when a word-final alveolar stop (/t/, /d/, or /n/) was followed by a velar or labial consonant. Assimilation was indicated relatively infrequently, while deletion, glottalization, or canonical pronunciations were more often indicated. Moreover, lexical frequency was shown to affect pronunciation; high frequency lexical items showed more types of variation. Second, acoustic analyses showed that neither place of articulation cues (indicated by second formant variation) nor relative amplitude was sufficient to distinguish assimilated from deleted and canonical variants; only when closure duration was additionally taken into account were these three variant types distinguishable. Implications for theories of word recognition are discussed. © *2007 Acoustical Society of America.* [DOI: 10.1121/1.2772226]

PACS number(s): 43.71.An, 43.70.Fq, 43.71.Es [MSS] Pages: 2340–2353

I. INTRODUCTION

There can be a great deal of variability in how words are pronounced (e.g., Dalby, 1986; Bell *et al.*, 2003; Shockey, 2003; Johnson, 2004), yet listeners rarely experience difficulty in understanding what is said. One type of pronunciation variation that occurs in connected speech is regressive place assimilation, in which the final alveolar segment of a word is produced at the same place of articulation as a following segment. For instance, the /n/ at the end of *green* may take the labial place of the following /b/ in the phrase *green boats*, so that *green* appears to be pronounced as *greem*.

How do listeners recognize the intended word (i.e., *green* given *greem*) when assimilation occurs? This question has received considerable attention in the experimental literature, with multiple theoretical accounts being put forth (e.g., Lahiri and Marslen-Wilson, 1991; Gaskell and Marslen-Wilson, 1998; Gow, 2003). Although this work has been informed by a large body of research on the articulatory and acoustic characteristics of assimilation (e.g., Wright and Kerswill, 1989; Holst and Nolan, 1995; Zsiga, 1995; Ellis and Hardcastle, 2002), there remain unanswered questions about the scope and nature of the problem that assimilation poses for recognition. For example, at a phonological level, little is known about the frequency of assimilation relative to other types of pronunciation variation that might also occur in contexts in which assimilation is possible (i.e., in assimilable environments). Furthermore, most articulatory and acoustic studies have examined assimilation in read speech, raising questions about how representative these data are of the unscripted, conversational speech to which listeners are most frequently exposed. Thus, understanding the extent to which assimilation and other connected speech processes occur in spontaneous speech is necessary to ensure theories of spoken word recognition can adequately and accurately account for how pronunciation variants are recognized.

The use of speech corpora to inform theorizing and experimentation in speech perception and production has been increasing in recent years. One early study by Dalby (1986) investigated effects of speaking rate on the likelihood of deleting schwa vowels in American English. The results showed schwas deleted more often at fast speech rates and in word-medial position, but no differences were found as a function of lexical stress environment. Subsequently, Patterson *et al.* (2003) examined frequency of schwa deletion in conversational American English. They found that lexical stress environment was the most important factor in predicting deletion of schwa vowels, in contrast to Dalby (1986). [See Crystal and House (1988) for a comparison of stress and speech rate effects on segment realizations in speech corpora.]

Moreover, Patterson and Connine (2001) investigated frequency of allophonic variants of word-medial /t/ in conversational American English. They showed that lexical items which were high frequency and/or less complex morphologically were more likely to show /t/ realized as the flapped variant [ɾ].[1] In Dutch, Mitterer and Ernestus (2006) used read and spontaneous speech corpora to investigate optional word-final /t/-lenition. They found that /t/-lenition occurred more often in spontaneous than read speech, but could identify no context which consistently induced /t/-lenition.

[a]Electronic mail: dilley@bgsu.edu. The author is now with the Department of Communication Disorders and Department of Psychology, Bowling Green State University, 247 Health Center, Bowling Green, OH 43403.

Additional studies on Dutch have shown that word frequency influences the likelihood of voicing assimilation (Ernestus *et al.*, 2006) and of short durations for affixes (Pluymaekers *et al.*, 2005). Moreover, the studies by Ernestus *et al.* (2006) as well as Snoeren *et al.* (2006) for French each suggest that voicing assimilation is graded, rather than categorical; both studies used read speech. Among other things, these results highlight the fact that speech style (e.g., read versus spontaneous) affects rates of variant modifications. The current research builds on this work by analyzing the types of variation occurring in environments where regressive place assimilation is phonologically possible in spontaneous, unscripted American English.

Three key findings about regressive assimilation have emerged from the empirical literature, almost all of which has used read speech. First, an alveolar stop which assimilates to a following labial or velar consonant often shows acoustic or articulatory evidence of both alveolar place of articulation, as well as labial or velar places of articulation, respectively (e.g., Kohler, 1990; Ohala, 1990; Barry, 1992; Ellis and Hardcastle, 2002; Gow, 2001, 2002, 2003). For example, electropalatography (EPG) and electromagnetic articulography data have demonstrated evidence of partial alveolar assimilations in alveolar-velar (e.g., /d#g/) sequences, indicating a tongue blade/body gesture along with velar contact of the tongue dorsum (Wright and Kerswill, 1989; Barry, 1992; Nolan, 1992; Zsiga, 1995). Similarly, acoustic studies have demonstrated that for assimilated alveolar stops followed by labials, the mean formant frequencies are intermediate between those of unassimilated alveolar and labial sounds (Gow, 2001, 2002, 2003).

A second finding is that assimilation often leads to gradient cues to place of articulation of the word-final segment (Gow, 2001, 2002, 2003; Holst and Nolan, 1995; Nolan *et al.*, 1996). For example, in an EPG study of alveolar-velar sequences (e.g., *road collapsed*) Wright and Kerswill (1989) found that speakers produce varying alveolar and velar contact, ranging from full alveolar with no velar contact, to a mixture of alveolar and velar contact, to full velar with no alveolar contact. This gradience is at least partially due to the fact that the degree of assimilation produced often varies both within and across talkers (e.g., Nolan *et al.*, 1996; Ellis and Hardcastle, 2002). That this gradience is relevant for processing has been demonstrated through perceptual studies of assimilation cues (Nolan, 1992; Gow and McMurray, in press). Such findings are reminiscent of work showing variability within phonetic categories and listener sensitivity to this variation (e.g., Miller, 2001), as well as variability in phonetic realizations as a function of lexical items, e.g., the tendency in American English to produce /t/ as [ɾ] in *pretty* (Patterson and Connine, 2001; Connine, 2004).

A third finding is that in cases where alveolar assimilation is extreme, assimilated forms and underlying nonalveolar forms can be indistinguishable (Holst and Nolan, 1995; Nolan *et al.*, 1996). For example, Holst and Nolan (1995) found on the basis of acoustic measurements that the assimilation of /s/ to [ʃ] was rarely differentiable from the underlying form (/ʃ/ to [ʃ]). Other studies have replicated this result using articulatory data (Nolan *et al.*, 1996; Ellis and Hardcastle, 2002).

These three findings collectively suggest that except in cases of the most extreme alveolar assimilation, assimilated segments are likely differentiable acoustically and/or articulatorily from canonical forms. A few perceptual studies have shown listeners can distinguish between alveolars realized as assimilated versus unassimilated as well (e.g., Nolan, 1992; Gow and McMurray, in press). However, the usefulness of remnant cues to underlying alveolar place for assimilated consonants in perception depends in part on these being reliably present. In this regard, it is not clear whether in spontaneous, unscripted speech alveolars in assimilable environments will retain remnants of their underlying place of articulation. Previous work suggests that place assimilation is more common during casual speaking styles, as well as at faster speech rates (e.g., Barry, 1992); however, these results were obtained using scripted speech. Spontaneous speech is known to show different acoustic-phonetic attributes from read speech, including greater gestural overlap, a higher degree of segmental deletion, and different strength of consonantal gestures (e.g., Browman and Goldstein, 1990; Johnson, 2004; Shockey, 2003).

Our goal in the present investigation was to build on prior research by examining a corpus of spontaneous speech to determine for assimilable environments the extent to which assimilation versus other forms of pronunciation variation occurred. This broad perspective enabled us to define and compare assimilation relative to other forms of word-final variation. These data should in turn inform theorizing about recognizing assimilated word forms by clarifying the challenges that the perceptual system faces.

There were two parts to the study. First, we examined the relative frequencies of distinct phonetic variants occurring in assimilable environments as given by phonetic labels in the speech corpus, in order to determine the consistency with which assimilation occurs. Second, we evaluated the extent to which assimilated alveolars are acoustically differentiable from unassimilated alveolars, labials, and velars. To ensure the generality of the results, we examined variation in multiple word-final alveolar segments (/t/, /d/, and /n/) in the context of a following velar (/g/ or /k/) or labial (/p/, /b/, or /m/) segment. In addition, this was done when the context preceding the alveolar was labeled as a high front vowel ([ɪ]) and a nonhigh front vowel ([æ] or [ɛ]).

