Available online at www.sciencedirect.com

ScienceDirect

Lingua 152 (2014) 24–44

Lingua

www.elsevier.com/locate/lingua

Imperceptible incomplete neutralization: Production, non-identifiability, and non-discriminability in American English flapping

Aaron Braver [*]

Texas Tech University, Department of English, P.O. Box 43091, Lubbock, TX 79409-3091, United States

Received 1 November 2012; received in revised form 6 September 2014; accepted 9 September 2014
Available online

Abstract

Flapping in American English has been put forward as a case of incomplete neutralization—in other words, /d/-flaps and /t/-flaps differ at the phonetic level. This paper first presents a production experiment which shows that, in line with previous work, flapping in American English is incompletely neutralizing: vowels before /d/-flaps are slightly longer than those before /t/-flaps—even in nonce words.

Early studies on the perceptibility of this difference, almost exclusively identification tasks, have shown mixed results. However, recent identification experiments (including one reported here) show that listeners are unable to properly categorize /d/- and /t/-flaps. Listeners' poor performance on identification tasks can be due to two factors: either (a) listeners' grammars lacking the relevant phonological categories, or (b) an effect of the type of perception tasks employed. In a 2AFC discrimination task presented here, listeners were unable to distinguish between /d/- and /t/-flaps, suggesting that poor perception performance generalizes to multiple task types.

Keywords: Incomplete neutralization; Flapping; Phonetics; American English

1. Introduction

In American English flapping, underlying /d/ and /t/ become [ɾ] in certain prosodic configurations (e.g., Kahn, 1980). A number of studies have examined the possibility that flapping is a case of incomplete neutralization—that is, that the underlying voicing status of a flap might be discernible on the surface, most notably by means of the duration of the preceding vowel (see, e.g., Joos, 1942; Fisher and Hirsh, 1976; Port, 1976; Fox and Terbeek, 1977; Zue and Laferriere, 1979; Huff, 1980; Patterson and Connine, 2001:264–267; Herd et al., 2010). The experiments presented in this paper test two main claims: (i) flapping in American English is incompletely neutralizing, and (ii) listeners are unable to label and categorize /d/-flaps and /t/-flaps based on their surface distinction. The production experiment described in section 2 provides additional evidence that speakers of (Mid-Atlantic) American English maintain a trace of the underlying laryngeal contrast in /d/-flaps and /t/-flaps. Further, on the basis of both identification and discrimination tasks (sections 3 and 4), I argue that listeners are unable to perceive the surface distinction between /d/-flaps and /t/-flaps.

Flapping has traditionally been described as a phonological rule which maps intervocalic /t/ and /d/ to [ɾ]; however, the precise mechanism and environment of this mapping varies from one description to the next. Turk (1992) divides these various rule-based accounts into three groups based on the prosodic environment of flapping: (i) flaps are ambisyllabic

[*] Tel.: +1 806 834 7127.
E-mail address: aaron.braver@ttu.edu.

http://dx.doi.org/10.1016/j.lingua.2014.09.004
0024-3841/© 2014 Elsevier B.V. All rights reserved.

(Kahn, 1980; Gussenhoven, 1986), (ii) flaps are syllable final (Selkirk, 1982; Inouye, 1989), and (iii) flaps are non-foot-initial (Kiparsky, 1979). What these three accounts have in common, though, is that they all assume that (complete) neutralization takes place—an underlying contrast between /t/ and /d/ neutralizes to [ɾ] in the given prosodic environment. Under all such traditional analyses, flaps that result from lenition of /t/ and flaps that result from lenition of /d/ are predicted to be identical. Similarly, these traditional analyses do not predict any effects to surrounding vowel duration.

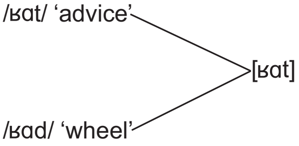
The rest of the paper is organized as follows. The remainder of this section discusses incomplete neutralization both in flapping and more generally. Section 2 describes a production experiment (Experiment 1) comprised of two tasks: a minimal pair reading task and a 'wug'-type task (Berko, 1958). In section 3, I present an identification task (Experiment 2) using stimuli from the production task. A two-alternative forced-choice (2AFC) discrimination task (Experiment 3) is presented in section 4. The concluding section discusses the theoretical claims which have been made about incomplete neutralization and evaluates them in light of the present experiments.

1.1. Incomplete neutralization

Traditionally, phonologists posited that languages were rife with complete neutralization—when two underlyingly distinct sounds become identical on the surface. However, researchers have suggested for some time that even the ''classic'' cases of complete neutralization are not quite as they seem.

Take as an example German final devoicing, which has classically been cited as a case of complete neutralization (Bloomfield, 1933/1984:218–219; Trubetzkoy, 1939/1969:235; Jakobson et al., 1952/1975:9; Hyman, 1975:29, 71–72). Early views of this process can be characterized as follows: an underlying voicing contrast is not preserved on the surface, where only voiceless segments are allowed syllable-finally (schematized in (1)).

(1) The traditional picture of German coda devoicing



A number of studies show, however, that the picture is not quite so simple. While the contrast between underlyingly voiced and voiceless segments is reduced (or partly neutralized) when they are 'devoiced', a trace of the underlying voicing distinction remains on the surface (Dinnsen and Garcia-Zamor, 1971 (in disyllables only); Taylor, 1975 (for some places of articulation); Mitleb, 1981a,b; Port and O'Dell, 1985; Smith et al., 2009; Kleber et al., 2010; Röttger et al., 2011, 2014). In other words, German /ʁɑt/ is not identical to /ʁɑd/–even on the surface.[1] Port and O'Dell (1985), for example, show that the vowel preceding a devoiced segment in German is in fact longer than the vowel preceding an underlyingly voiceless segment by about 15 ms on average.

Phenomena such as this one have been termed *incomplete neutralization* (Port et al., 1981; Fourakis and Iverson, 1984). While two underlyingly distinct segments move toward convergence, some trace of their underlying distinction remains on the surface (schematized in (2b), with the surface trace notated as a superscript feature value). In the German example above, the 'trace' of voicing is the vowel duration distinction. In contrast, in *complete* neutralization (2a), two underlyingly distinct segments become phonetically identical—with no trace of the underlying distinction.

(2) a. Complete neutralization

$\underset{[\alpha F]}{/X/} \rightarrow [Z]$ / (Context A)

$\underset{[\beta F]}{/Y/} \rightarrow [Z]$ / (Context A)

b. Incomplete neutralization

$\underset{[\alpha F]}{/X/} \rightarrow [Z^{(\alpha F)}]$ / (Context A)

$\underset{[\beta F]}{/Y/} \rightarrow [Z^{(\beta F)}]$ / (Context A)

[1] Fourakis and Iverson's (1984) study of German final devoicing argues that when a speaker is aware of an experiment's focus on pronunciation, incomplete neutralization results, but that when this focus is disguised, final devoicing is completely neutralizing. See section 1.2.3 for further discussion of this and related views.

The surface trace in incomplete neutralization trends in the same direction as the canonical realization of the underlying distinction in contexts where the contrast is fully realized (Port and O'Dell, 1985:460; van Rooy et al., 2003:60). In cases of final devoicing, this means that vowels before underlyingly voiced stops are longer than vowels before underlyingly voiceless stops, since this is the same direction as the cross-linguistic canonical realization of vowels preceding voiced and voiceless stops in non-neutralizing contexts (Chen, 1970). Port and O'Dell (1985) also find that there is less aspiration, longer voicing into closure, and marginally shorter closure duration in underlyingly voiced stops in German—all traits traditionally associated with voiced stops in non-neutralizing contexts (Lisker, 1986; Kluender et al., 1988; Kingston and Diehl, 1994). If incomplete neutralization is indeed occurring in flapping, we should expect similar distinctions between /d/- and /t/-flaps, with the former maintaining some /d/-like properties, and the latter maintaining some /t/-like properties.

1.2. Previous studies

1.2.1. Incomplete neutralization of flapping: previous acoustic results

While final devoicing has received much attention in experimental studies (see section 1.2.3), flapping in American English has also been put forward as a case of incomplete neutralization. In flapping, underlying intervocalic /d/ and /t/ become output [ɾ] in certain prosodic configurations (Kahn, 1980; on the production and perception of flaps, see Warner et al., 2009; Warner and Tucker, 2011)—a situation that looks at first glance like a case of complete neutralization. However, a number of studies report that for some speakers flaps stemming from underlying /d/ ('/d/-flaps') and those that stem from underlying /t/ ('/t/-flaps') are in fact phonetically distinct—in other words, they maintain some trace of the voicing status of their input correspondent. These reported distinctions between /d/-flaps and /t/-flaps include differences in the duration of the preceding vowel, degree of intensity dip during closure, and duration of closure (Fisher and Hirsh, 1976; Fox and Terbeek, 1977; Zue and Laferriere, 1979; Patterson and Connine, 2001:264–267). Other studies, however, report that flapping is completely neutralizing, at least for some speakers (Joos, 1942; Port, 1976, the latter of which examined only flapping contexts including [ɪ]). Huff (1980) examines New York City speakers, and finds mixed results: the F1 and F2 of /æ/ and /aɪ/[2] preceding /d/-flaps and /t/-flaps patterned with those preceding non-flapped /d/ and /t/, though this was not the case for /aw/. In his study, all pre-flap vowels in monosyllabic words were significantly longer before /d/ than before /t/, but in disyllabic words, only /aɪ/ was significantly longer before /d/ than before /t/.

A more recent study of flapping in American English, Herd et al. (2010), found evidence of incomplete neutralization. Vowels preceding /d/-flaps were on average 6 ms longer than vowels preceding /t/-flaps—a statistically significant difference. This result trended stronger in bimorphemic words ('wetting' vs. 'wedding'; 8 ms average difference) as compared to monomorphemic words ('petal' vs. 'pedal'; 3 ms average difference). However, this distinction between monomorphemic and bimorphemic words is based on a small subset of stimuli (three monomorphemic pairs and five bimorphemic pairs), and did not reach statistical significance. Additionally, unlike previous studies, the durations of /d/-flaps themselves were found to be on average 0.9 ms longer than /t/ flaps—a difference in duration opposite of the canonical realization of unflapped /t/ and /d/. While this difference was found to be statistically significant, Herd et al. (2010) suggest that this is likely due to the large sample size, and that flap duration is not a cue to the underlying voicing status of a flap.

1.2.2. Incomplete neutralization of flapping: previous perception results

Early studies of listeners' ability to categorize /d/- and /t/-flaps show mixed results (Oswald, 1943; Sharf, 1960; Malécot and Lloyd, 1968; Fisher and Hirsh, 1976). Sharf (1960) presents an identification task with English minimal pairs from a male and a female speaker. For tokens from the female speaker, listeners were accurate more than 86% of the time, while for the male speaker, they were accurate 61% of the time. Sharf notes, however, that the female speaker may have produced a [t] rather than a flap in /t/ words whereas the male speaker used a flap in these words, perhaps accounting for the difference across speakers. On the other hand, Malécot and Lloyd (1968), who used a similar procedure as Sharf (1960), found that their 50 listeners performed at only 56.6% accuracy. Fisher and Hirsh (1976), in addition to their production study, found that five out of six phonetically-trained judges were able to correctly categorize /d/- and /t/-flaps to some degree, though some judges were better than others (χ^2 values ranged from 4.24 to 24.52).

More recently, along with their production study, Herd et al. (2010) present a forced-choice identification task using four pairs of actual English words. Using recordings from the production experiment, three conditions were created: the 'mean' condition (with duration differences between pre-/d/ and pre-/t/ vowels of 7–9 ms), the 'enhanced' condition (with duration differences between 22 and 34 ms), and the 'opposite' condition (in which, unlike the other two conditions, vowels were longer before /t/-flaps than before /d/-flaps, ranging from a difference of 5 ms to a difference of 21 ms). None of the tokens' vowel durations were manipulated—the three conditions were created simply by selecting tokens that fit the criteria of each condition.

[2] Huff (1980) renders this diphthong as /ay/.

