Investigating the Incomplete Neutralization of Flaps in North American English

Michael Colley

Rice University, Houston, TX, USA

Introduction

This study examines claims of incomplete neutralization of flaps in North American English, in words like "writing" vs. "riding" (see Fox and Terbeek 1977). Specifically, it tests whether listeners are able to use differences in vowel durations before flaps to determine the underlying voicing of the flap.

Neutralization is a process by which normally contrasting sounds in a language become indistinguishable in certain phonological contexts. This occurs in the devoicing of final obstruents in many languages, for example in Dutch, *meet* "measures (sg.)" and *meed* "avoided (sg.)" are both pronounced [meˑt], although the same stems are distinguishable in other contexts, such as *meten* [meˑtən] "to measure" and *meden* [meˑdən] "avoided (pl.)." Some studies have suggested that not all reported cases of neutralization are complete, that the supposedly neutralized sounds may retain some characteristics of their non-neutralized counterparts and thus retain some degree of contrast. This has been reported in the case of final obstruent devoicing in Dutch [Warner et al. 2004], Polish [Jassem and Richter 1989; Slowiaczek and Dinnsen 1985], Russian [Pye 1986], and Afrikaans [Van Rooy et al. 2003]. All of these are cases that have been assumed to be complete neutralizations, but the authors present evidence that vowels before phonologically voiced final obstruents are longer than before phonologically voiceless ones; in some cases, the final obstruents themselves are shorter when phonologically voiced.

Although neutralization due to devoicing of final obstruents is not a common feature of English, a similar type of neutralization occurs in most varieties of North American English in the flapping of apical stops, as in *writing* and *riding,* both generally [ɹaɪɾiŋ]. In non-neutralized forms, vowel length is an important part of the contrast between such words, thus *write* [ɹaɪt] vs. *ride* [ɹaɪːd]. Some evidence has suggested that a similar distinction in vowel length is also made before flaps, meaning that vowels before /d/-flaps are longer than vowels before /t/-flaps. Fox and Terbeek [1977] examined the effect of underlying voicing of flaps on the duration of both the preceding vowel and the flap itself in word-list data. They found a significant

difference in vowel duration, but not flap duration. Similarly, Zue and Laferriere [1979] found a significant difference between the pre-flap vowel duration in word list data, but not for the flap duration itself. On average, vowels were 9 ms longer before [d] flaps than before [t] flaps. Patterson and Connine [2001] found similar results using data from the SWITCHBORAD corpus, consisting of two-way telephone conversations on prompted topics. Only minimal pairs, such as *latter* and *ladder* were included. All speakers in the corpus were included, thus not restricting the analysis to any particular regional variety of American English. Vowels before /d/-flaps were 16 ms longer on average than before /t/-flaps (significant at $p < 0.05$). They did not compare the difference in the durations of the flaps themselves. These studies suggest that flap neutralization in American English is a case of incomplete neutralization, rather than one of complete neutralization, since the words may retain their distinctiveness.

Perception studies have shown that it is possible for listeners to distinguish between pairs of words in cases of incomplete neutralization. Port and Crawford [1989] showed that German listeners performed above chance when asked to identify words that were suspected of being involved in incomplete neutralization. However, the speech data used for the study were from speakers who were consciously trying to make a contrast between the neutralized forms. Although the authors also recorded utterances under more natural conditions, they did not use these in the perception test. Furthermore, it is not clear from the study what phonetic contrast in the data listeners used to make the distinctions, whether for example it was vowel duration, consonant duration, or a combination of several factors.

The idea of incomplete neutralizations has not been universally accepted. Manaster Ramer [1996], for example, suggested that the claims of incomplete neutralization are due to an influence from orthography, and are thus examples of hypercorrection. In almost all of the cases where incomplete neutralization has been claimed, the supposed distinction is maintained orthographically, as in Dutch *meet* "measures (sg.)" and *meed* "avoided (sg.)". Manaster Ramer suggests that a more reasonable explanation to incomplete neutralization is that, for literate speakers, speech production draws from both the phonological representations of the words as well as spelling.[1]

---

[1] The argument that orthography may influence pronunciation, however, cannot apply to Patterson and Connine [2001], whose data come from prompted conversations, and not from a reading list like most other studies on incomplete neutralization.