II. PHONETIC LABELING ANALYSIS

A. Method

The relative frequency of regressive place assimilation was investigated using phonetic labels from the Buckeye Corpus of Conversational Speech (Pitt *et al.*, 2006); this corpus is comprised of conversations from talkers in the Columbus, OH area. Phonetic transcriptions were made by a group of trained phonetic labelers, who were paid for corpus preparation. Labelers used spectrogram and waveform displays in Xwaves software (Entropies Corp.), as well as auditory cues, to label the phonetic segments present in the speech accord-

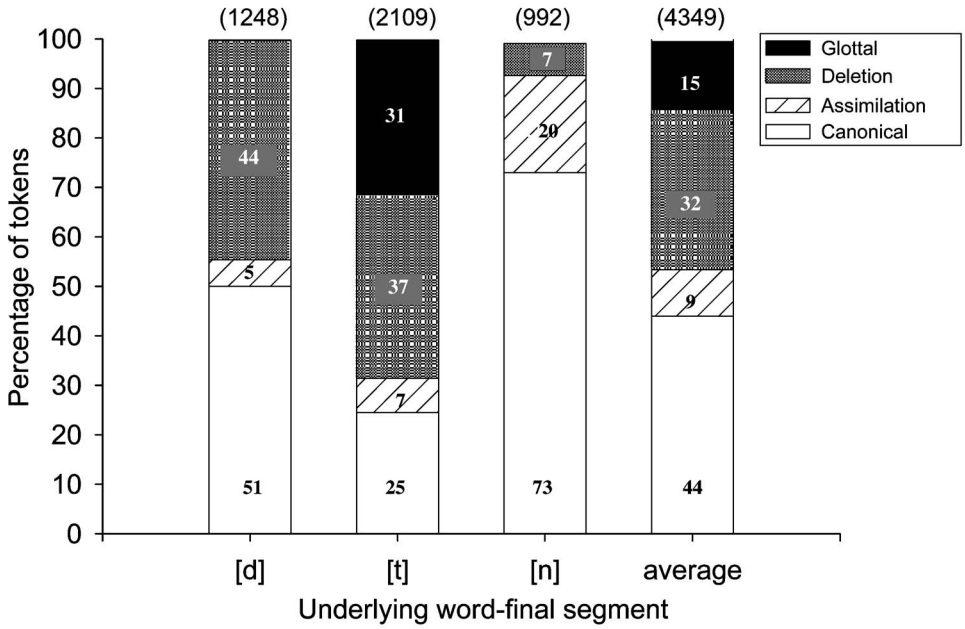

FIG. 1. Percentage of tokens labeled as each of the four variants for the three word-final alveolars in an assimilable context. The number of tokens of each alveolar is listed in parentheses.

ing to the labeling conventions described in Kiesling *et al.* (2006). Of relevance to the present study were conventions for labeling segments as assimilated, deleted, glottalized, and canonical. Labelers indicated a segment to be *assimilated* when (1) the perceptual evidence was consistent with a word-final alveolar stop adopting a labial or velar place of articulation, and (2) there was spectral evidence of a local change in F2 in a sonorant segment just preceding the stop closure (i.e., F2 fell or rose in the case of an alveolar stop adopting a labial or a velar place or articulation, respectively). Moreover, labelers indicated a segment to be *glottalized* when the segment had perceptually creaky voicing accompanied by irregularity in pitch period timing in the waveform. (See Dilley *et al.*, 1996; Redi and Shattuck-Hufnagel, 2001.) In addition, labelers indicated a segment to be *deleted* when it could not be heard when a short context was played and there was no clear visual evidence in the spectrogram that the segment was present. Finally, labelers indicated a segment to be *canonical* when it was perceived as present and unassimilated, lacked creaky voicing, and/or when they were uncertain about the variant type.[2]

Tests of transcription consistency and agreement among labelers were performed periodically during creation of the corpus, and have indicated a high degree of reliability in the use of the phonetic labels. In particular, in a published study of intertranscriber reliability (Pitt *et al.*, 2005) using four labelers annotating 4 min of speech, overall agreement for all phonetic labels was 80.3%, with agreement of 92.9% for stops. A more recent unpublished test of intertranscriber reliability using eight labelers and 1 min of speech allowed us to specifically investigate agreement among canonical, deleted, and glottal variants. Agreement for these variants was 85.2%, indicating high reliability in line with previous findings of good interrater agreement (e.g., Irwin, 1970; Eisen, 1991).[3]

The speech of 19 talkers (9 male, 10 female; approximately 138 000 words) was used to identify lexical sequences constituting assimilable environments, i.e., two-word sequences in which the place of articulation of the word-final phoneme could assimilate to that of a following word-initial phoneme. Analyses were limited to word-final alveolars (/t/, /d/, or /n/) that were followed by word-initial labials (/b/, /p/, or /m/) or velars (/g/ or /k/), since these environments are subject to processes of place assimilation in English (Shockey, 2003). These phonological environments were identified using citation pronunciations obtained from a phonetic dictionary (CMU pronouncing dictionary 0.6), using the orthographic transcriptions of the conversations. Note that two-word sequences which were labeled by phonetic analysts as showing a fluent or nonfluent pause at the word boundary were eliminated from analysis. [See Kiesling *et al.* (2006) and footnote 2 for more details.]

To determine what phonetic changes occurred in assimilable environments, the "underlying" or citation pronunciations were compared with the "surface" or actual pronunciations of the words in the corpus. The four surface forms that underlying word-final alveolar sounds could take based on transcription conventions were assimilated, deleted, glottalized, or canonical. Finally, note that word-final alveolars could not be realized as a flap in the contexts under investigation, because such variations are limited to positions between vocalic or sonorant segments (De Jong, 1998).

B. Results and discussion

There were a total of 4349 assimilable contexts. These are shown in Fig. 1, which gives the rate of each of four possible variant realizations (assimilation, deletion, glottal, canonical) as a function of underlying segment type (/t/, /d/, or /n/). The number of tokens included for each of the three

TABLE I. Percentage of tokens in assimilable environments as a function of form class, number of syllables, and assigned label. All percentages have been rounded.

	Function word		Content word	
	Monosyllabic	Polysyllabic	Monosyllabic	Polysyllabic
Canonical	15	3	18	6
Assimilation	5	0	4	1
Deletion	15	4	9	4
Glottal	9	1	4	0
SUM	44	8	35	11

segments is in parentheses at the top of the graph. The canonical variant constituted the most frequent type of surface realization, occurring 44% of the time overall. Noncanonical surface realizations were quite frequent as well, constituting 27%–75% of instances across the three segments. Assimilation was indicated infrequently, occurring only 9% of the time across the three segmental environments; it was least common for /t/ and /d/ (5% and 7%, respectively) and most common for /n/ (20%). Deletions were also quite common, especially for the oral stops; they constituted 45% and 37% of the /d/ and /t/ realizations, respectively. Finally, glottal variants were found almost exclusively for /t/, for which they occurred almost as often as deletions (31%).

The high rate of deletion for /t/ and /d/ was somewhat unexpected. It has been reported that /t/ and /d/ readily delete in the context of a preceding /n/ both word-medially (Raymond *et al.*, 2006) and word-finally (Guy, 1980; Neu, 1980). Consistent with this earlier work, a high percentage (53%, $N=1408$) of deleted /t/ and /d/ tokens were found to have been preceded either by /n/ or syllabic /n/. No other systematicities were identified in the remaining cases of deletion.

The data in Fig. 1 were analyzed further to provide a more complete picture of phonological variation in assimilable environments. We began by calculating the frequency of the four realizations as a function of word class (function or content word) and length. Monosyllables, which constituted 79% of the tokens, were compared with polysyllabic words. The data were combined over underlying segment identity (/t, d, n/) because all showed a consistent pattern at this level of analysis. The data are shown in Table I.

Comparison of the totals in the last row shows that function and content words occur equally often in assimilable environments. Given the small number of function words in English relative to content words, these data partially forecast what will become clear shortly, that a small number of function words make up the majority of items in assimilable environments. Table I also reveals usage statistics in the language: Monosyllables make up a higher percentage of assimilable tokens for both word form types, consistent with the fact that monosyllables make up 80% of the tokens in the Buckeye Corpus (Pitt *et al.*, 2005).

More interesting in Table I, however, is how the percentages of tokens differ across pronunciation variants. For monosyllabic function words, for instance, deletions are as frequent as canonical pronunciations. In contrast, monosyllabic content words show a canonical bias of approximately 2:1 relative to deletions. This asymmetry in frequency of canonical versus deleted realizations is present for polysyllabic words as well, only to a much smaller degree. Thus, function words appear more likely to deviate from their canonical pronunciations than content words. [See Raymond *et al.* (2006) for a similar finding.] Moreover, rates of assimilation do not differ appreciably for monosyllabic function and content words; both are relatively rare. Finally, glottal realizations were more common for function words than for content words.

The next analysis examined the frequency of different variant realizations for individual words represented in Table I. Given a particular word, was that word more likely to be realized in mainly one way (e.g., canonical only) or in more than one way (e.g., both canonical and deleted)? The data are shown in Table II, with the upper half reflecting word types and the lower half number of tokens. The values in each cell are percentages of the total number of types or tokens. The column labels designate the type(s) of variation exhibited by each word final segment (D=deletion, A=assimilation, C=canonical, G=glottal), with multiple letters indicating that the word-final segment showed each of those realizations at least once. Because the segments /d/ and /n/ were

TABLE II. Percentage of word types (upper half) and word tokens (lower half) labeled as having been realized as each of four variant categories (D=deletion, A=assimilation, C=canonical, G=glottal), broken down by final segment (/t/, /d/, or /n/). Categories with two or more labels (e.g., DA) indicate the word-final segment was indicated as having been realized as more than one variant type. Note that possible categories for /t/ are listed separately because glottals rarely occurred for /d/ or /n/. See the text for further details.

	Labeling categories							
Types	D	A	C	DA	DC	AC	DAC	
/d/	8	5	60	2	8	9	8	
/n/	4	10	58	3	3	14	9	
	DG	AG	CG	DAG	DCG	ACG	DACG	G
/t/	32	5	29	1	22	2	9	1
Tokens	D	A	C	DA	DC	AC	DAC	
/d/	6	1	15	1	35	12	32	
/n/	1	2	14	1	2	15	65	
	DG	AG	CG	DAG	DCG	ACG	DACG	G
/t/	7	1	5	1	22	1	65	0

almost never labeled as glottal variants (occurring just twice for /d/), only seven categories are given for these phones, as indicated by seven columns for single or multiple realization types. In contrast, /t/ was additionally labeled as a glottal variant very frequently, so up to 15 combinations of single or multiple realizations were possible. However, glottal variants were almost always a possible realization of /t/-final words in addition to at least one other variant type, so that many possible cells involving single realization types for /t/ were empty. Thus, only 8 of 15 possible combinations of realizations are shown in Table II (since the other possible combinations had $N=0$). In many ways, the results across categories were similar for /t/-final words as for /d/ and /n/-final words, except for the additional glottal variation.