Listeners heard tokens from each of the four word pairs, and were asked to select which member of the pair they heard (e.g., when listeners heard [ˈlirɚ], they chose between ‘leader’ and ‘liter’). On average, listeners fell near chance, scoring 52% correct, 52% correct, and 48% correct in the mean, enhanced, and opposite conditions, respectively. Tokens that contained an underlying /d/ (e.g., ‘leader’) were more often identified correctly (57%) than words containing underlying /t/ (e.g., ‘liter’, 44% correct).

1.2.3. Incomplete neutralization generally

The results of the studies on flapping are broadly similar to those of final devoicing—perhaps the most commonly cited instance of incomplete neutralization. Acoustic studies finding incomplete neutralization in final devoicing have been conducted on a number of languages, including Afrikaans (van Rooy et al., 2003), Catalan (Dinnsen and Charles-Luce, 1984), Dutch (Warner et al., 2004, though see that article itself and Warner et al., 2006 for caveats; Ernestus and Baayen, 2006, 2007), German (Dinnsen and Garcia-Zamor, 1971; Taylor, 1975; Port and O’Dell, 1985; Mitleb, 1981a,b; Smith et al., 2009; Kleber et al., 2010; Röttger et al., 2014), Polish (Slowiaczek and Dinnsen, 1985; Slowiaczek and Szymanska, 1989), and Russian (Dmitrieva et al., 2010; Kharlamov, 2012, 2014).

Other production studies, though, have either not found incomplete neutralization, or have suggested that it results from extragrammatical factors. For example, Fourakis and Iverson (1984) found support for claims of incomplete neutralization in German devoicing only in a word-list reading task, but not when the target words were elicited as part of a morphological paradigm completion task, in which the experiment’s focus on pronunciation was masked and speakers did not see the written form of the target word. Inozuka (1991) also reports acoustic evidence that shows complete neutralization of the voicing contrast in German codas. Similarly, Kopkallı (1993), on the basis of an experimental study, argues that devoicing in Turkish is completely neutralizing. Along the same lines as Fourakis and Iverson (1984), Warner et al. (2006) show that in Dutch, an underlying /t/ vs. /t-t/ distinction is completely neutralized in words where this distinction is not represented orthographically, even though Warner et al. (2004) show that subphonemic distinctions are present in words where the same contrast is primarily orthographic, rather than underlying (e.g., *baten* /batən/ ‘to avail’ vs. *baatten* /batən/ ‘they availed’).

In order to test whether German listeners are able to identify whether a given word contains a devoiced or an underlyingly voiceless consonant, Port and O’Dell (1985) ran a forced-choice identification task. Listeners were given an answer sheet with minimal pairs written for each token, and were asked to circle which word they heard on each trial. On average, listeners performed better than chance (59% correct). From this, Port and O’Dell (1985) conclude that German listeners can, in fact, make use of the incompletely neutralized voicing distinction. This result contrasts with the findings of Kopkalli’s (1993) similar identification tasks on Turkish devoicing: participants in that study performed nearer to chance.

Warner et al. (2004), in addition to their acoustic experiments on Dutch final devoicing, present several perception tasks. In their production experiment, they found that underlying voicing significantly affected vowel duration in the neutralization environment: pre-/d/ vowels were on average 3.5 ms longer than pre-/t/ vowels (though see Warner et al., 2006). Warner et al. (2004) find in an identification task that listeners perform better than chance in distinguishing final /d/ from final /t/ on tokens from two speakers, but that listeners’ accuracy is still quite poor (in the condition in which listeners performed best, $d' = 0.33$). In further identification tasks, composed of vowel and consonant closure duration continua, they found that listeners can use both as cues to underlying voicing—even though only vowel duration is linked to final voicing in their production data.

These studies leave open a number of questions which apply to incomplete neutralization in both devoicing and flapping. First, to what degree are findings of incompletely neutralized distinctions in production studies generalizable to other situations? In particular, do they extend to tasks in which the effects of hyperarticulation and/or orthography are less expected? Second, can listeners perceive these small distinctions? In all cases, or just some? And finally, does listeners’ poor performance in identification tasks on incomplete neutralization extend to other task types?

1.2.4. Challenges and remaining questions

While much evidence has been adduced for incomplete neutralization in flapping (see sections 1.2.1 and 1.2.2), final devoicing (see section 1.2.3), and other contexts (Gerfen, 2002; Yu, 2007), it has been suggested that putative cases of incomplete neutralization may be the result of extragrammatical or experimental factors (Fourakis and Iverson, 1984; Mascaro, 1987; Manaster Ramer, 1996a; Warner et al., 2006; though see Port and Crawford, 1989). This concern, however, has not been addressed in the literature on flapping. The production tasks in section 2 were designed to tease apart such factors, and in so doing support previous acoustic studies of flapping that have found that there is indeed a relatively small surface distinction between /d/-flaps and /t/-flaps (Fisher and Hirsh, 1976; Fox and Terbeek, 1977; Zue and Laferriere, 1979; Herd et al., 2010).[3]

[3] It should be noted that the small size of the distinction between /d/-flaps and /t/-flaps is dependent on the context of production (Charles-Luce, 1997; Charles-Luce and Dressler, 1999; Charles-Luce et al., 1999). In other words, small distinctions cannot be dismissed as linguistically non-significant (Dinnsen and Charles-Luce, 1984:56).

Given that the surface distinction between /d/- and /t/-flaps has been found to be relatively small (Fisher and Hirsh, 1976; Fox and Terbeek, 1977; Zue and Laferriere, 1979; Patterson and Connine, 2001), there has been some interest in the perceptibility of this distinction (Oswald, 1943; Sharf, 1960; Malécot and Lloyd, 1968; Fisher and Hirsh, 1976; Herd et al., 2010). Such studies are generally motivated by the assumption that ability to identify the distinction is suggestive of categoricality—even in the face of a seemingly gradient phenomenon like incomplete neutralization. Perception studies, however, have not conclusively determined whether it is, indeed, perceptible. These studies, which tend to use a small number of English words, focus almost exclusively on identification tasks. While early identification studies find mixed results (Oswald, 1943; Sharf, 1960; Malécot and Lloyd, 1968; Fisher and Hirsh, 1976), a more recent study by Herd et al. (2010) presents an identification task in which listeners were unable to categorize /d/-flaps and /t/-flaps—a result further supported by the identification study described in section 3. In order to further probe listeners' ability to distinguish /d/-flaps from /t/-flaps, a discrimination task is presented in section 4.

1.3. Theoretical implications

Incomplete neutralization can serve as a window into the shape of the phonetics–phonology interface. Early models of the grammar relegated phonetics and phonology to completely different domains, with most computation done in the phonology, and a relatively automatic phonetics. As Keating (1988:287) summarized the early situation, ''[t]he phonetic component consisted mostly of automatic, universal rules for implementing [the] feature matrices [provided by the phonology] as continuous physical events.'' Under these early models, the phonetic module was granted access only to a limited subset of the phonological computations. While the phonology had inputs, outputs, morphological structure, and prosodic structure, along with other byproducts of phonological computation, the phonetic module was granted access to only a small selection of this structure, such as outputs and pieces of prosodic structure.

In such classical modular feedforward architectures (see e.g., Bermúdez-Otero, 2007:501–503), the existence of consistent differences—however small—entails that words with /d/-flaps and those with /t/-flaps cannot possibly have completely identical phonological surface representations (SRs) that feed the phonetic module. However, given listeners' poor performance on identification tasks, /d/-flaps and /t/-flaps should not be assigned different SRs as this would predict (contrary to fact) an ability to categorize. A possible explanation in terms of the classical architecture plays on the fact that the surface distinction between /d/-flaps and /t/-flaps is mainly realized in terms of preceding vowel duration. Under this view, /d/-flaps and /t/-flaps themselves have identical feature specifications in the SR, but the preceding vowel's representation has been affected in some way during the course of the derivation.

Two important caveats to this 'vowel representation' explanation are necessary. First, listeners must somehow fail to interpret the effects on the vowels' representations as cues to the underlying laryngeal specification of the flaps—otherwise we would expect that listeners are able to easily categorize /d/-flaps and /t/-flaps. Second, if distinctions as small as the vowel duration differences found in flapping are encoded in the SR, one might expect that the (larger) distinction in non-neutralizing contexts between pre-voiced and pre-voiceless vowels should similarly have representation at SR. The encoding of such non-categorical distinctions in the phonological SR, however, seems contrary to the spirit of the modular feedforward architecture.

Since the phonetics seems to play on underlying phonological contrasts in incomplete neutralization, *contra* the modular feedforward model described above, the phenomenon has been used to argue for a number of modifications and alternative proposals. Such claims fall broadly into four categories: (i) variation in phonetic realization of phonological categories, (ii) questions of representation and category membership, (iii) adoption of non-categorical models (e.g., Exemplar Theory; Bybee, 2001; Pierrehumbert, 2001, 2002), and (iv) the rejection of formal phonology, at least so far as incomplete neutralization is concerned. (See section 5 for discussion of treatments of incomplete neutralization within a classical model.)

As an instance of the first category, Yu (2011) argues that incomplete neutralization is better analyzed in a model that allows the phonological module more control over the variation in phonetic implementation of contrasts. In such a model (like the one proposed by Kingston and Diehl (1994)), the subphonemic differences that are the hallmark of incomplete neutralization are ''qualitatively not different from those observed between allophones appearing in different phonetic contexts'' (Yu, 2011:311).

A different approach is to modify the representation of the traditional categories being neutralized, or similarly, to challenge the membership of such categories. For example, van Oostendorp (2008) argues that in German final devoicing, the phonetic module can distinguish between representations of underlyingly voiceless and devoiced segments. Under this model, devoiced segments maintain a 'projection' relation to their [voice] feature (even in the phonological output), while voiceless segments have no such relation. This difference in representation is then interpreted at the phonetic level as, e.g., a distinction in the duration of preceding vowels. Relatedly, Steriade (1997:48–49) argues that phonetically voiceless segments have a voicing target, while devoiced segments are phonetically targetless. The targetless segments become phonetically devoiced by means of an automatic, passive process which renders these segments largely voiceless, yet still distinct from underlyingly voiceless segments. A further extension of this idea is explored in Ernestus (2000). To account for Dutch final devoicing, Ernestus argues that word- and syllable-final obstruents (i.e., those that have the potential to neutralize) ''have no phonological [voice]-specification at all'' (p. 157). Rather, the realization of these non-specified coda obstruents is determined by ease of articulation: either voiced or voiceless depending on which articulation requires no additional effort. The non-specified obstruents are distinct, however, from phonologically specified [-voice] obstruents, as found in onsets.

A further departure from traditional models of phonology and phonetics are models which eschew categories *per se*. Most notable among these are Exemplar Theoretic models (Bybee, 2001; Pierrehumbert, 2001, 2002), which build labels by comparing new instances to instances already stored in episodic memory. As (Hansson, 2011:335) puts it, ''in the production of a given category in a given environment, stored exemplars of that category from *any* environment can in principle be activated and contribute to the calculation of a production target. . ..'' Thus, in the case of final devoicing, for example, incomplete neutralization results from the activation of exemplars of obstruents—some of which are in neutralizing contexts, but the majority of which are not.

A related approach adopts aspects of the pure modular feedforward architecture into a 'hybrid' model, in which phonetic implementation is conditioned by both a categorical phonological parse as well as episodic memory (i.e., exemplars). Pierrehumbert (2002) describes just such a model, in which it would be possible for /d/-flaps and /t/-flaps to have identical, categorical phonological parses yet result in distinct phonetic realizations. Under Pierrehumbert's (2002) model, flapped words correspond to exemplar clouds with different statistical distributions—when *writer* is activated, activation spreads to *write*; when *rider* is activated, activation spreads to *ride*. *Writer* is then probabilistically produced with a slightly shorter vowel, as the vowel in *write* displays pre-fortis clipping. The results of the acoustic experiment presented in this paper, which makes use of nonce words, have significant implications for this model (see section 2.5.2). Since novel nonce words presumably have no exemplars to call upon, it is unclear how word-based exemplar models can account for incomplete neutralization of such words.