To avoid any possible influence from orthography or even orthographic representations, an experiment was devised using pseudo-words, which were presented as auditory stimuli to the participants. The responses were recorded orally, thus avoiding any exposure to written language during the experiment.

Since previous studies have found a significant difference in vowel duration before flaps, but not for flap duration [Fox and Terbeek 1977, Zue and Laferriere 1979], this study will maintain a fixed duration for the flaps, but systemically vary the duration of the preceding vowels.

Material

The stimuli consisted of twenty phrases of the type "he is ___-ing", with the blank representing a pseudo-word ending in an alveolar stop, such as "he is chotting." The initial portion varied among "he is," "she is " and "they are."

There were ten pairs of pseudo-words, with each pair having identical codas, such as "chot" and "zot." In addition, thirty pseudo-words ending in non-alveolar stops were used as filler. The stimuli were grouped into ten blocks with five words each: two target words containing flaps and three filler words containing non-alveolar stops. To reduce the chances that participants might discover the purpose of the study by noticing that each block contained exactly two items with a flap, other patterns were intentionally created with the filler items. These include: (1) all of the blocks contain exactly one filler with the same initial segment as one of the target items, (2) the target items never share the same vowel, and (3) at least two of the filler items share the same vowel, a different vowel from the target items. In addition, the pairs of target words with identical codas (e.g. "chot" and "zot") were included in separate blocks, so that no block contained rhyming target words. Both the ordering of the blocks and the items within the blocks were randomized, with the constraint that the items containing flaps did not occur immediately following each other. The complete stimuli list is included in Appendix A.

The phrases were read by a native speaker of English (a native of Anchorage, Alaska, currently living in Houston, Texas). The flaps in all of the target stimuli were presented to the reader orthographically as "t"; since the reader was familiar with the International Phonetic Alphabet (IPA), the target stimuli were accompanied by an IPA using the alveolar flap symbol [ɾ]. To assure that participants were only using cues from vowel length to determine whether the flap represented a voiced or voiceless stop, the duration of each flap was modified using the Praat program. The duration was changed by setting the duration points in Praat's Manipulation program. The flaps were

all set to 31 ms, the average flap duration of all of the stimuli as recorded. This duration is within the normal range of flap durations found in the TIMIT corpus of American English, reported in Byrd [1993] as a mean of 29 ms (s.d. = 8 ms). For the purposes of measuring and modifying the flap, the aperiodic or quasi-periodic portion of the waveform between the two vowels was used to determine the location of the flap.

The vowel lengths were also modified using Praat's Manipulation program. Two sets of stimuli were created, one in which all vowel durations are equal, and one in which half of the vowel durations are exactly 20 ms. longer than the other half. Thus for the control stimuli, the vowel durations were all set to 128 ms., the average vowel length of the stimuli as recorded. For the experimental stimuli, the data was divided in half, with each half containing exactly ten words each having a different vowel. In one half, the vowel durations of the tokens were set to 118 ms., while in the other half, they were set to 138 ms.

The difference of 20 ms. was selected because it is the maximum difference in vowel length reported in the literature on incomplete neutralization, in this case, that of Russian reported by Pye [1986] as 5-20 ms. Many of the differences reported were much smaller than this, for example Jassem and Richter [1989] reported a difference in vowel length of only 4 ms. for a case of incomplete neutralization in Polish. The largest difference was selected to increase the chances of finding a difference in the results; if it can be shown that 20 ms. produces a difference in perception, further studies can determine the threshold needed to produce this difference. This difference is slightly longer than the largest difference found in previous studies on American English, namely 16 ms in Patterson and Connine [2001].

Fifty clip art pictures were chosen depicting people performing various activities, such as running, singing, and reading. Each picture was associated with one of the pseudo-words. As a pre-test, the stimuli list was presented to five speakers of English, and they were asked to determine if any of the pseudo-words resembled the words for the actions that they were intended to represent. If any of them found a similarity, the word was changed.

Participants

Thirty native speakers of North American English participated in the study; nine were males and twenty-one were females. All were undergraduate-level university students. None reported any known hearing or speaking disorder, and none were aware of the purpose of the study prior to the experiment.