Consider the data in the first three columns in which a variant was realized in only one way (two for /t/). For deletions, assimilations, and glottals the percentage of types and tokens is small (less than 14%), indicating few words are spoken only in a reduced form. The one exception is /t/, for which 40% of the /t/-final words were never spoken canonically. For canonical realizations, there is a marked difference in frequency between types and tokens. Type frequency is enormous, greater than 50% in the cases of /d/ and /n/, indicating that a large number of words were spoken only in their canonical form. The corresponding token frequencies are four times smaller, indicating that these words were spoken infrequently in the corpus. Note also that this pattern was greater overall for /d/ and /n/ than for /t/.

The final segment was realized in two ways (three for /t/) for items represented in the middle section of Table II, and the same bias for the canonical pronunciation is present here. Word-final segments that were realized as DA (deleted or assimilated) occurred less often than the combinations of DC and AC, both for types and for tokens. In general, percentages were low across the DA, DC, and AC columns, and there is not much change between type and token values, except for DC. Here, the token frequency for /d/ is fourfold that of type frequency, just the reverse of what was found when only one realization occurred (first three columns). In this case, a small set of words is responsible for a disproportionately large amount of variation. This same pattern of a few types being responsible for most variation is found in the DAC (DACG) column, where words were realized in all three (or four) ways. The result is particularly dramatic for the /n/ and /t/ words. Although they constituted less than 10% of the word types in the corpus, they accounted for almost two-thirds of the tokens that underwent variation.

These data indicate that most of the word-final variation in assimilable environments occurs in a small subset of words whose final segment can assimilate, delete, and in the case of /t/, glottalize. What are these words? Shown in Fig. 2 are the 21 most frequent words in our sample broken down by type of variation. They are grouped by the identity of the final segment. Together, they constitute 67% of the sample (2956 tokens). Most of the 21 are monosyllabic (19) and function words (14). Most of the /d/-final and /n/-final words show evidence of assimilation and deletion, though more of the former, except for *and*, whose final /d/ deletes most often (Neu, 1980). Many /t/-final words are realized in all four ways. The preponderance of glottal forms for /t/-final words is particularly striking. These data provide a highly representative depiction of the variation found in assimilable environments. Furthermore, it is representative across talkers, too, as most speakers (mean=16.9 out of 19) contributed tokens to the counts of each word.

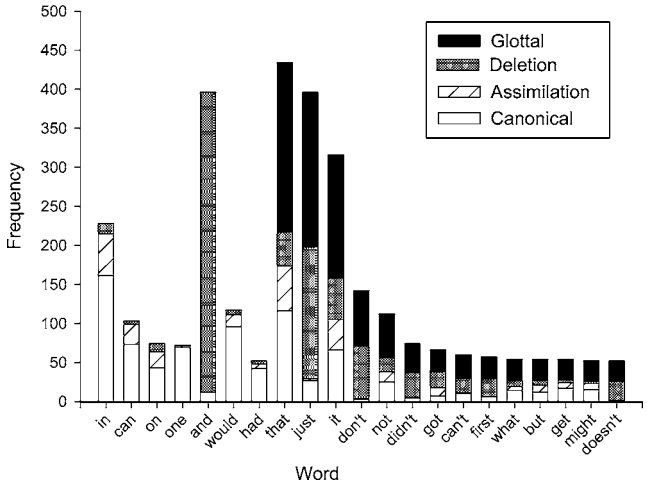

FIG. 2. Variant rates for the 21 most frequent words in the phonetic labeling analysis.

In sum, the phonetic analysis demonstrates that assimilated segments were heard relatively infrequently in environments where they might be expected to be found. By contrast, other phonetic realizations were more common, with canonical pronunciations and deletions predominating. In addition, variant realization varied by final segment type (/t/, /d/, or /n/). Finally, there was considerable variation according to word types, with a small number of very frequent items exhibiting all forms of variation and many other lower frequency items being realized only canonically.

These results suggest that recognizing a word in an assimilable environment is not just a matter of distinguishing between a surface (assimilated) form and an underlying (canonical) form. The large number of deletions, as well as glottal variants, indicate that the problem of recognition, even in this highly constrained context, is more complex than has often been assumed, and suggests that a high degree of flexibility in variant recognition is necessary for word recovery to succeed. In Sec. III we take a closer look at this variation by analyzing some acoustic characteristics in the vicinity of key word-final segments.

III. ACOUSTIC ANALYSIS

A. Method

Acoustic analyses were undertaken in order to probe further the nature of variation in word-final alveolars in assimilable environments. The second formant (F2) of a vowel is strongly affected by the place of articulation of an upcoming or preceding consonant (Stevens, 1998). Therefore, F2 was measured in the vicinity of the underlying word-final alveolars to estimate the degree to which cues to the place of articulation of the following word-initial consonant were indicated. To ensure comparable phonological contexts, analy-

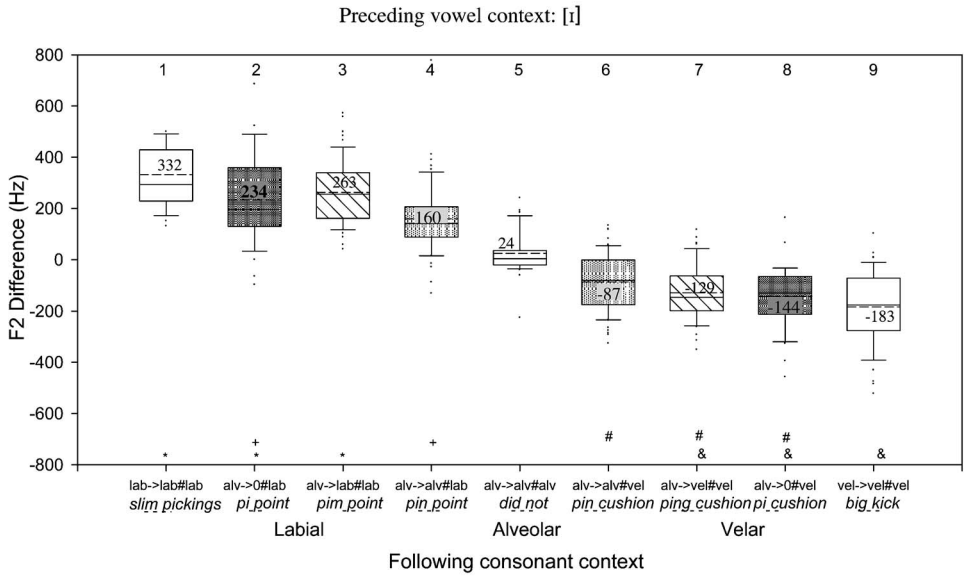

FIG. 3. Box plots of the F2 difference in an [ɪ] vowel show nine labeling conditions in which segments were followed by labial (bars 1–4), alveolar (bar 5), and velar (bars 6–9) contexts. Labels on the *x* axis describe the conditions, as well as how the word final tokens in those conditions were labeled. The underlying segment is listed first, followed after the arrow by its surface (i.e., labeled) realization. The following context is listed after the word boundary symbol (#). lab, alv, and vel refer to labial, alveolar, and velar places of articulation, respectively. 0 denotes the segment was deleted. Conditions with the white bars contain data from control contexts in which the place of articulation across the word boundary is the same. In bars 2–4 and 6–8, an underlying alveolar segment is in the context of a following labial or velar segment, respectively. In these conditions, the light-grey bars represent data from tokens labeled as canonical or assimilated, and the dark grey bars, deletions. The solid line in the middle box is the median of the distribution. The dashed line is the mean, with its value listed. Symbols at the bottom of the graph denote which conditions are statistically indistinguishable from each other.

ses were limited to tokens in which the underlying word-final alveolar was preceded by a vowel. Moreover, because the tongue height and advancement of vowels also affect F2 (Stevens, 1998), the identity of vowels in contexts preceding the underlying word-final alveolars was also controlled. Tokens in two kinds of preceding vocalic environments were selected for analysis: (1) the front high vowel [ɪ] as in *sit*, (2) the front nonhigh vowels [æ] and [ɛ] as in *sat* and *set*. These contexts were selected because they were most numerous in the corpus. Tokens with these vowels were identified using the phonetic labels in the corpus.

For each of these two vowel contexts, F2 difference values were determined for underlying alveolar consonants in a total of six assimilable conditions, arising from three possible segmental realizations for these alveolars (canonical, assimilated, or deleted), given either of two kinds of following word-initial consonants (labial or velar).[4] To assess the degree of F2 variation in underlying alveolar tokens, F2 difference measurements were obtained from vowels in control contexts in three nonassimilable environments which had homorganic places of articulation across the word boundary: Word-final labials preceding labials (e.g., *him pay*), word-final alveolars preceding alveolars (e.g., *did not*), and word-final velars preceding velars (e.g., *anything comes*). This gave nine token conditions for each vowel context, for a total of 737 tokens across both vowel contexts. For the [ɪ] context, the average number of tokens in each condition was 49.7 (range 22–71). For [æ] and [ɛ] contexts, the average number was 32.2 (range 9–50).

To calculate the F2 difference (measured in hertz), F2 was measured at two points: (1) the vowel midpoint and (2) at the final pitch period of the vowel, just before consonantal closure (cf. Pitt and Johnson, 2003). A combination of automatic and hand measurements were used to estimate the F2 difference. Automatic measurements were obtained by extracting formant frequencies using an Xwaves script. Hand measurements were made from spectra generated from a 25 ms Hanning window centered on the zero crossings of the pitch periods closest to the vowel midpoint and vowel end point, and/or by measuring F2 values from wide-band spectrograms.

B. Results

Figure 3 shows box plots of the differences in F2 values for a preceding [ɪ] vowel for the six assimilable and three control conditions. The bars are ordered with respect to place of articulation of the following word-initial consonant, with the data from labial contexts in bars 1–4 and velar contexts in bars 6–9; bar 5 shows the data from the alveolar context. The dark grey bars (2 and 8) represent instances of deletion. The bars with hash marks represent cases in which underlying alveolars were classified as assimilated or canonical. The white bars (1, 5, 9) are the control conditions in which place of articulation was the same underlyingly across the word boundary (e.g., *him pay*). Recall that these were included as referents against which to compare the direction and extent of F2 deviation in the other conditions. The horizontal bar through the middle of each box is the median of the distribution. The dashed line is the mean, with it value given.