The final type of claim is the rejection of formal phonology, or at least its relevance to neutralization. Proponents of this view argue that putative cases of *complete* neutralization are actually incomplete, and that therefore a categorical phonology in the traditional sense cannot hold. For example, Dinnsen (1985) argues that ''there appear to be no empirically defensible cases of the Type A [i.e., complete] neutralizations'' (p. 273) and that ''the construct [complete] 'neutralization' (and any principles of grammar formulated in terms of it) may be empirically indefensible'' (p. 276). In a similar spirit, Port (1996) and Port and Leary (2005) suggest that incomplete neutralization serves as evidence against formal phonological systems that employ categorical phonetic units. Port (1996:508) points to German final devoicing and English flapping as instances in which the mapping between abstract symbol and phonetic reality is not entirely clear: ''whether they [the neutralized segments] are the same or different depends on how you ask the question.'' Port and Leary (2005:947) follow this line of reasoning, arguing that ''if incomplete neutralization phenomena [are] correct, then it would imply that linguists cannot rely on their own or anyone's auditory transcription.'' To summarize: this view holds that the reliance on discrete categories as the basic units of the theory means that the gradient realm of neutralization cannot be accounted for.

In addition to the four types of claims described above, an additional possible view is that the apparent preservation of (at least some) underlying contrasts is an effect of methodological, pragmatic, or extragrammatical factors, and will not generalize beyond the specific task in which the contrast was found (Fourakis and Iverson, 1984; Manaster Ramer, 1996a,b; van Rooy et al., 2003; Warner et al., 2006 among others; cf. Port and Crawford, 1989). By way of example, Fourakis and Iverson (1984:149) argue that ''German final devoicing is a bona fide instance of phonological neutralization,'' dismissing experimental results to the contrary as ''hypercorrect manifestations of linguistic insecurity'' not generalizable to situations outside of the laboratory. They further claim that most putative cases of complete neutralization are indeed complete so long as the context is 'natural', but that ''under unnatural conditions, one may expect unnatural results.'' In other words, findings of incomplete neutralization may not generalize beyond the conditions tested in the laboratory. Along similar lines, van Rooy et al. (2003:62, 63) conclude that certain specific task types can lead to incomplete neutralization but that ''[n]ormal speech in Afrikaans is characterized by complete neutralization [in final devoicing].''

1.4. Motivation for the current study

The acoustic experiment reported in this paper attempts to provide additional evidence for incomplete neutralization in American English flapping by addressing a number of questions raised by previous studies. Chief among these questions are the issues of hyperarticulation and effects of orthography (Fourakis and Iverson, 1984; Port and Crawford, 1989; Warner et al., 2006)—do previous findings generalize to situations where such effects are less likely? Of the two tasks in Experiment 1, one was designed to have relatively more effects of hyperarticulation and orthographic influence (the 'minimal pair task'), while the other (the 'wug task', Berko, 1958) was designed to have relatively fewer such effects. This paradigm was largely inspired by Fourakis and Iverson's (1984) study of incomplete neutralization in German final devoicing, which reports two tasks: an 'elicitation task' in which speakers conjugated verbs based on oral cues, and a 'reading task' in which speakers read target words from index cards. By showing that speakers incompletely neutralize under both conditions in Experiment 1, I argue that incomplete neutralization of flapping in American English occurs in both situations with and without direct exposure to an orthographic representation or minimal pair at the time of production.

An additional concern is the effect of lexical frequency on both production and perception. For example, Herd et al. (2010) find lexical frequency effects of two sorts. First, words with high lexical frequency were more often correctly identified (59% accurate) than words with low lexical frequency (42%). Additionally, they found an interaction between lexical frequency and underlying voicing: /d/ words behaved relatively similarly regardless of lexical frequency (62% correct for words with high lexical frequency, 51% correct for words with low lexical frequency), while /t/ words behaved quite differently based on lexical frequency (55% correct for words with high lexical frequency, 33% correct for words with low lexical frequency).

In order to mitigate possible effects of lexical frequency, which can apply in both production and perception, the tasks reported in this paper use nonce words only. Using nonce words also allows for a greater number of different stimuli—English has only a limited number of minimal pairs that differ just by /d/ or /t/ in a flapping environment.

Additionally, one can ask whether results like those in Herd et al. (2010)—poor performance in perceiving /d/- and /t/-flaps—are due to the type of task used (identification). In order to probe this possibility, two perception tasks are presented here: an identification task (Experiment 2) similar to the one in Herd et al. (2010), as well as a 2-Alternative Forced Choice (2AFC) discrimination task (Experiment 3). The tasks were designed to give listeners a good chance of perceiving the intended contrast. As such, a practice phase was included at the start of each task in this experiment, including a section with real English words and a section with nonce words, in order to acclimate the listeners to both the format of the task, and the sort of tokens that they would be hearing. Additionally, feedback was provided after each trial.

As a further consideration, it is worth noting that a strong bias for /d/ was found by Herd et al. (2010). /d/-words in their perception task were accurately perceived 57% of the time, while /t/-words were accurately identified 44% of the time. The results of all perception tasks presented in this paper are given in d', which unlike the percent-correct measure reported by Herd et al. (2010), teases apart sensitivity from bias (Macmillan and Creelman, 2005). Bias of this sort could potentially lead to misinterpretation of experimental results. For example, if a listener says that they heard a /d/ word on all trials—regardless of what they had actually heard, they would still be accurate on 100% of /d/ trials. (They would, of course, also score 0% accuracy on /t/ trials). The percent-correct measure could hypothetically lead to an interpretation that listeners are good at finding /d/ words and bad at finding /t/ words, when in reality the results are due only to the listener's bias toward responding /d/.

Finally, listeners' poor performance on identification tasks of flapping can be an effect of either (a) listeners' grammars lacking the relevant phonological categories, or (b) an effect of the task employed. The use of discrimination tasks such as the 2AFC task[4] in section 4, shed light on these possibilities. If listeners perform well on the 2AFC task, it will be clear that listeners' poor performance in the identification of flapping is not generalizable to other task types. If, however, listeners perform poorly on discrimination tasks, listeners' poor performance on identification tasks is a consequence—robust across tasks—of an inability to distinguish the sounds.

2. Experiment 1: Acoustic experiment

This experiment was comprised of two tasks, with the goals of (a) verifying whether speakers produce words with /d/-flaps differently from those with /t/-flaps, and (b) showing that this result is robust both in contexts with relatively more extragrammatical factors at play, and those with relatively fewer. The 'minimal pair task' was designed to increase the effects of hyperarticulation and orthographic influence, while the 'wug task' was designed to reduce these effects (see Fourakis and Iverson, 1984 for a similar design in an experiment on German final devoicing). By showing that speakers incompletely neutralize the voicing distinction in flapping under both conditions, it is argued that these and previous results are robust even in contexts where direct exposure to orthographic representation of the flapping context and minimal pairs is minimized.

2.1. Stimuli

The nonce-word stimuli in this experiment were used in both the wug task and the minimal pair task, with minor differences described below. Minimal pairs of disyllabic nonce words were created, ending in either /d/ or /t/, such that adding the '-ing' suffix to these stimuli would place the alveolar stop in a post-tonic intervocalic context where flapping occurs.

[4] 2AFC is a discrimination task because there are two stimuli per trial; however this task type also involves explicit labeling.

Table 1
Stimuli from the acoustic experiment.

/σ-pi{t,d}/		/σ-tæ{t,d}/	
puhPEET	puhPEED	puhTAT	puhTAD
buhPEET	buhPEED	buhTAT	buhTAD
tuhPEET	tuhPEED	tuhTAT	tuhTAD
duhPEET	duhPEED	duhTAT	duhTAD

/σ-pɛ{t,d}/		/σ-ki{t,d}/	
puhPEHT	puhPEHD	puhKEET	puhKEED
buhPEHT	buhPEHD	buhKEET	buhKEED
tuhPEHT	tuhPEHD	tuhKEET	tuhKEED
duhPEHT	duhPEHD	duhKEET	duhKEED

/σ-pæ{t,d}/		/σ-kɛ{t,d}/	
puhPAT	puhPAD	puhKEHT	puhKEHD
buhPAT	buhPAD	buhKEHT	buhKEHD
tuhPAT	tuhPAD	tuhKEHT	tuhKEHD
duhPAT	duhPAD	duhKEHT	duhKEHD

/σ-ti{t,d}/		/σ-kæ{t,d}/	
puhTEET	puhTEED	puhKAT	puhKAD
buhTEET	buhTEED	buhKAT	buhKAD
tuhTEET	tuhTEED	tuhKAT	tuhKAD
duhTEET	duhTEED	duhKAT	duhKAD

/σ-tɛ{t,d}/	
puhTEHT	puhTEHD
buhTEHT	buhTEHD
tuhTEHT	tuhTEHD
duhTEHT	duhTEHD

The initial (non-target) syllable in each token was composed of a simple onset (one of /p, b, t, d/) with a schwa nucleus and no coda. The second (target) syllable in each token was composed of a simple onset (one of /p, t, k/), a vocalic nucleus (one of /æ, ɛ, i/), and a final /d/ or /t/. These combinations yielded 72 tokens, comprised of 36 minimal pairs. The stimuli were written in English orthography, with the second syllable capitalized to indicate stress, and with schwas indicated by 'uh' (for example, puh-PAT~puh-PAD, and tuh-KEET~tuh-KEED). The stimuli are listed in Table 1.[5]

36 filler pairs were also created. These stimuli consisted of an initial unstressed syllable followed by a stressed (closed) syllable, none of which ended in /d/ or /t/. The fillers were written in the same orthography as the target stimuli (e.g., nuh-SOOF~nuh-SOON, and zuh-DOOK~zuh-DOON).

2.2. Participants and equipment

12 speakers participated in both tasks in this experiment. All speakers reported being raised primarily in New Jersey and adjacent areas. Each speaker made two visits to the Rutgers Phonetics Laboratory, seven days apart, with task order balanced across speakers. Recordings were made in a sound-attenuated booth, using an Audio-Technica AT4040 cardioid capacitor microphone with a pop filter, amplified through an ART TubeMP microphone pre-amplifier and a JVC RX 554V amplifier. The speech was digitized as WAV files at a sampling rate of 44.1 kHz using Audacity (Audacity Team, 2008). Acoustic analysis was performed using Praat (Boersma and Weenink, 2009).

[5] One stimulus item, puh-TEET, is similar to the actual English adjective 'petite'. 'Petite', however, is ungrammatical in the context in which speakers were asked to produce these items. Neither the flap closure duration nor the pre-flap vowel duration of puh-TEET differed significantly from other stimuli ending in TEET (closure duration: $t(29.82) = -1.78$, *n.s.*; vowel duration: $t(15.47) = -0.79$, *n.s.*).

2.3. *Procedure*

2.3.1. *The minimal pair reading task*

The minimal pairs constructed as described in section 2.1 were put into the progressive '-ing' form (e.g., the tokens 'puh-PAT' and 'puh-PAD' became 'puh-PAD-ing' and 'puh-PAT-ing', respectively). The stimuli were then presented visually to speakers on a computer screen. On each trial, speakers saw two sets of sentences, consisting of two sentences each, on the screen. In the first set, speakers saw one sentence with a nonce word in its 'plain' form (i.e., without the '-ing' suffix), and a second sentence with the same nonce word in its '-ing' form. The second set of sentences was identical to the first set, except that the nonce word from the first set was replaced with the other member from its minimal pair. Whether the /d/- or /t/-member of the minimal pair was displayed in the first or second set was randomized. On each trial, speakers saw something like (3):

(3) John learned how to puh-PAT this week.
He was puh-PAT-ing this whole week.

John learned how to puh-PAD this week.
He was puh-PAD-ing this whole week.

Speakers were asked to read each sentence aloud naturally. This procedure was repeated for all 36 minimal pairs (=72 target items), randomized, with 36 filler pairs (=72 total filler items).