Experiment procedure

A Windows-based program was created to conduct the experiment. The experiment consisted of three phases: an exposure phase, a memorization phase, and an experimental phase. In each of the phases, participants saw a depiction of a particular action, such as a person swimming, or several people hiking. The pseudo-words were presented to the participants in the exposure phase in the "-ing" form, such as "he's chotting". This form was also used in the memorization phase. Finally, in the experimental phase, the base form of the pseudo-word was elicited, as in "he likes to chot/chod", thus testing whether the participant interpreted the flap as a "t" or "d".

Since the program required some input from both the experimenter and the participant, the experimenter maintained control of the keyboard during the experiment, and the participant maintained control of the mouse. The experimenter determined whether the participants' responses were correct or incorrect, and responded by pressing the "r" and "w" keys respectively (for "right" and "wrong"). The participant used the mouse to start each of the ten blocks. After each response, either a "happy" face or a "sad" face was displayed on the computer screen, depending on whether the response was correct or incorrect. In addition, the participant heard one of four pre-recorded phrases: "good job", "that's right", "very good", or "you're doing great" for correct responses; "listen again", "listen carefully", "sorry try again", or "try it again" for incorrect responses. The phrases were recorded by the same person who recorded the stimuli. Although the experimenter was present during the entirety of the epxeriment, interaction between the experimenter and the participant was kept to a minimum during the blocks. The responses were recorded using a Microtrack digital recorder and a lavaliere microphone.

Participants were randomly assigned to one of two groups: a control group and an experimental group.[2] The control group heard the stimuli in which all of the vowel durations were identical; the experimental group heard the stimuli in which half of the vowel durations were 20 ms. longer than the other half. The visual stimuli and experimental procedure for each group were identical.

Before beginning the experiment, a practice block consisting of two real English words was run in order for participants to become familiar with the format of the study.

In the exposure phase, the participant saw one of the clipart pictures and heard one of the stimuli phrases, such as "he's chotting." This was followed by a question of

[2] Due to the increased difficulty of locating subjects who are able to return for a second experiment, the experiment was conducted across subjects, rather than within subjects.

the type "what is he doing?" Participants were instructed to repeat the phrase with the pseudo-word.

The memorization phase was identical to the exposure phase, except that participants did not hear the stimulus again (eg. "he's chotting"). Thus, the participants were expected to start memorizing the pseudo-words. In cases where they get one wrong, the phrase was repeated (eg. "he's chotting"), following by the question (eg. "what is he doing?"). The program automatically adjusted to the errors by repeating more often the items that participants got wrong. The words were considered to be fully memorized when participants were able to go through the block two consecutive times with no mistakes.

Participants required on average 3.1 repetitions of each block to memorize the words during this phase. The highest number of repetitions required occurred on average in the first and last blocks presented to the participant[3], with average of 3.5 and 3.3 repetitions respectively. This is presumably because the task was still new at the beginning of the experiment, and because some participants may have become tired by the end. To see if this had any effect on the responses, ANOVA tests were used to compare the proportion of responses containing "t" in the first five blocks versus the last five blocks (referred to as "half"). In the experimental group, neither half [$F(1, 4) = 1.987$, $p = 0.231$] nor vowel length [$F(1, 4) = 0.234$, $p = 0.654$] were significant factors. In the control group (in which vowel lengths were identical), half was not a significant factor [$F(1,9) = 0.219$, $p = 0.651$].

Finally, in the experimental phase, the question types were changed to "what does he like to do?" thus forcing the participant to use either a voiced or voiceless stop in the response, eg. "he likes to chot/chod." If participants did not respond with the correct word, they were given one more chance at the end of the block. In cases where participants did not remember the word after two tries, it was excluded from the analysis ($n = 19$ of 600 cases).

The responses during the experimental phase were analyzed impressionistically as either [t] or [d]. Ambiguous or inaudible responses were not counted ($n = 5$ of 600 cases). Responses that contained minor variations from the target word, such as [zet] instead of [zæt], were counted as correct.

[3] Since the blocks were presented in random order, "first" and "last" refers to the order in which they were presented, not to the block number.