First consider the data on the left-hand side of the graph, where segment realization could be influenced by a following labial. For alveolar segments labeled as assimilated (bar 3), the extent of the F2 transitions approaches that of true

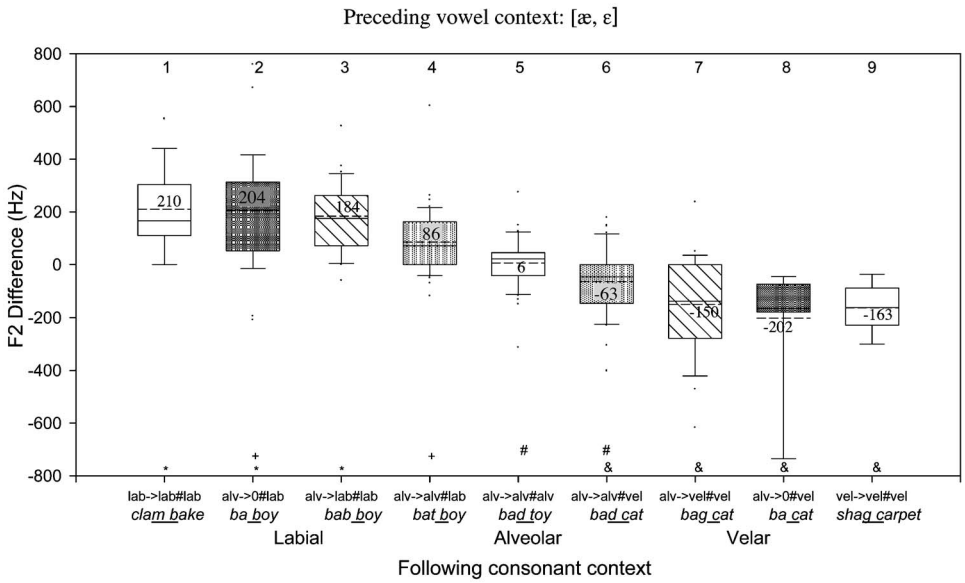

FIG. 4. Box plots of the F2 difference in [æ] and [ɛ] vowels show nine labeling conditions in which segments were followed by labial (bars 1–4) alveolar (bar 5), and velar (bars 6–9) contexts. Labels on the *x* axis describe the conditions, as well as how the word final tokens in those conditions were labeled. The underlying segment is listed first, followed after the arrow by its surface (i.e., labeled) realization. The following context is listed after the word boundary symbol (#). lab, alv, and vel refer to labial, alveolar, and velar places of articulation, respectively. 0 denotes the segment was deleted. Conditions with the white bars contain data from control contexts in which the place of articulation across the word boundary is the same. In bars 2–4 and 6–8, an underlying alveolar segment is in the context of a following labial or velar segment, respectively. In these conditions, the light-grey bars represent data from tokens labeled as canonical or assimilated, and the dark grey bars, deletions.

labials (bar 1), but falls short on average by 79 Hz. Nevertheless, relative to the alveolar context (bar 5, 239 Hz difference), segments labeled as assimilated are much more similar to true labials. For underlying alveolar segments labeled as deleted (bar 2), not only is the F2 transition affected by the labial place of articulation of the following segment, it is affected to an extent almost as great as segments labeled as assimilated, falling short of this category by a mere 29 Hz. In terms of F2 transition information, assimilated and deleted categories are virtually indistinguishable, as indicated by the fact that their middle quartiles are highly similar. In addition, each distribution overlaps extensively with the labial distribution (bar 1).

Next, for alveolar segments labeled as canonical in the context of a following labial (bar 4), the amount of F2 change is not as great as in assimilated cases (bar 3), as expected. However, it is surprisingly large, with the bulk of the distribution clearly separate from that of the control alveolar context (bar 5). The frequency drop in F2 of these alveolars suggests some degree of assimilation has taken place, but the drop was not enough for labelers to classify the segment as assimilated. These data indicate that a substantial amount of F2 deviation is tolerated before a segment is classified as labial (or missing), despite the fact that this cue is almost certainly being used by labelers to distinguish alveolars from labials (comparing bars 1 and 5).

Similar results are found for alveolars in a following velar context (bars 6–9), but in the opposite direction. For alveolars labeled as having assimilated to a following velar (bar 7), F2 transitions frequently extend as far as true velars (bar 9), falling short on average by 54 Hz. For alveolars labeled as deleted (bar 8), F2 is affected by the place of the following segment to an even greater degree than segments labeled as assimilated, differing from true velars by 39 Hz. Just as in labial contexts, assimilated and deleted categories in velar contexts appear indistinguishable on the basis of F2 information. In addition, relative to the alveolar context (bar 5), deleted and assimilated segments are much more similar to velars. Finally, for alveolar segments labeled as canonical (bar 6), the extent of F2 change is again large, with the bulk of the distribution offset from that of control alveolar contexts (bar 5), but not entirely overlapping with that of a segment labeled as assimilated (bar 7). The data from velar contexts again suggest that large F2 transitions must be present before an assimilated segment (or no segment) is perceived.

Statistical analyses were carried out to determine which distributions were reliably different from one another. A one-way analysis of variance (ANOVA) showed a significant main effect of condition, $F(8,438)=111.4$, $p<0.01$. The symbols at the bottom of the graph specify which conditions differed reliably in Tukey's honestly significantly different (HSD) post hoc tests. Conditions that share the same symbol at the bottom of the graph are statistically indistinguishable (e.g., 1–3; 2 and 4). The results are virtually symmetrical across the labial and velar contexts. The deleted and assimilated distributions are not only statistically equivalent to each other, but also to the corresponding control (labial and velar) distributions. The main difference across contexts is that the alveolars (bars 4 and 6) are more distinguishable when the following segment is labial than velar.

F2 transition analyses in [æ] and [ɛ] vowel contexts are displayed in Fig. 4; these contexts replicate what was found for the [ɪ] context. For alveolars labeled as having assimilated to a following labial or velar (bars 3 and 7, respectively), the extent of the F2 transitions are comparable to

those of true labials and velars (bars 1 and 9), with the distributions overlapping almost completely. Similarly, for alveolars labeled as having been deleted in the context of a following labial or velar (bars 2 and 8), the distributions overlap fully in the labial context and greatly in the velar context. (Note that the large variability for bar 8 is likely due to the small number of cases of alveolars labeled as having been deleted in the context of velars in our corpus, $n=11$.) Relative to the alveolar context (bar 5), alveolars labeled as assimilated or deleted (bars 2, 3, 7, and 8) are much more similar to labials and velars (bars 1 and 9). These results provide further evidence that assimilated and deleted categories are indistinguishable on the basis of F2 information. Finally, alveolar segments labeled as canonical in labial and velar contexts (bars 4 and 6, respectively) show rather large F2 transitions, such that the bulk of the distributions are separate from that of control alveolar contexts (bar 5). However, these distributions do not entirely overlap with those of segments labeled as assimilated (bars 3 and 7, respectively). This again suggests that the extent of F2 transition can be substantial before a different category is indicated.

A one-way ANOVA across the nine conditions was reliable, $F(8,280)=30.75$, $p<0.01$. Post hoc Tukey's comparisons for labial contexts yielded an almost identical pattern of reliable differences as in Fig. 3. For velar contexts, the differences between distributions were not statistically reliable, probably because of the somewhat greater dispersion of values.

A final analysis of the F2 difference measurements involved combining the data across vowel contexts into a single ANOVA with two variables, preceding vowel context ([ɪ] vs [æ] and [ɛ]) and word-boundary condition (1–9). This analysis was performed mainly to increase the stability of the results, especially in the few cells in which observations were low. The main effect of the word-boundary condition was reliable, $F(8,718)=110.85$, $p<0.001$, and post hoc Tukey's tests showed even more clearly that deletions are indistinguishable from assimilations. Both assimilated and deleted distributions approximate quite closely the control distributions, be they labial or velar; conditions 1–3 were not reliably different, nor were 7–9. Conditions 4–6, in contrast, differed reliably from each other and all other conditions. The only exception to this was condition 6, which was not reliably different from 7. The ANOVA also yielded a significant main effect of vowel context, $F(1,718)=11.857$, $p<0.001$, with F2 differences being smaller overall for [æ] and [ɛ] than [ɪ]. The interaction of the two variables was also reliable, $F(8,718)=2.30$, $p<0.02$, with the F2 difference being smaller in [æ] and [ɛ] contexts when the following environment was labial, but much more comparable to [ɪ] when the environment was velar.

The results of the F2 analyses are quite consistent, replicating across two different preceding vowel contexts. Moreover, by detailing variation in informal, unscripted speech, the present results provide a snapshot of the kinds of variation that listeners typically encounter in conversation, and thus what models must account for in explaining how spoken words are recognized. These findings agree with studies using casually produced read speech, which have shown that assimilated alveolars often exhibit values in place of articulation metrics which are intermediate between alveolar and nonalveolar place (e.g., Gow, 2001, 2002, 2003). The present results for spontaneous speech therefore extend and validate previous studies using read speech.

In addition, these results increase our understanding of pronunciation variation in assimilable environments by demonstrating comparable degrees of acoustic modification of place information for perceptually distinctive phonetic realizations, i.e., assimilations versus deletions. In particular, measurements of F2 difference were indistinguishable for deleted and assimilated categories; these two in turn were statistically indistinguishable with respect to three of four control conditions (i.e., true labial or velar contexts). Only in the labial context in Fig. 3 is the control distribution noticeably shifted upward away from the assimilated and deleted distributions.