This task was designed to increase speakers' attention to both pronunciation and orthography (see the 'reading task' in Fourakis and Iverson, 1984 and the 'contrastive sentences' and 'dictation sentences' conditions in Port and Crawford, 1989). Since speakers saw both members of a minimal pair on the screen at the same time, the task highlighted the distinction between the voiced and voiceless stimuli.

2.3.2. *The wug task*

In the wug task, speakers were shown the 'plain' form of the 72 stimuli, and were asked to fill the '-ing' form of these words into a blank in a frame sentence. Unlike the minimal pair task, only one stimulus was shown on the screen at a time—the stimuli were not presented as minimal pairs. The order of stimuli was randomized, with 72 filler items. On a given trial, speakers saw a screen like (4):

(4) John learned how to puh-PAT this week.

He was ____________ this whole week.

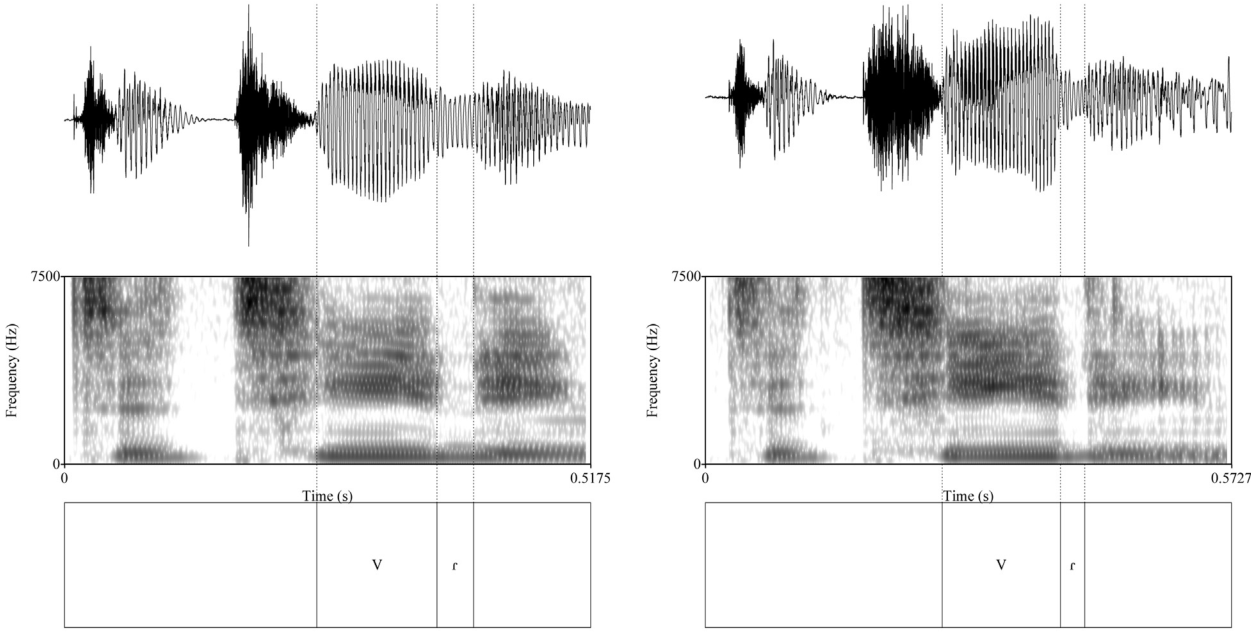
After an initial training phase, speakers reliably filled in the '-ing' form of the nonce word, producing, for example, (5):

(5) John learned how to puh-PAT this week.
He was puh-PAT-ing this whole week.

This task was designed with Fourakis and Iverson's (1984) 'elicitation task' in mind. While the morphology of the English progressive is simpler than the German strong verb system employed in their task, the task presented here seeks to reduce the influence of orthography, while at the same time ''diminish[ing] the obviousness of focus on pronunciation'' (Fourakis and Iverson, 1984:142). By instructing speakers to focus on filling in the blank, rather than on their pronunciation, the task aimed to elicit more natural productions of the target words. Similarly, while speakers did see the plain form of the nonce word on the screen, they did not see the target '-ing' form, mitigating somewhat the effects of orthography, especially as compared to the minimal pair reading task. Finally, unlike in the minimal pair task, speakers did not see both members of a minimal pair on a given trial, reducing the salience of the distinction between the voiced and voiceless stimuli.

2.3.3. *Acoustic measurements and statistical analysis*

The [voice] feature has a number of known acoustic correlates (Lisker, 1986), including preceding vowel duration (Chen, 1970), closure duration (Kluender et al., 1988), effects on the F0 and F1 of surrounding vowels (Hombert et al., 1979; Kingston and Diehl, 1994), and F1-F2 divergence in the offglide of closing (i.e., upgliding) diphthongs (Thomas, 2000; Moreton, 2004; before /t/-flaps specifically: Kwong and Stevens, 1999). First, the duration of vowels preceding the flapped segments was measured from the onset of voicing to the onset of the flap closure. The onset of flap closure was marked as the location on the spectrogram with a marked reduction in formant structure, accompanied by a drop in intensity and periodic energy as seen in the waveform. See Fig. 1 for representative examples.



'duhteet-ing'

'duhteed-ing'

Fig. 1. Representative spectrograms (from speaker 6, minimal pair task).

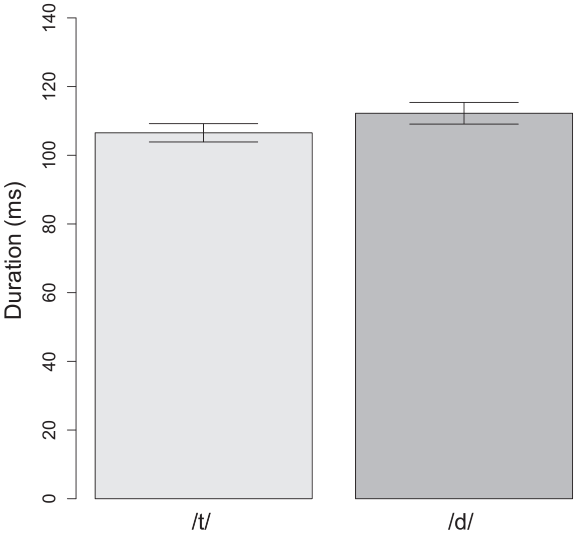
While preceding vowel duration tends to be the focus of studies of incomplete neutralization, it is far from the only known correlate of voicing (Hombert et al., 1979; Kluender et al., 1988; Kingston and Diehl, 1994; Kwong and Stevens, 1999; Thomas, 2000; Moreton, 2004). As such, the following measurements were also taken: closure duration, percent voicing during closure, and the slope of F0 and F1 at 10 ms preceding the onset of flap closure and 10 ms following the offset of flap closure. Closure duration was calculated as the distance between the onset of flap closure (as described above) and the offset of flap closure, which was marked as the location on the spectrogram where formant structure resumed, and intensity and periodic energy increased as seen in the waveform. Percent voicing during closure was determined by Praat's 'Voice Report' function. F0 and F1 were measured 10 ms on either side of the flap onset boundary and the flap offset boundary (as described above), and the slope between these points and their respective values at the flap onset/offset boundary itself was calculated.

Tokens more than 2 standard deviations from the mean on any of the recorded measurements were considered outliers and discarded (a total of 750 observations remained after this procedure). A linear mixed model (Baayen, 2008) was run using the `lme4` package (Bates and Maechler, 2009) in R (R Development Core Team, 2009). Pre-flap vowel duration was regressed against a model in which underlying voicing status and task were fixed factors. Vowel height (contrast coded) was also included as a fixed factor to soak up variability, since it is known to affect vowel duration (Peterson and Lehiste, 1960; Lehiste, 1970). An interaction term between underlying voicing status and task was also included in order to examine the effects of underlying voicing status on neutralization that might be present in one task but not the other. Speaker and item were included as random factors.[6] To examine the multiple acoustic correlates of voicing, identical models were run substituting the following measurements as the fixed factor: closure duration, percent voicing during closure, and the slope of F0 and F1 10 ms preceding/following flap closure onset/offset.

2.4. Results

I focus in this section on the results pertaining to preceding vowel duration, because the other measures investigated were not found to be significantly impacted by underlying voicing status. The results of these other measures are summarized in the appendix, in Table 4. The linear mixed model described in section 2.3.3 was compared with a simpler

[6] The procedure for calculating degrees of freedom in this type of model is unknown, so the significance of the coefficients was checked by the `pvals.fnc` function of R's `languageR` package (Baayen, 2009), which uses the Markov Chain Monte Carlo method.



(Error bars show 95% CI)

Fig. 2. Mean pre-flap vowel duration by underlying voicing status.

Table 2
Mean pre-flap vowel duration (ms) for all speakers in the acoustic experiment, across both tasks.

Speaker	1	2	3	4	5	6	7	8	9	10	11	12
Before /d/-flap	122.65	86.79	86.02	90.76	71.31	131.11	119.50	120.24	103.2	94.47	105.58	138.46
Before /t/-flap	114.61	87.61	88.43	102.11	73.98	115.83	114.94	118.88	88.6	93.44	98.71	132.92
Difference	8.04	−0.83	−2.41	−11.36	−2.67	15.28	4.55	1.36	14.60	1.04	6.87	5.54

model excluding underlying voicing status; a log-likelihood ratio shows a significant difference between the two models ($\chi^2(6) = 35.77$, $p < .001$), so the more complex model was retained.

2.4.1. Task

Task did not have a significant main effect on target vowel duration—pre-flap vowels were similar in duration in the wug task and the minimal pair task (wug mean: 110.20 ms, minimal pair mean: 108.45 ms, mean difference: 1.74 ms, $t = 0.54$, *n.s.*). Similarly, the effect of the interaction between task and underlying voicing on pre-flap vowel duration was not significant ($t = 0.82$, *n.s.*), suggesting that the difference in vowel duration before /d/- and /t/- flaps was consistent across tasks. Since task was not a significant factor in the model, the description of underlying voicing status below is not divided by task.

2.4.2. Underlying voicing status

Among all stimuli (pooled from both tasks), underlying voicing status had a significant effect on vowel duration. Vowels preceding /d/-flaps were longer than those preceding /t/-flaps (mean /d/: 112.23, mean /t/: 106.54, mean difference: 5.69 ms, $t = 3.77$, $p < 0.001$). The mean duration of vowels preceding /d/-flaps and /t/-flaps can be seen in Fig. 2, with these results broken down by speaker in Table 2.

2.5. Discussion

Acoustic results suggest that flapping in American English is incompletely neutralizing, at least for some speakers. The underlying voicing status of a flap is reflected on the surface in the duration of the preceding vowel—vowels before /d/-flaps are longer than vowels before /t/-flaps. This pattern mirrors the distinction in preceding vowel duration seen in voiced segments generally (Chen, 1970). The fact that the vowel duration distinction between /d/-flaps and /t/-flaps was not significantly different between the two tasks, when considered in the context of previous studies finding incomplete neutralization in flapping, suggests that there is some degree of independence from the extragrammatical factors manipulated here.

The fastest speakers (i.e., those with the shortest vowels before flaps: speakers 2, 3, 4, 5, 10) also produced either essentially no vowel duration distinction or a distinction in the unexpected direction. One potential explanation is that the slower speakers were hyperarticulating (hence the slower speech), and therefore consciously adjusted their performance to differentiate /d/-flaps and /t/-flaps. Informal interviews with the participants after the experiment, however, do not support this hypothesis: most speakers were unable to identify /d, t/ as the target of the experiment.

2.5.1. Effects of task on incomplete neutralization

The two tasks in the acoustic experiment were designed to show that incomplete neutralization in flapping is robust in tasks that vary in terms of their potential for extragrammatical effects along the dimensions manipulated. One possible explanation for the lack of a distinction between the 'wug' and 'minimal pair' tasks is that since the stimuli are nonce words, participants may be especially sensitive to orthography even during an oral task. In other words, speakers' knowledge about the nonce words in these tasks is extremely limited, so they turn to the only data they have been given: in the minimal pair task, this is the orthographic form of the target; in the wug task, this is the orthographic representation of the base form of the nonce verb to which they add '-ing'.