Hypotheses

The null hypothesis is that a 20 ms. difference in pre-flap vowel length has no effect on participants' interpretation of the flaps as phonologically [t] or [d]. The alternative hypothesis consists of two parts: (1) participants in the experimental group interpret the flaps before long vowels (those with durations of 0.138 seconds) less often as [t] than participants in the control group, and (2) participants in the experimental group interpret the flaps before short vowels (those with durations of 0.118 seconds) more often as [t] than those in the control group.

Results

Table 1 shows for each coda type (defined by the vowel used in the coda) the proportion of participants who interpreting the flap as "t". A repeated-measures ANOVA comparing the proportion of participants who interpreted the stops as "t" shows that there is a significant difference between the control group and the experimental group [$F(1,9) = 46.556$, $p < 0.001$] but not between the groups that heard long vowels and short vowels [$F(1,9) = 0.044$, $p = 0.838$]. In other words, while participants in the experimental group interpreted the flaps differently from those in the control group, neither group interpreted the flaps differently according to whether they heard a long vowel or short vowel. Of course, for the control group, this result is a natural consequence of the fact that the vowel durations were all equal, but in the experimental group, it shows that the participants were not able to use the vowel duration difference to determine whether the flaps were phonologically [t] and [d]. Overall, participants in the experimental group were more likely to interpret the flaps as "t", regardless of whether the vowel was long or short. Thus, the null hypothesis can be rejected on the grounds that vowel durations do have an effect on participants' interpretation of the flaps. But part (1) of the alternative hypothesis must also be rejected: participants in the experimental group did not interpret the flaps before long vowels less often as [t] than participants in the control group. The evidence only supports part (2) of the alternate hypothesis, that participants in the experimental group interpreted the flaps before short vowel more often as [t] than those in the control group.

In the control group, participants were more likely to interpret the flaps as "d" than "t". Of the 286 total responses, there were 219 instances of "d", 77% of the total. In the experimental group, however, the interpretations were more evenly divided between "t" and "d"; of the 281 total responses, 163 were interpreted as "d", 58% of the total. Since there is no significant difference in the experimental group between

items with long vowels and short vowels, this difference cannot be attributed to the participants' ability to consistently distinguish between the two sets of token on the basis of vowel duration. However, it does show that listeners are sensitive to the fact that there is an overall difference in vowel length, and perhaps even to the fact that the tokens were evenly divided among short vowels and long vowels. The perception that half of the target tokens had a vowel that was longer than the other half seems to have increased the likelihood that participants' interpretations of the flaps were more evenly split between "t" and "d" than participants who heard all target tokens with the same vowel length.

With respect to the idea that the flapping of /t/ and /d/ in American English results in an incomplete neutralization, the results show that while this is still possible, they do not show concrete evidence for it. The vowel length difference of 20 ms. is generous compared to the differences reported in the literature on similar types of incomplete neutralizations, which range from 4 to 20 ms. Thus this study does not provide evidence that vowel length alone is enough to distinguish between /t/-flaps and /d/-flaps in American English.

Discussion and Conclusion

Proponents of the idea of incomplete neutralizations often use it to challenge the idea of discreteness of phonetic categories. Port [1996], for example, claims that "[t]he phenomenon of incomplete neutralization *is* a worrisome chink in the dam that supports all of current symbol-based phonological theory." But as Manaster Ramer [1996] points out, claims about the reality of incomplete neutralizations have not yet been demonstrated conclusively since, among other reasons, they have not distinguished between orthographic and phonological representations.

Since this study looks only at perception, it is not intended to directly address whether incomplete neutralization is a real phenomenon in speech production. Previous studies, such as, Port and Crawford [1989] found that listeners are able to distinguish at above-chance levels, between words that have a supposed incomplete neutralization due to final devoicing, but is it not clear what phonetic factors are involved in making the distinction. Furthermore, as with most studies on incomplete neutralization, it does not address the question of whether the entire phenomenon is due to an influence from the orthography. This seems very likely in the case of Port and Crawford's study since the speech data comes from contrastive sentences such as "Ich habe 'Rat' gesagt; nicht 'Rad'" (*I said 'Rat'; not 'Rad'*), as well as from word lists. In these cases, speakers

who do not otherwise distinguish between the words are likely to pronounce them in a way that is contrastive for the sake of comprehension.