The data in the four deleted conditions are particularly surprising because labelers listen closely to the speech for evidence of segments, and they adopt a conservative criterion when classifying a segment as deleted (Kiesling *et al.*, 2006). Although the extent of F2 transitions were comparable for both deletions and assimilations, they were apparently not perceptible in the former case. Why? Analyses reported in the next sections were carried out to identify acoustic evidence that labelers might have used in distinguishing deletions from assimilations.

C. Amplitude differences in formant transitions

One possible reason why labelers reported a word-final segment to be deleted rather than assimilated could have been that the amplitude of the F2 transitions decreased more dramatically in the former case than the latter, making F2 cues less perceptible. We evaluated this idea by measuring amplitude in the vicinity of the underlying segment for a subset of tokens labeled as deleted and assimilated in assimilable environments; measurements from alveolar segments labeled as canonical in assimilable environments were also included for comparison. To ensure comparable samples across these categories, tokens were matched according to several dimensions, including the identity, gender, and/or age of the talker, the type of preceding vowel context, and the place of articulation of the following context (as labial or velar).[5] For [ɪ] contexts, 78 assimilated, 78 deleted, and 78 canonical tokens were examined, while for [æ] and [ɛ] contexts, 45 assimilated, 45 deleted, and 45 canonical tokens were examined.

The amplitude of the first formant, A1, and the amplitude of the second formant, A2, were measured at two points in the vowel, since we hypothesized that a reduction in one or both might conceivably alter perception enough to cause a change in labeling. A 15 ms Hanning window was used in Xwaves to generate a FFT spectrum used in amplitude measurements. The first measurement was taken near the middle of the vowel by centering the analysis window on the positive-going portion of the pitch period closest to the vowel midpoint; this gave rise to estimates of A1 and A2 in decibels. The second measurement was taken near the end of

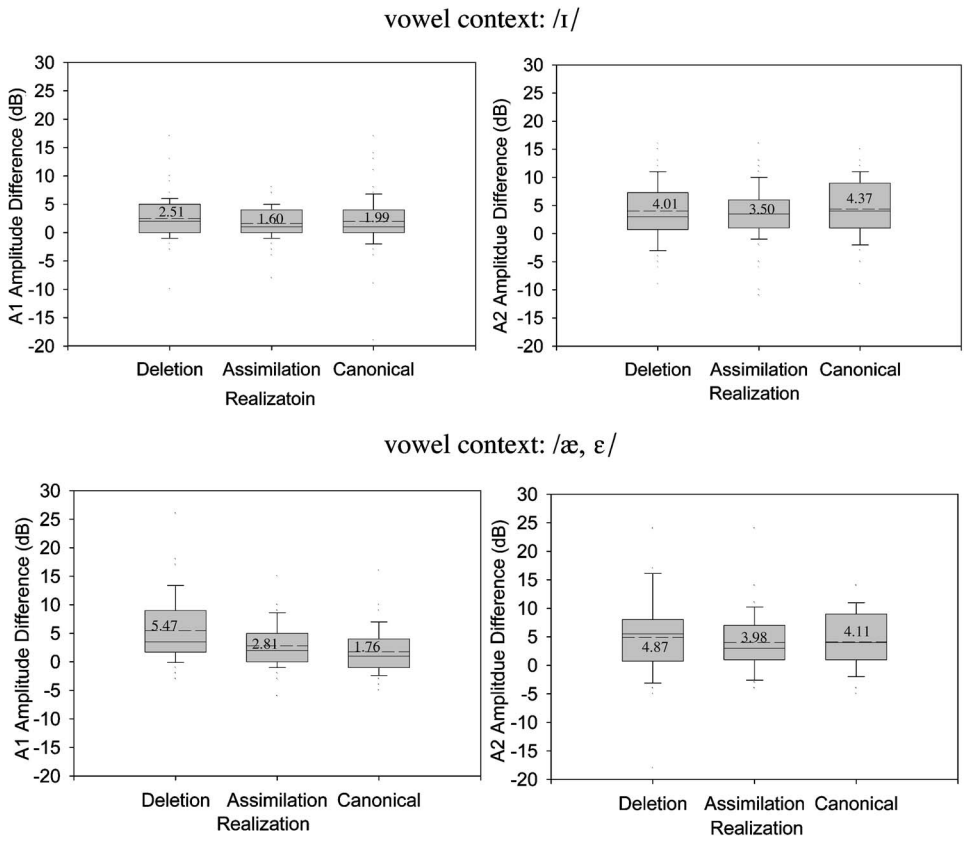

FIG. 5. Box plots of the drop in formant amplitude in tokens labeled as deleted, assimilated, and canonical. The top and bottom graphs differ with respect to the identity of the preceding vowel ([ɪ] or [æ,ɛ]). Data in the left graph are for the first formant, those in the right for the second. The solid line in the middle box is the median of the distribution, and the dashed line is the mean, with its value listed.

the vowel by centering the analysis window on the positive-going portion of the pitch period closest to but not less than 7.5 ms (half the size of the analysis window) from the end of the vowel segment. Two relative amplitude metrics, an A1 difference and an A2 difference, were then calculated by subtracting the values taken at the vowel end point from those taken at the vowel midpoint. No amplitude measurements were taken if the vowel was less than 30 ms in duration. This resulted in discarding eight tokens each from deleted and assimilated categories and seven from the canonical category in the [ɪ] context, as well as seven tokens from the deleted category and two from the assimilated category in the [æ] and [ɛ] contexts.

The results of the amplitude analysis are shown in Fig. 5. A three-way ANOVA was performed on the amplitude measurements, with vowel context and realization as between-item factors and formant amplitude as a within-item factor. The main effect of realization was marginally reliable, $F(2,332)=2.967$, $p<0.053$. None of the post hoc comparisons between pairs of conditions reached significance. Nevertheless, across all four graphs in Fig. 5 there is a trend in the predicted direction, with the drop in energy being greater for segments labeled as deleted than assimilated.

However, additional properties of the graphs do not instill confidence in the ability of listeners to discriminate reliably between these two realizations using amplitude drop-off. The bulk of the deleted and assimilated distributions in three of the four graphs (all but the lower left) overlap. Furthermore, these means differ by no more than 1 dB from each other. In fact, one-way ANOVAs on these triplets of distributions produced no reliable effect of token label. Only in the lower left graph was the ANOVA reliable, $F(2,123)=6.99$, $p<0.001$. Post hoc Tukey's tests confirmed that the 2.6 dB drop from the deleted to the assimilated condition is reliable. Although amplitude drop-off is greater for deletions than assimilations, its small magnitude combined with wide variability make this acoustic property, just like the F2 difference, a minimally informative cue to use in discriminating deletions from assimilations.

Other reliable effects in the omnibus ANOVA included a main effect of formant amplitude, $F(1,332)=26.766$, $p<0.001$, with drops generally being greatest for A1. This variable also interacted with realization, $F(2,332)=3.87$, $p<0.022$, with the A1 vs A2 differences being much larger for canonical than either deleted or assimilated realizations.

D. Closure duration

What other acoustic characteristic could distinguish tokens labeled as deleted from those labeled as being present (i.e., assimilated or canonical)? For tokens which were labeled as assimilated, deleted, and canonical, the acoustic realization of the consonant(s) at the word boundary, namely $(C_1)\#C_2$ in underlying $VC_1\#C_2(V)$ contexts, consistently corresponded to a single low- or zero-amplitude consonantal occlusion, with no release burst for C_1 if present. In other words, to the extent that the initial stop, C_1, might be present, its most salient acoustic hallmark was as an unreleased stop closure, followed immediately by the consonantal closure for C_2, giving rise in most cases to an otherwise undifferentiated silent interval. Previous research has shown that the perceptual salience of stop consonants depends in part on the dura-

tion of the closure, with listeners tending to perceive a stop C_1 as being absent in $VC_1\#C_2V$ contexts when the duration of the stop closure for $C_1\#C_2$ is short (Repp, 1978; Fujimura *et al.*, 1978; Schouten and Pols, 1983; Ohala, 1990). We reasoned, therefore, that a difference in the entire duration of the (silent) consonantal closure might distinguish instances labeled as deleted from those labeled as assimilated or canonical; if a consonant C_1 in $C_1\#C_2$ were produced with a very short overall closure duration, this might tend to make underlying C_1 segments be perceived as absent, and thus labelers would tend to code it as deleted. Conversely, we reasoned that a segment which was produced more carefully would be perceived as being present, and thus labeled as assimilated or canonical and exhibit a longer closure duration.

Using the same subsets of tokens for which amplitude measurements were taken, we compared closure durations associated with consonantal constrictions at the word boundary across instances labeled as deleted, assimilated, or canonical.[6] Xwaves displays of a waveform and wide-band spectrogram were used to identify the start and end of the consonantal constriction in the vicinity of the word boundary for each token. The starting and ending points of the closure were taken as the positions at which the amplitude suddenly dropped off or increased, respectively, across frequencies. For tokens classified as assimilated or canonical, the consonantal constriction included the duration of the word-final assimilated or canonical consonant, plus the closure period associated with the following word-initial consonant before any burst release. For tokens classified as deleted, the consonantal constriction included only the duration of the closure period associated with the following word-initial consonant prior to any burst release, since phonetic labels indicated zero duration for the word-final consonant.

Box plots of the closure durations for the deleted, assimilated, and canonical tokens are graphed in Fig. 6. The data are impressively consistent across the two vowel contexts. Relative to the deleted distribution, the assimilated distribution is shifted upward by an average of 20 ms into regions of longer durations. The canonical distribution is shifted into even longer regions by the same amount. In the [ɪ] context, a one-way ANOVA across realization conditions was reliable, $F(2,231)=18.62$, $p<0.001$. Statistical analyses in the [æ,ɛ] contexts yielded the same outcome, $F(2,219)=19.99$, $p<.001$.[7] Post hoc Tukey's HSD tests in both vowel contexts showed that all conditions were significantly different from one another. When the data are combined into a single two-way analysis with vowel as the additional factor, only an effect of realization emerged, $F(2,360)=33.19$, $p<0.001$.