Another extragrammatical factor that can affect incomplete neutralization is the presence of minimal pairs among the stimuli. It has been suggested that stimulus lists containing minimal pairs, even when members of a minimal pair are not presented simultaneously, can encourage the preservation of a phonological contrast (Jassem and Richter, 1989:318; Piroth and Janker, 2004:86; Kharlamov, 2014:48). As such, there are two minimal pair effects potentially at work in the acoustic tasks: (i) the effect of distant minimal pair counterparts, and (ii) the more obvious effect of minimal pairs with both members presented at the same time. Speakers should be expected to undergo greater influence from minimal pairs presented simultaneously than minimal pairs whose members are shown separately, as the contrast in the former case is more immediately apparent. As such, the lack of a task effect between the minimal pair and wug tasks in Experiment 1 suggests that minimal pair exposure of type (ii), at least, does not condition vowel duration in incomplete neutralization.

It should also be noted that the influence of minimal pairs not presented simultaneously relies on speakers' abilities to become aware of the experimental task: after seeing enough minimal pairs (however spread out they might be), speakers figure out the contrast being tested. Informal interviews with participants in the acoustic experiment seem to contradict this assumption: the majority of participants reported that they could not figure out what they were being tested on.[7]

2.5.2. Nonce words and exemplar models

This acoustic experiment used only nonce word stimuli presented orthographically. In this setup, participants' episodic memories contain no traces of previous exposure to the acoustic realization of the target stimuli: nonce words (by definition) have lexical frequency of zero, and orthographic presentation during the experiment does not provide participants an acoustic realization to remember. Cast in terms of Exemplar Theory, nonce words presented orthographically simply do not have exemplar clouds.

This presents a challenge for a particular class of exemplar-theoretic models—specifically those that rely on word-based, rather than category-based storage (see Sóskuthy, 2011, for further discussion on this distinction). The division between word- and category-based exemplar models comes down to a question of level of representation: are exemplars examples of words, or of (categorical) speech sounds? It is not clear how word-based exemplar models (e.g., Johnson, 1997; Bybee, 2001) can account for the incomplete neutralization of the nonce words in this experiment: the exemplar(s) to which the target should be attracted simply do(es) not exist (see Hansson, 2011:335 for discussion of exemplar models in the context of incomplete neutralization). If, on the other hand, exemplars are labeled by phonological category, participants will have been exposed in the past to the category being tested (and thus will have exemplars of this category). We should therefore expect a nonce word containing a member of that category to be attracted to that label.

3. Experiment 2: Identification perception experiment

The vowel duration distinctions of many speakers in the acoustic tasks approach the range of just-noticeable-difference (jnd): the jnd for vowel duration is between 5–10 ms, depending on context (Fujisaki et al., 1975; Nooteboom and Doodeman, 1980). As such, the perceptibility of these distinctions bears investigating. Herd et al. (2010) showed that on a basic identification task of flapped consonants, listeners perform poorly. The experiment reported in this section was designed to replicate this result with nonce words, taking bias into consideration through the use of d' as a measure of sensitivity. It follows the basic format of similar studies into the perception of /t/- and /d/-flaps (Sharf, 1960; Malécot and Lloyd, 1968; Fisher and Hirsh, 1976; Herd et al., 2010), with the important distinction that the stimuli in this task are nonce words, rather than actual words of English.

[7] It is not possible to know for sure whether participants truly could not figure out what they were being tested on. An alternative analysis is that participants were unwilling to provide an answer for fear of being wrong.

3.1. Participants and equipment

21 undergraduates participated in this experiment, none of whom had participated in the production experiment. All participants were native speakers of English. A plurality of participants were born in New Jersey (*n* = 9, 42.9%), and 76.2% (*n* = 16) of the participants report having been raised mostly in New Jersey. The experiment took place at the Rutgers Phonetics Laboratory, with stimuli presented and responses recorded by SuperLab 4.5 (Cedrus Corporation, 2010), through Sennheiser HD 280 Professional headphones.

3.2. Tokens

108 tokens (=54 pairs, 18 per speaker) were selected from among those recorded by the speakers in the acoustic experiment (see section 2.1). Stimuli were chosen from the three speakers who had the biggest difference between pre-/d/ and pre-/t/ vowel duration, and who accurately produced a sufficient number of items. Tokens were chosen from each speaker to maximize the pre-flap vowel duration difference between members of a pair, while at the same time balancing onset and vowel of the target syllable and voicing of the target segment (/d/ or /t/). Note, however, that in spite of selecting the pairs with the greatest vowel duration differences, not every stimulus pair contained particularly robust differences. Items were selected from across both the minimal pair reading task (66.67%) and the wug task (33.33%). Each block had only items from a single speaker. The full set of stimuli used in each of the perception experiments reported in this paper is listed in Table 3.

3.3. Procedure

Prior to the actual experimental task, participants read instructions for the task, practiced with both English and nonce words that had been recorded by the author, and were given an opportunity to ask the experimenter any questions about the procedure.

On each experimental trial, listeners heard a single token (as described above), and were directed to press one of two buttons indicating whether the sound immediately preceding the ‘-ing’ was a /d/ or a /t/. For example, if listeners heard ‘buhKEED-ing’, they would press the button indicating that the sound immediately preceding the ‘-ing’ was a /d/.

As an additional measure to increase performance on this task, visual feedback was given on each trial. After a participant’s response (or failure to respond after 1500 ms) the correct response was shown for 500 ms (colored green if they had been correct, and red if they had been incorrect).

The task consisted of three blocks (one for each of the three speakers from whom tokens were taken). Each block consisted of 36 trials (half /d/ and half /t/), randomized, with three repetitions (=108 trials per block). Listeners were allowed to take a short break between each block. Block order was balanced (Latin Square) across all listeners.

Table 3
Tokens used in all perception experiments. ‘Speaker’ indicates which speaker from the acoustic experiment produced the token. ‘Task’ refers to the task from the acoustic experiment during which the token was produced. The identification task used every token here as a stimulus. The 2AFC task used stimuli created from each minimal pair in the table.

Speaker 6			Speaker 11			Speaker 12		
/d/	/t/	Task	/d/	/t/	Task	/d/	/t/	Task
buhKADing	buhKATing	wug	buhKEEDing	buhKEETing	min	puhTEHDing	puhTEHTing	min
buhKEHDing	buhKEHTing	min	buhKEEDing	buhKEETing	wug	buhKADing	buhKATing	min
buhPADing	buhPATing	min	buhKEHDing	buhKEHTing	min	buhKEEDing	buhKEETing	wug
buhPEEDing	buhPEETing	wug	buhTADing	buhTATing	min	buhPADing	buhPATing	min
buhTADing	buhTATing	min	buhTEHDing	buhTEHTing	min	buhTADing	buhTATing	min
buhTEEDing	buhTEETing	min	buhTEHDing	buhTEHTing	wug	buhTEEDing	buhTEETing	wug
buhTEHDing	buhTEHTing	wug	duhPADing	duhPATing	wug	duhKEEDing	duhKEETing	wug
duhKEEDing	duhKEETing	min	duhPEEDing	duhPEETing	wug	duhKEHDing	duhKEHTing	min
duhKEHDing	duhKEHTing	min	puhKADing	puhKATing	min	duhPEHDing	duhPEHTing	min
duhPADing	duhPATing	min	puhKEHDing	puhKEHTing	min	duhTADing	duhTATing	min
puhPEEDing	puhPEETing	wug	puhTADing	puhTATing	min	puhKEHDing	puhKEHTing	wug
puhPEHDing	puhPEHTing	min	puhTEEDing	puhTEETing	min	puhPEEDing	puhPEETing	min
puhTADing	puhTATing	wug	puhTEHDing	puhTEHTing	wug	puhTEEDing	puhTEETing	min
puhTEEDing	puhTEETing	min	tuhKADing	tuhKATing	min	puhTEHDing	puhTEHTing	wug
tuhKADing	tuhKATing	min	tuhPADing	tuhPATing	wug	tuhKADing	tuhKATing	min
tuhKEEDing	tuhKEETing	min	tuhPEEDing	tuhPEETing	min	tuhPADing	tuhPATing	wug
tuhKEHDing	tuhKEHTing	wug	tuhPEHDing	tuhPEHTing	min	tuhPEEDing	tuhPEETing	min
tuhPEHDing	tuhPEHTing	min	tuhTEEDing	tuhTEETing	min	tuhPEHDing	tuhPEHTing	min

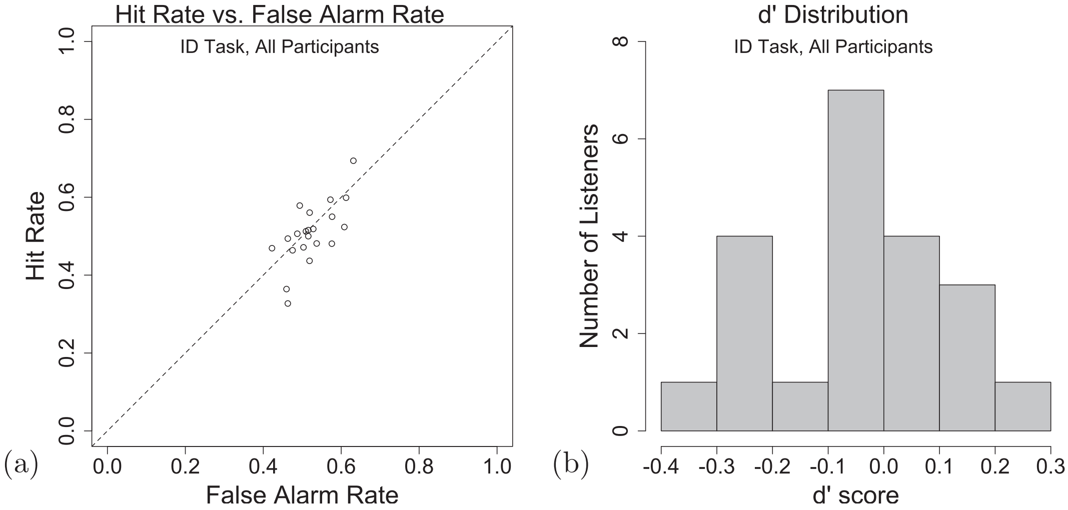


Fig. 3. Results of the identification task. (a) Hits vs. false alarms: ID task. (b) Histogram of d' scores: ID task.

3.4. Analysis

The results of this task are reported using d' as a measure of sensitivity (Macmillan and Creelman, 2005). Unlike the percent-correct measure, d' teases apart listeners' actual sensitivity from bias, because it takes both hit rate (how often a listener is correct) and false alarm rate (how often a listener mistakenly identifies a non-target as a target) into account. Using R, d' for this task was calculated as $d' = z(H) - z(F)$, where H is the hit rate, F is the false alarm rate, and the function z is the z-transform (Macmillan and Creelman, 2005). A d' score of zero indicates an inability to discriminate; as d' scores rise above zero, they indicate improving ability to discriminate. Given the strong bias in favor of /d/ found in Herd et al. (2010), d' is a more appropriate measure of sensitivity than percent-correct for this experiment.

3.5. Results

The mean d' score across all listeners for the identification task was $d' = -0.04$, which is not significantly different from zero (Wilcoxon test: $V = 76$, *n.s.*). In other words, listeners responded that they had heard a /d/ equally often in trials where /d/ was played as in trials where /t/ was played (/d/ responses to /d/ stimuli: 48.78%; /d/ responses to /t/ stimuli: 49.49%). Fig. 3(a) shows a plot of the hit rate vs. false alarm rate for participants in the identification task. Listeners cluster around the hit rate = false alarm rate line: they had as many hits (saying '/d/' when they heard a /d/) as they had false alarms (saying '/d/' when they had actually heard a /t/). In other words, saying that they had heard a /d/ had little relation to whether the trial's stimulus had actually contained a /d/. Given that d' was calculated as the difference between the z-transformed hit and false alarm rates, similar hit and false alarm rates will result in d' scores near zero. This is reflected in the frequency distribution of d' scores shown in Fig. 3(b), which center around zero.