Rather than being differentiated by amplitude or F2 transitions, assimilations and deletions were distinguished by the duration of consonantal closures. These results suggest that labelers' judgments about the presence of the word-final stop were likely to have been influenced by closure duration, a finding which mirrors results from experiments showing that perception of consonant presence and identity is mediated by closure duration (Repp, 1978; Fujimura *et al.*, 1978; Schouten and Pols, 1983; Ohala, 1990; Esposito and Di Benetto, 1999). What might be the cause of this significant change in percept? Because gap duration is shorter in the case of segments labeled as deleted, energetic masking of preconstriction material by the postconstriction segment could be greater. However, this explanation seems unlikely given the length of the gap (60 ms; Studdert-Kennedy *et al.*, 1970). A more intriguing possibility is that listeners' sensitivity to the timing in gesture coordination across a word boundary might have caused them to perceive a word-final stop for longer gap intervals (cf. Browman and Goldstein, 1990). Alternatively, a minimum intrinsic duration might be necessary in order for listeners to perceive certain phonemes (Klatt, 1979). In sum, these results indicate that whether an assimilated or deleted segment is perceived is likely to be influenced by closure duration.

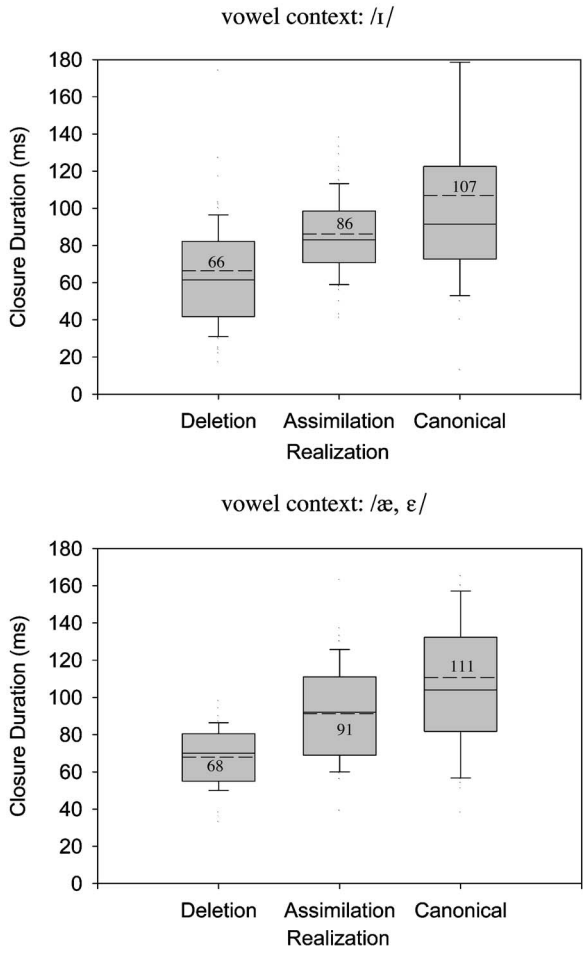

FIG. 6. Box plots of the closure duration in tokens labeled as deleted, assimilated, and canonical (alveolar). The top and bottom graphs differ with respect to the identity of the preceding vowel ([ɪ] or [æ,ɛ]).

IV. GENERAL DISCUSSION

Two studies investigated the nature of regressive place assimilation in a corpus of spontaneous speech. In one study, the frequency of assimilation was examined relative to other kinds of variation, including glottalization and deletion, using phonetic labels from the Buckeye Corpus of spontaneous speech. In a second study, an acoustic investigation was conducted to assess the extent of place modification for tokens

labeled as assimilated versus nonassimilated, as well as the ways in which assimilation differed from other types of variation.

These studies yielded several key findings. First, assimilation is a relatively rare form of pronunciation variation in assimilable environments, as estimated from labeling data produced by trained speech analysts. In particular, assimilation occurred in only 9% of all possible assimilable environments in our data. Tokens labeled as assimilated are likely to correspond to cases which have previously been labeled "extreme assimilations," "complete" or "near-complete assimilations," or "zero alveolars" (e.g., Wright and Kerswill, 1989; Nolan, 1992; Gaskell, 2003). Earlier estimates of rates of these extreme assimilations have varied. For example, Nolan (1992) estimated the prevalence of extreme assimilations to be between 20% and 90% depending on the speed and style of speech. Moreover, Gaskell and Marslen-Wilson (2001) proposed that approximately 50% of assimilations would be "fully assimilated" in speech corpora. In contrast, the present spontaneous speech corpus analysis suggests that the rate of complete or near-complete assimilation in environments where it might occur is low, indicating that the forms which are likely to be most problematic for speech perception are fairly infrequent. In contrast, other types of variation, namely deletion and glottal variants, were much more common in these environments, and canonical realizations also occurred quite often. Lexical frequency was shown to be a factor in the range of variation that a word exhibited: High-frequency lexical items were more likely to appear with multiple variant types than lower-frequency words. Such findings mirror previous results showing that higher frequency lexical items are more likely to exhibit missing segments and other types of modifications (e.g., Bell *et al.*, 2003; Pluymaekers *et al.*, 2005; Ernestus *et al.*, 2006).

Our second finding was that assimilation is a graded phenomenon in spontaneous speech. This can be seen by the fact that alveolars labeled as assimilated showed a range of places of articulation, as gauged by the distribution in F2 values. Moreover, even alveolars labeled as canonical but followed by labial or velar consonants showed a shift in F2 values relative to alveolars that were followed by another alveolar consonant (e.g., bars 4 and 5 in Fig. 3), indicating partial assimilation of the former. These results confirm and extend previous findings using read speech which have shown gradation in acoustic or articulatory markers of degree of assimilation (Gow, 2001, 2002, 2003; Zsiga, 1995; Holst and Nolan, 1995; Nolan, 1992; Wright and Kerswill, 1989). Our results demonstrating gradation in place assimilation are also mirrored by recent results for another conversational speech phenomenon, namely voicing assimilation, which is similarly graded in French as well as Dutch conversational speech (Snoeren *et al.*, 2006; Ernestus *et al.*, 2006).

The finding that assimilation is realized through graded acoustic cues has implications for interpreting the labeling data to estimate rates of assimilation in spontaneous speech and for using such data to study word recognition. Although labelers made categorical labeling judgments about the status of tokens as assimilated versus nonassimilated (e.g., bars 3 and 4 in Fig. 3), the acoustic data showed that tokens in assimilable contexts which were labeled as not assimilated (e.g., *pin point*, bar 4) nevertheless exhibited F2 values which were shifted away from nonassimilable, alveolar contexts (e.g., *did not*, bar 5). This observation suggests that estimates of assimilation rates will depend on labelers' thresholds for hearing tokens as assimilated or not (i.e., as canonical, deleted, or glottal forms). Given that labelers adopted a conservative criterion for the assimilated versus nonassimilated distinction (Kiesling *et al.*, 2006), estimates of assimilation rate based on labeling data provide a reasonable benchmark of near-complete and complete assimilation frequency in spontaneous speech. Finally, these results underscore the necessity of conducting acoustic-phonetic studies to obtain a full picture of connected speech processes, rather than relying on phonological data alone.

A third finding of the present study was that acoustic cues to place of articulation for alveolars labeled as assimilated had virtually identical distributions to those of underlying labial or dorsal sounds. This suggests that assimilation is often complete or nearly complete in spontaneous speech, consistent with findings from read speech (Holst and Nolan, 1995; Nolan *et al.*, 1996; Ellis and Hardcastle, 2002). Together with the finding that assimilation is graded in spontaneous speech, these data suggest that details of assimilation previously examined in read speech material generalize to casual, unscripted speech materials (Wright and Kerswill, 1989; Holst and Nolan, 1995; Zsiga, 1995; Ellis and Hardcastle, 2002).

Our fourth finding concerns acoustic factors which differentiate assimilated, canonical, and deleted variants. Perhaps surprisingly, similar degrees of F2 change were exhibited for assimilated and deleted variants, suggesting that both forms were taking the place of articulation of a following labial or dorsal sound. Moreover, similar relative amplitude levels were found for assimilated and deleted variants, as well as canonical segments (Fig. 5), underscoring the ambiguity of acoustic information in word-final position with respect to the surface realization of a segment. However, the duration of the consonant closure was found to distinguish variant types: both canonical and assimilated variants had longer closure durations than deleted variants. Previous work has shown that listeners are perceptually sensitive to closure duration in judgments related to consonantal context (Esposito and Di Benetto, 1999; Repp, 1978; Fujimura *et al.*, 1978; Schouten and Pols, 1983; Ohala, 1990), lending support that this acoustic cue was likely to be important in distinguishing assimilations from deletions. In sum, these results suggest that the word-final segment is perceptually quite fragile.

The current findings should be qualified in a few respects. First, the present study was limited to talkers from in and around Columbus, OH. While we cannot speculate in depth on how assimilation rate might be different for other dialects, the variety of English spoken in central Ohio is very similar to the General American dialect (Labov *et al.*, 2005). Thus these data provide a reasonable index of assimilation behavior for a large number of speakers in North America. Moreover, with respect to the acoustical study, our investiga-

tion of place assimilation cues was limited to measurements of F2. While information about place of articulation is most readily conveyed both perceptually and acoustically by F2 information in C-to-V and V-to-C transitions (e.g., Liberman *et al.*, 1954; Tartter *et al.*, 1983), F3 transitions, burst spectra, and other cues are also relevant to place perception (e.g., Harris *et al.*, 1958). The present study also did not take prosodic boundaries into account, although likely instances of the largest prosodic boundaries, e.g., full intonational phrase (IP) and utterance boundaries (cf. Beckman and Pierrehumbert, 1986) were excluded from analysis. Prosodic structure has been argued to mediate optional reduction phenomena, such that prosodic phrase-initial positions are realized with markers of articulatory strength, such as glottalization, increased linguopalatal contact, and/or shorter VOT (e.g., Dilley *et al.*, 1996; Fougeron and Keating, 1997). Such cues could be present in an assimilable context, and thus aid processing. However, recent work shows that word-initial alveolar /t/ and /d/ do not show enhanced cues to voicing contrasts in full IP-initial position in American English read speech (Cole *et al.*, 2007), and that prosodic junctures smaller than the full IP may not be important for processing cues to assimilation, at least in Korean (Cho and McQueen, in press). Thus, prosodic structure seems unlikely to have had a significant effect on our results.