3.6. Summary and discussion

Listeners in this experiment were unable to correctly categorize /d/-flaps and /t/-flaps.[8] Even though this experiment uses nonce word stimuli, the results comport with those seen in previous identification tasks of incompletely neutralized flaps using actual English words. Studies on the perception of incomplete neutralization have centered on identification, rather than discrimination (see Matsui, 2011:1342–1343 for discussion and a notable exception to this generalization).

[8] Since nonce words were used in this identification task, unlike the one presented in Herd et al. (2010), participants saw only 'd' and 't' on the screen, rather than whole words. In order to assuage the possible concern that the 'd or t' type task is easier than a task in which listeners see whole-word minimal pairs on the screen, an identical version of the task reported in this section was conducted with 12 new listeners, in which the nonce-word minimal pairs were presented on the screen instead of 'd' and 't'. The results are nearly identical to those of the task presented in the main body of this section: participants were unable to correctly categorize /d/-flaps and /t/-flaps ($d' = -0.02$), which is not significantly different from zero (Wilcoxon $V = 26$, *n.s.*).

The discrimination task presented in section 4, therefore, fills a critical lacuna by adducing evidence for the non-discriminability of an incompletely neutralized contrast.

A 2-alternative forced choice (2AFC) task (Experiment 3) is presented below. The goal of this task is to address the question of whether listeners' poor performance in identification studies of flapping is matched by similarly poor performance in discrimination tasks.

4. Experiment 3: 2AFC perception experiment

In order to test whether listeners' good performance in identification tasks is generalizable to other task types, a 2-Alternative Forced Choice (2AFC) task was run. In a 2AFC task listeners hear two stimuli per trial, which are always different from one another. Listeners are then asked to identify whether the target sound (in this case /d/ for half of the participants, and /t/ for the other half) came first or second. This experimental design provides listeners with more evidence upon which to base their decision than an identification task—listeners are able to compare two stimuli on each trial in a 2AFC task, while they have evidence from only one stimulus per trial in an identification experiment (Macmillan and Creelman, 2005:167–170).

4.1. Participants and equipment

24 undergraduates participated in this experiment, none of whom participated in the previous experiments. All participants were native speakers of English. 75% ($n = 18$) of the participants were born in New Jersey, and 95.8% ($n = 23$) reported being raised primarily in New Jersey. As with the previous perception task, this experiment took place at the Rutgers Phonetics Laboratory, with stimuli displayed and responses recorded by SuperLab 4.5 (Cedrus Corporation, 2010) through Sennheiser HD 280 Professional headphones.

4.2. Stimuli

All tokens were taken from the same set of stimuli used in the Identification task.

4.3. Procedure

On each trial, listeners heard two tokens—members of a minimal pair—separated by 250 ms. Listeners were instructed to pay attention to the sound immediately preceding the '-ing' in each word. Half of the listeners were asked whether the /d/-member of the pair came first or second. The other half of the listeners were asked whether the /t/-member of the pair came first or second. Participants in the 'find /d/' variation were instructed to push one of two buttons corresponding to the two words they saw on the screen, indicating whether the first word or the second word had the /d/. Participants in the 'find /t/' variation indicated whether the first word or the second word had the /t/. For example, if a listener who was told to 'find /d/' heard 'buhKEED-ing - buhKEET-ing', they would press the button corresponding to the first word, since it is the one that contains a /d/ immediately preceding '-ing'.

Each of the three blocks, separated by speaker, consisted of 36 randomized trials (half /d/ and half /t/). Block order was balanced (Latin Square) across all listeners. As with the Identification experiment, feedback was provided on each trial, and the next trial was presented if listeners did not respond within 1500 ms.

4.4. Analysis

As with the Identification experiment, d' was used as a measure of sensitivity, and was computed in R. Because 2-Alternative Forced Choice tasks such as this one provide listeners with more evidence upon which to base their decisions than identification tasks, d' is calculated here as $d' = (z(H) - z(F))/\sqrt{2}$ (Macmillan and Creelman, 2005:167–170).

4.5. Results

The mean d' score across all listeners for the 2AFC task was $d' = -0.016$, which is not significantly different from zero (Wilcoxon test: $V = 138$, *n.s.*). In other words, listeners indicated that the target sound (/d/ or /t/, depending on the listener) had been in the first word of a given trial just as often when the target sound had indeed been in the first word as when the target sound had actually been in the second word (49.48% correct). Fig. 4(a) shows a plot of the hit rate vs. false alarm rate for participants in the 2AFC task. Fig. 4(b) shows the frequency distribution of d' scores, again centering around zero.

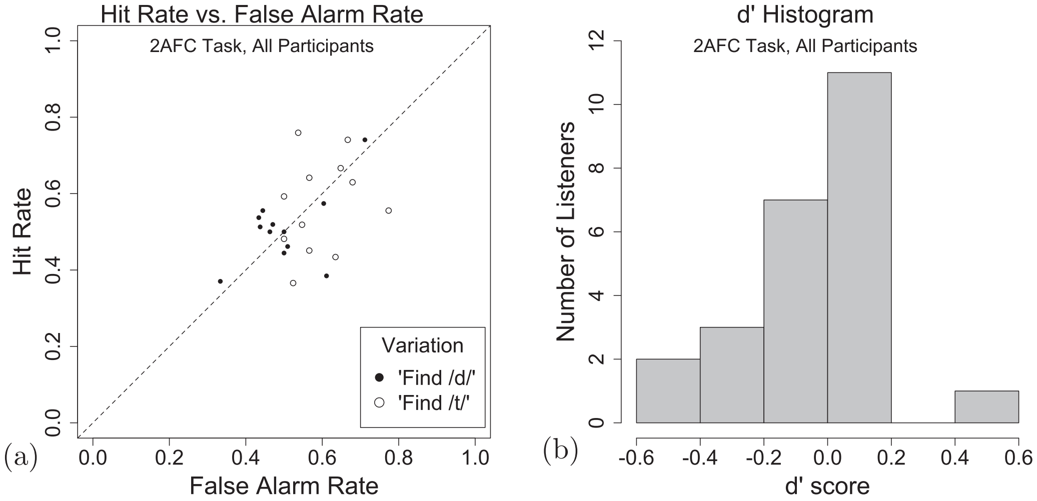


Fig. 4. Results of the 2AFC task. (a) Hits vs. false alarms: 2AFC task. (b) Histogram of *d*′ scores: 2AFC task.

4.6. Discussion

Listeners in the 2AFC task were unable to accurately distinguish between /d/-flaps and /t/-flaps. This result, in combination with the Identification task, suggests that listeners' inability to distinguish /d/-flaps from /t/-flaps is relatively robust across task type.

5. Conclusions

The results of the acoustic experiment suggest that flapping in American English is, indeed, incompletely neutralizing. Vowels preceding /d/-flaps were, on average, 5.69 ms longer than vowels preceding /t/-flaps—a small, yet measurable trace of the underlying voicing distinction. This result holds in both a minimal pair reading task and a paradigm completion 'wug' task, suggesting that this and previous findings of incomplete neutralization in American English are not due solely to the extragrammatical or task-related factors manipulated in these studies.

While there is a significant difference between vowels preceding /d/-flaps and /t/-flaps on the surface, listeners are unable to distinguish the two across different task types. This result was shown to hold for nonce words and with the *d*′ measure of sensitivity, which takes listener bias into account. In the Identification task, listeners showed an inability to distinguish between /d/- and /t/- flaps—a result in line with the recent study from Herd et al. (2010). Similarly, listeners were unable to distinguish between /d/- and /t/-flaps in the 2AFC task. This result is significant as previous studies on the perceptibility of this distinction have focused solely on identification tasks.

Taken together, these results suggest that listeners' poor performance on identification tasks of /d/- and /t/-flaps is robust across task type. Given that some other reported cases of incomplete neutralization show larger surface distinctions which may be identifiable (e.g., Port and O'Dell (1985) report a 15 ms vowel duration difference in German final devoicing, and report that listeners can identify better than chance), it is conceivable that there are two types of incomplete neutralization: (a) one in which identification and discrimination are not possible (like flapping), and (b) one in which identification and discrimination are possible, and where some degree of categorical knowledge is expected (like German devoicing).[9]

Further, incomplete neutralization of the sort described in this paper shows that the phonetic module must have access to some degree of information about underlying phonological representations. In order for the phonetics to implement vowel lengthening on the basis of an underlying voicing distinction, for example, it must be privy to underlying voicing status. This information might, however, be passed to the phonetics in an indirect or roundabout way (see, e.g., van Oostendorp's (2008) analysis of incomplete neutralization in final devoicing, on which more below).

A traditional analysis of the distinction between pre-/d/-flaps and pre-/t/-flaps under the classical modular feedforward architecture is that the process that lengthens vowels before voiced segments (or, as suggested by the term *pre-fortis clipping*, the process that shortens vowels before voiceless segments) is crucially ordered with respect to flapping: flapping must come last (i.e., flapping counterbleeds the vowel-duration process) (Anderson, 1975; Fox and Terbeek,

[9] These two types of incomplete neutralization align approximately with Dinnsen's (1985:274) Type B and Type C neutralizations, respectively.

1977; Zue and Laferriere, 1979). The challenge for this view is that the vowel duration rule is gradient, yet must precede the phonological flapping rule (Fox and Terbeek, 1977:33; Zue and Laferriere, 1979:1047). Anderson (1975:54) describes the situation thus: ''The flapping rule itself is still phonological. . . The fact that it must nevertheless follow the phonetic [vowel] length rule provides an instance of the intermixture of phonological and phonetic rules.'' Such 'intermixture', however, is not permitted in a modular feedforward architecture.

More recently, Bermúdez-Otero (2004) has developed this counterbleeding proposal in a manner more amenable to the classical model. He argues that pre-fortis clipping is, in fact, a categorical phonological process, and that its domain excludes class-two suffixes: it applies at the stem level, and therefore underapplies when the conditioning voiceless segment belongs to a word-level suffix. Since flapping occurs across word boundaries (e.g., 'sit in the park' → si[ɾ] in the park', Kaisse and Shaw, 1985), it must be a post-lexical process (Kiparsky, 1982a,b, 1985; Mohanan, 1982; Kaisse and Shaw, 1985). Since post-lexical processes occur after stem-level ones, Bermúdez-Otero's (2004) proposal correctly orders flapping after the vowel-duration rule. Crucially, Bermúdez-Otero (2004:section 21) argues that a 'gradient, nonneutralizing phonetic process of durational adjustment' reduces the duration of the categorically non-clipped vowels which precede /d/-flaps. Because this process gradiently applies to non-clipped vowels before /d/-flaps, their duration comes to approximate (but not match) that of the clipped vowels preceding /t/-flaps, as shown in the derivation in (6) below (modified from Bermúdez-Otero, 2004:section 21).

(6)	*utter*	*udder*
UR	/ʌtəɹ/	/ʌdər/
Clipping	ʌ̆təɹ	—
Flapping	/ʌ̆ɾəɹ/	/ʌɾər/
. . .	. . .	. . .
SR	[ʌ̆ɾə˞]	[ʌɾə˞]
Phonetic vowel duration adjustment rule	—	Gradiently shortens [ʌ]
Phonetic implementation	[ʌ] in *udder* has shortened, but is not as short as [ʌ̆] in *utter*	

Another, possibly complementary, approach capitalizes on the variability of incomplete neutralization within a population. Since (at least some) speakers do produce incompletely neutralized /t/- and /d/-flaps, it follows that they have somehow acquired this distinction—in spite of their poor performance on perception tasks. Those who produce the distinction must register the distinction at some level, but clearly do not use it for categorization or lexical recognition. Following Kirby (2011a,b), this state of affairs might come about by virtue of population-level variation in the durational distinction maintained in pre-flap vowels. Under this view, the covert contrast is acquired thanks, on the one hand, to those speakers who maintain a relatively large duration difference (i.e., these speakers alert the learner to the link between pre-flap vowel duration and the underlying laryngeal contrast), but on the other hand speakers with no real distinction render vowel duration an unreliable cue for lexical recognition.