What are the implications of these data for theories of spoken word recognition, and in particular, for accounts developed specifically for recognizing assimilated variants (Gaskell and Marslen-Wilson, 1998; Lahiri and Marslen-Wilson, 1991; Gow, 2003)? It seems that they change the nature of the problem that must be solved, in several respects. First, spoken word recognition theories must allow for the fact that there are graded degrees of assimilation in contexts where assimilation is possible, ranging from zero assimilation to complete (i.e., extreme) assimilation. This gradation has yet to be dealt with satisfactorily in most spoken word recognition accounts. However, recent models that have begun to incorporate findings of gradedness in assimilation cues into accounts of how listeners perceive the intended word, such as Gaskell (2003) and Gow (2003), seem the most promising. To facilitate empirical work, future studies might benefit from the use a graded labeling scale (e.g., 1–7) to more explicitly indicate varying degrees of assimilation, as demonstrated in our data.

Second, the present data show that other types of variation in addition to assimilation can occur in assimilable contexts. This suggests that reliance of these models on acoustic cues to resolve the place of articulation of the final segment will need to be revised or augmented. Recognizing a variant pronunciation apparently requires taking into account multiple kinds of acoustic information spread out over time, since place of articulation cues alone are insufficient to distinguish assimilated, deleted, canonical, and glottal variants.

How might spoken word recognition theories account for these other types of variation in assimilable environments, in particular, for deleted and glottal variants? Two possibilities come to mind. Under the view closest to existing accounts of processing of assimilated variants (e.g., Gaskell and Marslen-Wilson, 1998), separate mechanisms may be involved in recognizing variants which are realized through distinct acoustic dimensions. According to this view, different processes are entailed in recovering segments after modifications to place of articulation, duration, voice quality, and so forth. This is the view implied by traditional linguistic analyses (e.g., Kager, 1999), according to which categorically distinctive phonological processes apply to canonical forms to yield variant types (glottal, deleted, or assimilated) with rather divergent acoustic structures; recovering the intended segment then involves processes of "undoing" individual rule applications to yield the underlying phoneme.

Alternatively, a somewhat different view that was suggested by a reviewer is that assimilated and deleted variants might be dealt with through the same mechanism. This alternative proposal arises from the observation that both assimilated and deleted variants in this study showed comparable degrees of F2 modification, indicating that both could potentially be handled in the same way by the recognition system. Under this proposal, the assimilation rate would be higher, and only glottal variants would require special treatment. However, it does not seem to us that assimilated and deleted variants could be processed in exactly the same way, since the two were distinguished by a timing variable, namely, the duration of a consonantal closure. This suggests that speech timing must be taken into account in spoken word recognition, consistent with recent findings showing that temporary lexical activation of embedded words (e.g., *ham* in *hamster*) depends partly on temporal cues (e.g., Salverda *et al.*, 2003). Moreover, the importance of timing information in speech perception more generally is well established (e.g., Repp *et al.*, 1978; Tartter *et al.*, 1983).

Because current theoretical accounts were not designed to explain the manifold variation in assimilable environments, it might be more appropriate to ask how well-positioned they are to incorporate these new findings. The importance of closure duration leads us to believe that Gow's (2003) account could fare particularly well. This is because the feature parsing account proposed by Gow specifically invokes perceptual principles of auditory scene analysis, an area which focuses on the problem of assigning sounds to their environmental sources. Sounds which are temporally closer tend to preferentially group together into auditory "streams" (e.g., Bregman, 1990), thus potentially providing the rudiments of an account of perception of a deleted segment. Nevertheless, the prevalence and distinct acoustic manifestation of glottal variants pose further challenges for this and other processing accounts, particularly since glottalization may mark vowel onsets and/or prosodic phrase boundaries, in addition to realizing [ʔ] variants of /t/ and /d/ (Dilley *et al.*, 1996; Redi and Shattuck-Hufnagel, 2001).

Given our findings that the final segment is often acoustically unclear, it seems likely that in such cases the surrounding context is the most frequent source of information with which to ensure correct recognition, further increasing the complexity of a theoretical account. One possibility in particular is that the preceding lexical context could aid in interpreting the ambiguous acoustic information. Such a lexically driven restoration process could be highly successful because the segment occurs at the end of the word, where

lexical influences are greatest (Pitt and Samuel, 1993, 1995). However, the success of a lexically guided account depends on the context itself being unique enough to specify how the ambiguous word-final cues should be interpreted. Content words that become lexically unique before word offset would be maximally informative in this regard. Examination of the assimilated and deleted tokens from the acoustic analyses showed that such ideal conditions for restoration occur infrequently. Although word length ranged from 2 to 11 phonemes, 85% were two and three phonemes long, with "it," "in," and "that" being the most frequent. Only 36% of the tokens are lexically unique at or before word offset, but over half of these instances are due to a single word, "that." Removal of this word reduces the number to 19%. These statistics suggest that lexical information will be of only moderate help in recovering the word-final segment. Most likely the larger sentential context will also be needed to determine the word's identity, and thus the identity of the segment (Gaskell, 2001). Of course, when the acoustic cues unambiguously specify an assimilated segment, lexical and sentential information might not suffice.

V. CONCLUSIONS

In conclusion, the current research shows that regressive place assimilation is only one of several types of variation which occur in assimilable environments, with deletion and glottal variants being others. Assimilation in spontaneous speech was graded, ranging from full to partial to none. Although acoustic cues to place of articulation were consistently present in the signal, they did not determine the type of variant that was heard. Rather, it is necessary to take other kinds of acoustic information into account, such as closure duration, to explain word perception in assimilable contexts.

ACKNOWLEDGMENTS

The authors would like to thank Cynthia Connine and an anonymous reviewer for very helpful feedback which has greatly strengthened the paper. Additionally, we thank Anne Pier Salverda, Holger Mitterer, Mirjam Ernestus, David Gow, and Gareth Gaskell for providing feedback on earlier drafts of the paper and/or help with queries. Moreover, we thank Mallory Crapo for help with data analysis. Finally, we thank Keith Johnson for many contributions to an early version of this work. This work was supported by research Grant No. DC004330 from the National Institute on Deafness and Other Communication Disorders. Portions of this project were presented at the 46th Annual Meeting of the Psychonomic Society, Toronto, Canada.

[1]Consistent with standard linguistic conventions, slashes are used throughout the paper to indicate underlying segment types (e.g., /t/), while square brackets are used to indicate surface phonetic realizations (e.g., a flapped realization, [ɾ]).

[2]Phonetic analysts also selected among labels such as SIL (for silence) or HES (for hesitation) to indicate locations of nonfluent or fluent pauses between portions of running speech which could not be attributed to a stop closure. See Kiesling *et al.* (2006) for more details.

[3]These labeling consistency data compare favorably with other studies (e.g., Irwin, 1970; Eisen, 1991). For example, Eisen (1991) found labeling accuracy of 88% for obstruent consonants.

[4]Glottal variants were not subjected to further acoustic analysis, since their phonetic realizations are quite different from other variant realizations (Redi and Shattuck-Hufnagel, 2001), in a way which did not lend them to comparison along the dimensions of interest. Moreover, critical acoustic analysis for these tokens already took place during the phonetic labeling process, where each token labeled as glottalized was identified as having irregularly timed pitch pulses in the waveform.

[5]For both vowel contexts, 100% of tokens were matched according to the proportion of following labial and velar contexts. Moreover, 94% of tokens preceded by [ɪ] vowels and 91% of tokens preceded by [æ,ɛ] vowels were matched by the gender and age of a talker, while 79% and 66% of these were matched according to the exact identity of a talker, respectively.

[6]It is important to note that the criteria of labelers about the presence and variant status of a particular underlying word-final alveolar token did not expressly take closure duration into account. (See Sec. II A.) Thus, there is no *a priori* reason why one might expect a longer closure duration for assimilated and canonical variants than deleted variants. In fact, the majority of tokens measured here involved underlying stop consonants followed by another stop; the acoustic realization of the closure in such cases was a single, silent interval spanning both segments. Labeling conventions for the Buckeye Corpus (Kiesling *et al.*, 2006) dictated that in such contexts when an underlying stop was judged to be realized as assimilated or canonical, the midpoint of the closure period was taken as the end of the word-final stop, but when an underlying stop was judged to be deleted, the entire closure period was attributed to the following, word-initial stop.

[7]When closure duration is normalized by dividing by the duration of the preceding vowel, which has been proposed as one means of equating for speech rate (e.g., De Jong, 1998), we found that the pattern of results was quite similar.

Barry, M. (**1992**). "Palatalisation, assimilation, and gestural weakening in connected speech," Speech Commun. **11**, 393–400.

Beckman, M. E., and Pierrehumbert, J. B. (**1986**). "Intonational structure in Japanese and English," Phonology Yearbook **3**, 255–309.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (**2003**). "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation," J. Acoust. Soc. Am. **113**, 1001–1024.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Browman, C., and Goldstein, L. (**1990**). "Tiers in articulatory phonology, with some implications for casual speech," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. E. Beckman (Cambridge University Press, Cambridge), pp. 341–376.

Cho, T., and McQueen, J. "Not all assimilated sounds are perceived equally: Evidence from Korean," J. Phonetics, in press.

CMU pronouncing dictionary (v. 0.6) [www.speech.cs.cmu.edu/cgi-bin/cmudict], Pittsburgh, PA, Carnegie Mellon University (distributor). Date last viewed: 11 June, 2007.

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

Connine, C. (**2004**). "It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition," Psychonomic Bulletin and Review **11**, 1084–1089.

Crystal, T. H., and House, A. S. (**1988**). "Segment durations in connected-speech signals: Syllabic stress," J. Acoust. Soc. Am. **83**, 1574–1585.

Dalby, J. M. (**1986**). "Phonetic structure of fast speech in American English," Ph.D. dissertation, Indiana University, Bloomington, IN.

De Jong, K. (**1998**). "Stress-related variation in the articulation of coda alveolar stops: Flapping revisited," J. Phonetics **26**, 283–310.

Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M. (**1996**). "Glottalization of vowel-initial syllables as a function of prosodic structure," J. Phonetics **24**, 423–444.

Eisen, B. (**1991**). "Reliability of speech segmentation and labeling at different levels of transcription," in *Proceedings of Eurospeech-91*, Berlin, pp. 673–676.

Ellis, L., and Hardcastle, W. J. (**2002**). "Categorical and gradient properties of assimilation in alveolar to velar sequences: Evidence from EPG and EMA data," J. Phonetics **30**, 373–396.

Ernestus, M., Lahey, M., Verhees, F., and Baayen, R. H. (**2006**). "Lexical frequency and voice assimilation," J. Acoust. Soc. Am. **120**, 1040–1051.

Esposito, A., and Di Benetto, M. G. (**1999**). "Acoustical and perceptual study of gemination in Italian stops," J. Acoust. Soc. Am. **106**, 2051–2062.

Fougeron, C., and Keating, P.A. (**1997**). "Articulatory strengthening at edges of prosodic domains," J. Acoust. Soc. Am. **101**, 3728–3740.
Fujimura, O., Macchi, M. J., and Streeter, L. A. (**1978**). "Perception of stop consonants with conflicting transitional cues: A cross-linguistic study," Lang Speech **21**, 337–346.
Gaskell, M. G. (**2001**). "Phonological variation and its consequences for the word recognition system," Lang. Cognit. Processes **16**, 723–729.
Gaskell, M. G. (**2003**). "Modelling regressive and progressive effects of assimilation in speech perception," J. Phonetics **31**, 447–463.
Gaskell, M. G., and Marslen-Wilson, W. D. (**1998**). "Mechanisms of phonological inference in speech perception," J. Exp. Psychol. Hum. Percept. Perform. **24**, 380–396.
Gaskell, M. G., and Marslen-Wilson, W. D. (**2001**). "Lexical ambiguity and spoken word recognition: Bridging the gap," J. Mem. Lang. **44**, 325–349.
Gow, D. W. (**2001**). "Assimilation and anticipation in continuous spoken word recognition," J. Mem. Lang. **45**, 133–159.
Gow, D. W. (**2002**). "Does English coronal place assimilation create lexical ambiguity?" J. Exp. Psychol. Hum. Percept. Perform. **28**, 163–179.
Gow, D. W. (**2003**). "Feature parsing: Feature cue mapping in spoken word recognition," Percept. Psychophys.. **65**, 575–590.
Gow, D. W., and McMurray, B. "Word recognition and phonology: The case of English coronal place assimilation," Papers in Laboratory Phonology 9, in press.
Guy, G. R. (**1980**). "Variation in the group and the individual: The case of final stop deletion," in *Locating Language in Time and Space*, edited by W. Labov (Academic, New York).
Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., and Cooper, F. S. (**1958**). "Effect of third-formant transitions on the perception of the voiced stop consonants," J. Acoust. Soc. Am. **30**, 122–126.
Holst, T., and Nolan, F. (**1995**). "The influence of syntactic structure on [s] to [ʃ] assimilation," in *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvaniti (Cambridge University Press, Cambridge), pp. 315–333.
Irwin, R. B. (**1970**). "Consistency of judgments of articulatory productions," J. Speech Hear. Res. **13**, 548–555.
Johnson, K. (**2004**). "Massive reduction in conversational American English," in *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, edited by K. Yoneyama and K. Maekawa (The National International Institute for Japanese Language, Tokyo), pp. 29–54.
Kager, R., (**1999**). *Optimality Theory*. (Cambridge University Press, Cambridge).
Kiesling, S., Dilley, L., and Raymond, W. (**2006**). "The Variation in Conversation (ViC) Project: Creation of the Buckeye Corpus of conversational speech," Department of Psychology, Ohio State University, Columbus, OH, available at www.buckeyecorpus.osu.edu. Last viewed 8/20/07.
Klatt, D. H. (**1979**). "Synthesis by rule of segmental durations in English sentences," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhmann (Academic, New York), pp. 287–299.
Kohler, K. (**1990**). "Segmental reduction in connected speech in German: Phonological facts and phonetic explanations," in *Proceedings of the NATO Advanced Study Institute on Speech Production and Speech Modelling*, Bonas, France (Kluwer Academic, Dordrecht).
Labov, W., Ash, S., and Boberg, C. (**2005**). *The Atlas of North American English: Phonetics, Phonology and Sound Change* (Mouton de Gruyter, Berlin).
Lahiri, A., and Marslen-Wilson, W. (**1991**). "The mental representation of lexical form: A phonological approach to the recognition lexicon," Cognition **38**, 245–294.
Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (**1954**). "The role of consonant-vowel transitions in the perception of stop and nasal consonants," Psychol. Monogr. **68**, 1–13.
Miller, J. L. (**2001**). "Mapping from acoustic signal to phonetic category: Internal category structure, context effects and speeded categorization," Lang. Cognit. Processes **16**, 683–690.
Mitterer, H., and Ernestus, M. (**2006**). "Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch," J. Phonetics **34**, 73–103.
Neu, H. (**1980**). "Ranking of constraints on /t,d/ deletion in American English: A statistical analysis," in *Locating Language in Time and Space*, edited by W. Labov (Academic, New York), pp. 37–54.
Nolan, F. (**1992**). "The descriptive role of segments: Evidence for assimilation," in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, edited by D. R. Ladd and G. J. Docherty (Cambridge University Press, Cambridge), pp. 261–280.
Nolan, F., Holst, T., and Kühnert, B. (**1996**). "Modeling [s] to [ʃ] accommodation in English," J. Phonetics **24**, 113–137.
Ohala, J. J. (**1990**). "The phonetics and phonology of aspects of assimilation," in *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, edited by J. Kingston and M. Beckman (Cambridge University Press, Cambridge), pp. 258–275.
Patterson, D. and Connine, C. M. (**2001**). "A corpus analysis of variant frequency in American English flap production," Phonetica **58**, 254–275.
Patterson, D., LoCasto, P. C., and Connine, C. M. (**2003**). "A corpus analysis of schwa vowel deletion frequency in American English," Phonetica **60**, 45–68.
Pitt, M., and Johnson, K. (**2003**). "Using pronunciation data as a starting point data in modeling word recognition," Paper presented at the 15th International Congress of Phonetic Sciences, Barcelona.
Pitt, M., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (**2005**). "The Buckeye Corpus of conversational speech: Labeling conventions and a test of transcriber reliability," Speech Commun. **45**, 89–95.
Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (**2006**). "Buckeye Corpus of conversational speech" (1st release) [www.buckeyecorpus.osu.edu], Department of Psychology, Ohio State University (distributor), Columbus, OH. Last viewed 8/20/07.
Pitt, M. A., and Samuel, A. G. (**1993**). "An empirical and meta-analytic evaluation of the phoneme identification task," J. Exp. Psychol. Hum. Percept. Perform. **19**, 699–725.
Pitt, M. A., and Samuel, A. G. (**1995**). "Lexical and sublexical feedback in auditory word recognition," Cognit Psychol. **29**, 149–188.
Pluymaekers, M., Ernestus, M., and Baayen, R. H. (**2005**). "Lexical frequency and acoustic reduction in spoken Dutch," J. Acoust. Soc. Am. **18**, 2561–2569.
Raymond, W., Dautricourt, R., and Hume, E. (**2006**). "Word-medial /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors," Lang. Var. Chg. **18**, 55–97.
Redi, L., and Shattuck-Hufnagel, S. (**2001**). "Variation in realization of glottalization in normal speakers," J. Phonetics **29**, 407–429.
Repp, B. (**1978**). "Perceptual integration and differentiation of spectral cues for intervocalic stop consonants," Percept. Psychophys. **24**, 471–485.
Repp, B. H., Libernan, A. M., Eccardt, T., and Pesetsky, D. (**1978**). "Perceptual integration of acoustic cues for stop, fricative, and affricate manner," J. Exp. Psychol. Hum. Percept. Perform. **4**, 621–637.
Salverda, A. P., Dahan, D., and McQueen, J. M. (**2003**). "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," Cognition **90**, 51–89.
Schouten, M. E. H. and Pols, L. C. W. (**1983**). "Perception of plosive consonants. The relative contributions of bursts and vocalic transitions," in *Sound Structures: Studies for Antonie Cohen*, edited by M. P. R. van den Broecke, V. J. van Heuven, and W. Zonneveld (Foris, Dordrecht), pp. 227–243.
Shockey, L. (**2003**). *Sound Patterns of Spoken English* (Blackwell, Malden, MA).
Snoeren, N. D., Hallé, P. A., and Segui, J. (**2006**). "A voice for the voiceless: Production and perception of assimilated stops in French," J. Phonetics **34**, 241–268.
Stevens, K. (**1998**). *Acoustic Phonetics* (MIT, Cambridge, MA).
Studdert-Kennedy, M., Shankweiler, D. P., and Schulman, S. (**1970**). "Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables," J. Acoust. Soc. Am. **48**, 588–602.
Tartter, V. C., Kat, D., Samuel, A. G., and Repp, B. H. (**1983**). "Perception of intervocalic stop consonants: The contributions of closure duration and formant transitions," J. Acoust. Soc. Am. **74**, 715–25.
Wright, S., and Kerswill, P. (**1989**). "Electropalatography in the study of connected speech processes," Clin. Linguist. Phonetics **3**, 49–57.
Zsiga, E. (**1995**). "An acoustic and electropalatographic study of lexical and postlexical palatalization in American English," in *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvaniti (Cambridge University Press, Cambridge), pp. 282–283.