This approach is consistent with the data across the production and perception experiments reported here. Recalling the results of the production experiment (section 2.4.2, Table 2), there was a high degree of variation among speakers: Speaker 10 had, on average, a distinction of 1.04 ms between vowels preceding /d/-flaps and /t/-flaps, while Speaker 6 had an average distinction of 15.28 ms. This inter-speaker variability is what we might expect when learners have inconsistent input. The presence of reversals, too, supports the population-variation hypothesis: a learner who hears longer vowels before /d/-flaps from some speakers, but longer vowels before /t/-flaps from others, should not learn to use the duration distinction as a cue for lexical recognition. Learners in this situation are predicted to perform poorly on perception tasks, as was the case in those presented here.

Further support for the population-variation hypothesis comes from priming experiments reported in Luce et al. (1999) and McLennan et al. (2003). In these studies, response time to /t, d/ was found to be facilitated by prior exposure to flapped pronunciations of these segments (even in non-alternating monomorphemic pairs such as *udder* and *utter*: Luce et al., 1999:1890). This suggests that flaps activate laryngeally specified underlying representations for /t, d/, at least at a level that can impact priming. Under the population-variation hypothesis, (at least some) speakers are also predicted to have some limited categorical knowledge that does not reach the level of a full laryngeal contrast.

5.1. Alternative proposals

The first category of alternative proposals argues that the phonological module has a higher degree of control over the variation of phonetic implementation of contrasts than is normally assumed (Kingston and Diehl, 1994; Yu, 2011). On this view, subphonemic distinctions are ''qualitatively not different from those observed between allophones appearing in different phonetic contexts'' (Yu, 2011:311). A strength of this model is that speakers have more active control over the fine details of the realization of a given feature (e.g., in utterance-initial position, [+voice] may be realized as lowered F1, or lowered F0, or weaker burst, or some combination of these). The implication for incomplete neutralization is that speakers simply apply the correct 'allophone' in neutralizing contexts. It seems, however, that under such a model the perception results presented in this paper are unexpected. If speakers map phonological features to an array of possible phonetic categories (as opposed to gradient variation along the phonetic dimension), listeners might be expected to mirror this behavior.

The second class of proposals question representation of and membership in phonological categories. These models generate incomplete neutralization by presenting the phonetics with distinct representations of the neutralized category (e.g., in German, devoiced and underlyingly voiceless segments are represented differently). Under the Turbidity Theory (Goldrick, 2001) model developed by van Oostendorp (2008), for example, underlyingly voiceless segments have no association to a [voice] feature, while devoiced segments have a 'projection' relation to a [voice] feature. This distinction is accessible to the phonetics, which articulates the two types of segment differently. The results presented here are largely compatible with this view: speakers maintain a distinction between the target segments for the purposes of production, but since this distinction consists only of the presence or absence of a projection relation (and doesn't include, in this theory's terminology, a 'pronunciation' relation), it might be argued that listeners cannot access this information for the purposes of categorization or lexical recognition. A criticism of this class of model, however, is that for the most part it is not clear what requires the phonetic module to realize the two distinct representations in precisely the manner found in incomplete neutralization. In the case of flapping, it is only a stipulation of the theory that causes the phonetic module to interpret the distinct representations of /d/-flaps and /t/-flaps as a small duration distinction in the preceding vowel.

Let us now turn to arguments that reject the relevance of formal phonology for neutralization (Dinnsen, 1985; Port, 1996; Port and Leary, 2005). These arguments are not explanations of incomplete neutralization *per se*, but rather claims about the nature of phonology made on the basis of incomplete neutralization. The thrust of such proposals is that the existence of incomplete neutralization, and the claimed non-existence of complete neutralization, mean that formal, categorical phonology should give way to a model more prominently driven by gradient units of analysis. It is not clear, however, how such models would analyze the results of the priming experiments described above, in which response time for /t, d/ stimuli was facilitated by prior exposure to flapped pronunciations (Luce et al., 1999; McLennan et al., 2003). These results show that flaps activate the underlying, distinct categories /t/ and /d/, suggesting that such categories do exist, at least at some level.

The present experiments also shed light on the claim that incomplete neutralization is largely an extragrammatical affair (Fourakis and Port, 1986; Manaster Ramer, 1996a,b; van Rooy et al., 2003; Warner et al., 2006). The two acoustic tasks—one highlighting the contrast, and one masking it—showed similar results. It is quite possibly the case that extragrammatical factors not controlled for play a role in the distinction between vowels preceding /d/-flaps and /t/-flaps. For example, even though participants did not see an orthographic representation of the target '-ing' verb, they did see a representation of the verb's stem. However, the tasks presented here represent a first foray into examining potential effects of task with respect to the incomplete neutralization of flapping in particular.

Acknowledgments

I thank three anonymous reviewers for their helpful comments, especially regarding the implementation of incomplete neutralization in modular feedforward and exemplar models, alternative analyses of the results, and discussion of possible strategies used by experiment participants. Thanks also to Shigeto Kawahara, Ricardo Bermúdez-Otero, and the audiences of NELS 42, MFM 18, and LabPhon 12 for useful discussion.

Appendix

See Table 4.

Table 4
Summary of results from the acoustic experiment. *t*-values are from the linear mixed model described in section 2.3.3. *p*-values are calculated using the `pvals.fnc` function of R's `languageR` package (Baayen, 2009), which uses the Markov Chain Monte Carlo method.

Main effect of underlying voicing status					
Fixed factor	Mean /d/	Mean /t/	Mean difference	*t*	*p*
Pre-flap vowel duration (ms)	112.23	106.54	5.69	3.77	<0.001
Flap closure duration (ms)	28.54	28.16	0.38	0.04	n.s.
Percent closure voicing (%)	98.14	97.40	0.74	−0.68	n.s.
F0 slope at flap onset (Hz/ms)	0.40	0.43	−0.03	−0.45	n.s.

Table 4 (*Continued*)

Main effect of underlying voicing status

Fixed factor	Mean /d/	Mean /t/	Mean difference	*t*	*p*
F1 slope at flap onset (Hz/ms)	4.64	4.91	−0.27	−0.80	n.s.
F0 slope at flap offset (Hz/ms)	−0.06	0.07	−0.13	−1.29	n.s.
F1 slope at flap offset (Hz/ms)	−0.75	0.21	−0.96	−0.79	n.s.

Main effect of task

Fixed factor	*t*	*p*
Pre-flap vowel duration	0.59	n.s.
Flap closure duration	0.03	n.s.
Percent closure voicing	−2.79	<0.01
F0 slope at flap onset	−0.70	n.s.
F1 slope at flap onset	1.63	n.s.
F0 slope at flap offset	0.18	n.s.
F1 slope at flap offset	0.21	n.s.

Interaction: task × underlying voicing status

Fixed factor	*t*	*p*
Pre-flap vowel duration	0.82	n.s.
Flap closure duration	−0.31	n.s.
Percent closure voicing	1.59	n.s.
F0 slope at flap onset	−0.26	n.s.
F1 slope at flap onset	−1.21	n.s.
F0 slope at flap offset	−0.23	n.s.
F1 slope at flap offset	0.12	n.s.

References

Anderson, S.R., 1975. On the interaction of phonological rules of various types. J. Linguist. 11, 39–62.

Audacity Team, 2008. Audacity v. 1.2.6. Computer program. http://audacity.sourceforge.net.

Baayen, H., 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge University Press, Cambridge.

Baayen, H., 2009. languager: data sets and functions with ‘‘analyzing linguistic data: a practical introduction to statistics’’. R package. http://CRAN.R-project.org/package=languageR.

Bates, D., Maechler, M., 2009. lme4: linear mixed-effects models using s4 classes. R package. http://CRAN.R-project.org/package=lme4.

Berko, J., 1958. The child’s learning of English morphology. Word 14, 150–177.

Bermúdez-Otero, R., 2004. Raising and flapping in Canadian English: grammar and acquisition. In: Paper Presented at CASTL Colloquium, University of Tromsø. www.bermudez-otero.com/tromsoe.pdf.

Bermúdez-Otero, R., 2007. Diachronic phonology. In: de Lacy, P. (Ed.), The Cambridge Handbook of Phonology. Cambridge University Press, Cambridge, pp. 497–517.

Bloomfield, L., 1933/1984. Language. University of Chicago Press, Chicago.

Boersma, P., Weenink, D., 2009. Praat: doing phonetics by computer. Computer program. http://www.praat.org.

Bybee, J., 2001. Phonology and Language Use. Cambridge University Press, Cambridge.

Cedrus Corporation, 2010. Superlab v. 4.5. Computer program.

Charles-Luce, J., 1997. Cognitive factors involved in preserving a phonemic contrast. Lang. Speech 40, 229–248.

Charles-Luce, J., Dressler, K.M., 1999. The effects of semantic predictability in non-pathological older adults’ production of a phonemic contrast. Clin. Linguist. Phon. 13, 199–217.

Charles-Luce, J., Dressler, K.M., Ragonese, E., 1999. Effects of semantic predictability on children’s preservation of a phonemic voice contrast. J. Child Lang. 26, 505–530.

Chen, M., 1970. Vowel length variation as a function of the voicing of the consonant environment. Phonetica 22 .

Dinnsen, D., 1985. A re-examination of phonological neutralization. J. Linguist. 21, 265–279.

Dinnsen, D., Charles-Luce, J., 1984. Phonological neutralization, phonetic implementation and individual differences. J. Phon. 12, 49–60.

Dinnsen, D.A., Garcia-Zamor, M., 1971. The three degrees of vowel length in German. Pap. Linguist. 4, 111–126.

Dmitrieva, O., Jongman, A., Sereno, J., 2010. Phonological neutralization by native and non-native speakers: the case of Russian final devoicing. J. Phon. 38, 483–492.

Ernestus, M., 2000. Voicing Assimilation and Segment Reduction in Casual Dutch: A Corpus-based Study of the Phonology–Phonetics Interface. LOT, Utrecht, The Netherlands.

Ernestus, M., Baayen, H., 2006. The functionality of incomplete neutralization in Dutch: the case of past-tense formation. In: Goldstein, L.M., Whalen, D., Best, C.T. (Eds.), Laboratory Phonology, vol. 8. Mouton de Gruyter, Berlin, pp. 29–51.

Ernestus, M., Baayen, H., 2007. Intraparadigmatic effects on the perception of voice. In: van de Weijer, J., van der Torre, E. (Eds.), Voicing in Dutch: (De)voicing–Phonology, Phonetics, and Psycholinguistics. John Benjamins Publishing Company, Amsterdam, pp. 153–173.

Fisher, W.M., Hirsh, I.J., 1976. Intervocalic flapping in English. In: Papers from the Twelfth Regional Meeting of the Chicago Linguistic Society, Chicago Linguistic Society.

Fourakis, M., Iverson, G., 1984. On the 'incomplete neutralization' of German final obstruents. Phonetica 41, 140–149.

Fourakis, M., Port, R., 1986. Stop epenthesis in English. J. Phon. 14, 197–221.

Fox, R.A., Terbeek, D., 1977. Dental flaps, vowel duration, and rule ordering in American English. J. Phon. 5, 27–34.

Fujisaki, H., Nakamura, K., Imoto, T., 1975. Auditory perception of duration of speech and non-speech stimuli. In: Fant, G., Tatham, M.A.A. (Eds.), Auditory Analysis and Perception of Speech, vol. 197–219. Academic, London.

Gerfen, C., 2002. Andalusian codas. Probus 14, 247–277.

Goldrick, M., 2001. Turbid output representations and the unity of opacity. In: Hirotani, M., Coetzee, A., Hall, N., Kim, J.-Y. (Eds.), Proceedings of the North East Linguistics Society 30. GLSA Publications, Amherst, MA, pp. 231–245.

Gussenhoven, C., 1986. English plosive allophones and ambisyllabicity. Gramma 10, 119–141.

Hansson, G.Ó., 2011. Diachronic explanations of sound patterns. In: Goldsmith, J., Riggle, J., Yu, A.C.L. (Eds.), The Handbook of Phonological Theory. 2nd ed. Blackwell, Oxford, pp. 319–347.

Herd, W., Jongman, A., Sereno, J., 2010. An acoustic and perceptual analysis of /t/ and /d/ flaps in American English. J. Phon. 38, 504–516.

Hombert, J.-M., Ohala, J.J., Ewan, W.G., 1979. Phonetic explanations for the development of tones. Language 55, 37–58.

Huff, C.T., 1980. Voicing and flap neutralization in New York City English. Res. Phon. 1, 233–256.

Hyman, L., 1975. Phonology: Theory and Analysis. Holt, Rinehart and Winston, New York.

Inouye, S., 1989. The flap as contour segment. UCLA Working Pap. Linguist. 72, 39–81.

Inozuka, E., 1991. The realization of German neutralized word-final plosives /g,k/: an acoustic analysis. Sophia Linguist. 30, 119–134.

Jakobson, R., Fant, G., Halle, M., 1975. Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates. MIT Press, Cambridge.

Jassem, W., Richter, L., 1989. Neutralization of voicing in Polish obstruents. J. Phon. 17, 317–325.

Johnson, K., 1997. The auditory/perceptual basis for speech segmentation. OSU Working Pap. Linguist. 50, 101–113.

Joos, M., 1942. A phonological dilemma in Canadian English. Language 18, 141–144.

Kahn, D., 1980. Syllable-based Generalizations in English Phonology. Garland, New York.

Kaisse, E.M., Shaw, P.A., 1985. On the theory of lexical phonology. Phonol. Yearbook 2, 1–30.

Keating, P., 1988. The phonology–phonetics interface. In: Newymeyer, F. (Ed.), Linguistics: The Cambridge Survey, vol. 1. Cambridge University Press, Cambridge, pp. 281–302.

Kharlamov, V., 2012. Incomplete neutralization and task effects in experimentally-elicited speech: evidence from the production and perception of word-final devoicing in Russian (Ph.D. Thesis). University of Ottawa, Ottawa, Canada.

Kharlamov, V., 2014. Incomplete neutralization of the voicing contrast in word-final obstruents in Russian: phonological, lexical, and methodological influences. J. Phon. 43, 47–56.

Kingston, J., Diehl, R.L., 1994. Phonetic knowledge. Language 3, 419–454.

Kiparsky, P., 1979. Metrical structure assignment is cyclic. Linguist. Inq. 10, 421–441.

Kiparsky, P., 1982a. From cyclic phonology to lexical phonology. In: van der Hulst, H., Smith, N. (Eds.), The Structure of Phonological Representations, vol. 1. Foris, Dordrecht, pp. 131–175.

Kiparsky, P., 1982b. Lexical phonology and morphology. In: Lee, I.-H. (Ed.), Linguistics in the Morning Calm. Hanshin, Seoul, pp. 3–91.

Kiparsky, P., 1985. Some consequences of lexical phonology. Phonol. Yearbook 2, 85–138.

Kirby, J.P., 2011a. Modeling the acquisition of covert contrast. In: Paper Presented at the 19th Manchester Phonology Meeting. http://www.lel.ed.ac.uk/mfm/19mfm-abbk.pdf.

Kirby, J.P., 2011b. Modeling the acquisition of covert contrast. In: Lee, W.-S., Zee, E. (Eds.), Proceedings of the International Congress of Phonetic Sciences XVII.

Kleber, F., John, T., Harrington, J., 2010. The implications for speech perception of incomplete neutralization of final devoicing in German. J. Phon. 38, 185–196.

Kluender, K.R., Diehl, R.L., Wright, B.A., 1988. Vowel-length differences before voiced and voiceless consonants: an auditory explanation. J. Phon. 16, 153–169.

Kopkallı, H., 1993. A phonetic and phonological analysis of final devoicing in Turkish (Ph.D. Thesis). University of Michigan.

Kwong, K., Stevens, K.N., 1999. On the voiced-voiceless distinction for *writer/rider*. In: Speech Communication Group Working Papers, vol. 11, Research Laboratory of Electronics at MIT, pp. 1–20.

Lehiste, I., 1970. Suprasegmentals. MIT Press, Cambridge.

Lisker, L., 1986. ''voicing'' in English: a catalog of acoustic features signaling /b/ versus /p/ in trochees. Lang. Speech 29, 3–11.

Luce, P.A., Charles-Luce, J., McLennan, C.T., 1999. Representational specificity of lexical form in the production and perception of spoken words. In: Proceedings of the XIV International Congress of Phonetic Sciences (ICPhS99).

Macmillan, N.A., Creelman, C.D., 2005. Detection Theory: A User's Guide, 2nd ed. Lawrence Erlbaum Associates Inc., Mahwah, NJ.

Malécot, A., Lloyd, P.M., 1968. The /t/:/d/ distinction in American alveolar flaps. Lingua 19, 264–272.

Manaster Ramer, A., 1996a. A letter from an incompletely neutral phonologist. J. Phon. 24, 477–489.

Manaster Ramer, A., 1996b. Report on Alexis' dreams—bad as well as good. J. Phon. 24, 513–519.

Mascaro, J., 1987. Underlying voicing recoverability of finally devoiced obstruents in Catalan. J. Phon. 15, 183–186.

Matsui, M., 2011. The identifiability and discriminability between incompletely neutralized sounds: evidence from Russian. In: Proceedings of the International Congress of Phonetic Sciences XVII.

McLennan, C.T., Luce, P.A., Charles-Luce, J., 2003. Representation of lexical form. J. Exp. Psychol. Learn. Mem. Cogn. 29, 539–553.

Mitleb, F.M., 1981a. Segmental and non-segmental structure in phonetics: evidence from foreign accent (Ph.D. Thesis). Indiana University, Bloomington.

Mitleb, F.M., 1981b. Temporal correlates of 'voicing' and its neutralization in German. Res. Phon. 2, 173–192.

Mohanan, K., 1982. Lexical phonology (Ph.D. Thesis). Massachusetts Institute of Technology.

Moreton, E., 2004. Realization of the English postvocalic [voice] contrast in F_1 and F_2. J. Phon. 32, 1–33.

Nooteboom, S.G., Doodeman, G.J.N., 1980. Production and perception of vowel length in spoken sentences. J. Acoust. Soc. Am. 67, 276–287.

van Oostendorp, M., 2008. Incomplete devoicing in formal phonology. Lingua 118, 1362–1374.

Oswald Jr., V.A., 1943. ''voiced *t*'': a misnomer. Am. Speech 18, 18–25.

Patterson, D.J., Connine, C.M., 2001. Variant frequency in flap production: a corpus analysis of variant frequency in American English flap production. Phonetica 58, 254–275.

Peterson, G.E., Lehiste, I., 1960. Duration of syllable nuclei in English. J. Acoust. Soc. Am. 32, 693–703.

Pierrehumbert, J., 2002. Word-specific phonetics. In: Gussenhoven, C., Warner, N. (Eds.), Laboratory Phonology, vol. 7. Mouton de Gruyter, Berlin.

Pierrehumbert, J.B., 2001. Exemplar dynamics: word frequency, lenition, and contrast. In: Bybee, J., Hopper, P. (Eds.), Frequency Effects and the Emergence of Linguistic Structure. John Benjamins, Amsterdam/Philadelphia, pp. 137–157.

Piroth, H.G., Janker, P.M., 2004. Speaker-dependent differences in voicing and devoicing of German obstruents. J. Phon. 32, 81–109.

Port, R., 1976. The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words (Ph.D. Thesis). University of Connecticut.

Port, R., 1996. The discreteness of phonetic elements and formal linguistics: response to A. Manaster Ramer. J. Phon. 24 .

Port, R., Crawford, P., 1989. Incomplete neutralization and pragmatics in German. J. Phon. 17, 257–282.

Port, R., Leary, A., 2005. Against formal phonology. Language 81, 927–964.

Port, R., Mitleb, F., O'Dell, M., 1981. Neutralization of obstruent voicing in German is incomplete. J. Acoust. Soc. Am. 70 (Suppl. 1), 13.

Port, R., O'Dell, M., 1985. Neutralization and syllable-final voicing in German. J. Phon. 13, 455–471.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

van Rooy, B., Wissing, D., Paschall, D.D., 2003. Demystifying incomplete neutralization during final devoicing. South. Afr. Linguist. Appl. Lang. Stud. 21, 49–66.

Röttger, T.B., Winter, B., Grawunder, S., 2011. The robustness of incomplete neutralization in German. In: Lee, W.-S., Zee, E. (Eds.), Proceedings of the 17th International Congress of Phonetic Science.

Röttger, T.B., Winter, B., Grawunder, S., Kirby, J.P., Grice, M., 2014. Assessing incomplete neutralization of final devoicing in German. J. Phon. 43, 11–25.

Selkirk, E., 1982. The syllable. In: van der Hulst, H., Smith, N. (Eds.), The Structure of Phonological Representations, vol. 2. Foris, Dordrecht.

Sharf, D.J., 1960. Distinctiveness of 'voiced T' words. Am. Speech 35, 105–109.

Slowiaczek, L.M., Dinnsen, D., 1985. On the neutralizing status of Polish word-final devoicing. J. Phon. 13, 325–341.

Slowiaczek, L.M., Szymanska, H., 1989. Perception of word-final devoicing in Polish. J. Phon. 17, 205–212.

Smith, B.L., Hayes-Harb, R., Bruss, M., Harker, A., 2009. Production and perception of voicing and devoicing in similar German and English word pairs by native speakers of German. J. Phon. 37, 257–275.

Sóskuthy, M., 2011. Why phonologists should care about exemplar theory. In: Paper Presented at the 19th Manchester Phonology Meeting http://www.lel.ed.ac.uk/~s0954634/soskuthy11_mfm19.zip

Steriade, D., 1997. Phonetics in Phonology: The Case of Laryngeal Neutralization. Ms. University of California, Los Angeles.

Taylor, D.Q., 1975. The inadequacy of bipolarity and distinctive features: the German ''voiced/voiceless'' consonants. In: Reich, P.A. (Ed.), The Second LACUS Forum. Hornbeam Press, Inc., Columbia, SC, pp. 107–119.

Thomas, E.R., 2000. Spectral differences in /ai/ offsets conditioned by voicing of the following consonant. J. Phon. 28, 1–25.

Trubetzkoy, N.S., 1939/1969. Grundzüge der Phonologie [Principles of Phonology]. Vandenhoeck and Ruprecht [Translated by Christiane A.M. Baltaxe 1969, University of California Press], Güttingen.

Turk, A., 1992. The American English flapping rule and the effect of stress on stop consonant durations. Cornell Working Pap. Phon. 7, 103–133.

Warner, N., Fountain, A., Tucker, B.V., 2009. Cues to the perception of reduced flaps. J. Acoust. Soc. Am. 125, 3317–3327.

Warner, N., Good, E., Jongman, A., Sereno, J., 2006. Orthographic vs. morphological incomplete neutralization effects. J. Phon. 34, 285–293.

Warner, N., Jongman, A., Sereno, J., Kemps, R., 2004. Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. J. Phon. 32, 251–276.

Warner, N., Tucker, B.V., 2011. Phonetic variability of stops and flaps in spontaneous and careful speech. J. Acoust. Soc. Am. 130, 1606–1617.

Yu, A.C.L., 2007. Understanding near mergers: the case of morphological tone in Cantonese. Phonology 24, 187–214.

Yu, A.C.L., 2011. Contrast reduction. In: Goldsmith, J., Riggle, J., Yu, A.C.L. (Eds.), The Handbook of Phonological Theory. 2nd ed. Blackwell, Oxford. (Chapter 9), pp. 291–318.

Zue, V.W., Laferriere, M., 1979. Acoustic study of medial /t, d/ in American English. J. Acoust. Soc. Am. 66, 1039–1050.