By manipulating only the vowel durations and maintaining a consistent contrast of 20 ms. between the phonologically voiced and voiceless groups of words, this study was able to test whether this factor alone is sufficient for listeners to perceive a distinction in a reported case of incomplete neutralization. Furthermore, by avoiding both the exposure to and the use of orthography throughout the study (except for the reading of the stimuli list, in which all contrasts to be made were spelled with the same grapheme), it eliminates the possiblity that listeners may interpret cues that are only present in unnatural "reading" pronunciations.

The fact that participants behave differently in each of the two groups—the one that heard identical vowel lengths and the one that heard contrastive vowel lengths—shows that the 20 ms. difference in vowel duration is enough to affect listeners' perceptions of flaps in North American English. But the results do not show that they are more likely to interpret flaps preceded by long vowels as phonologically voiced than those preceded by short vowels. This suggests that if the flapping of alveolar stops in North American English is indeed a case of incomplete neutralization, the distinction is likely not to be maintained by vowel length alone.

One possible confound on this study, however, may be the fact that all of the flaps were set to the same duration. Although neither Fox and Terbeek [1977] nor Zue and Laferriere [1979] found a significant difference in the duration of [t]-flaps and [d]-flaps, Zue and Laferriere [1979] did find that flap duration was affected by the proceeding vowel, specifically that the duration was greater after high vowels or vowels with high glides (eg. [i], [u], [eɪ], [aɪ]). Thus future studies should account for such differences.

In addition, measuring the vowel lengths of the participants responses during the study may help to explain the results obtained in the study. For example, if the participants' own pronunciations of the pseudo-words in the experimental group reflect the difference in vowel length in the exposure phase, but not in the memorization phase, this may indicate that the results were affected more by the participants' memory than their perception of the vowel lengths. If this is the case, then modifications to the experimental procedure will be necessary in future studies to account this issue.

Byrd, D.: 54,000 American Stops. UCLA Working Papers in Phonetics 83: 97-116 (1993)

Fox, R.; Terbeek D.: Dental flaps, vowel duration and rule ordering in American English. J. of Phonetics *5*: 27-34 (1977).

Jassem, W.; Richter, L.: Neutralization of voicing in Polish obstruents. J. of Phonetics *17*: 317-325 (1989).

Manaster Ramer, A.: A letter from an incompletely neutral phonologist. J. of Phonetics *24*: 477-489 (1996).

Patterson, D.; Connine, C.: Variant frequency in flap production: A corpus analysis of variant frequency in American English flap production. Phonetic 58: 254-275 (2001).

Port, R.: The discreteness of phonetic elements and formal linguistics: response to A. Manaster Ramer. J. of Phonetics *24*: 491-511 (1996).

Port, R., Crawford, P.: Incomplete neutralization and pragmatics in German. J. of Phonetics *17:* 257-282 (1989).

Pye, S.: Word-final devoicing of obstruents in Russian. Cambridge papers in phonetics and experimental linguistics. *5*: 1-10 (1986).

Slowiaczek, L.; Dinnsen, D.: On the neutralization status of Polish word-final devoicing. J. of Phonetics *13*: 325-341 (1985).

Van Rooy, B.; Wissing, D.; Paschall, D.: Demystifying incomplete neutralization during final devoicing. Southern African linguistics and applied language studies *21 (1-2)*: 49-66 (2003).

Warner, N.; Jongman, A.; Sereno, J.; Kemps, R.: Incomplete neutralization and other sub-phonemic durational differences and perception: evidence from Dutch. J. of Phonetics *32*: 251-276 (2004).

Zue, V.; Lafferriere, M.: Acoustic study of medial /t,d/ in American English. J. Acoust. Soc. Am. 66(4): 1039-1050 (1979).

Table 1. Proportions of participants who interpreted the flaps as "t" for each coda type.

Coda Type	Short vowel group		Long vowel group	
	Control	Experimental	Control	Experimental
i	21% (3/14)	50% (7/14)	7% (1/14)	27% (4/15)
ɪ	0% (0/15)	38% (5/13)	43% (6/14)	40% (6/15)
eɪ	21% (3/14)	23% (3/13)	27% (4/15)	73% (11/15)
ɛ	43% (6/14)	67% (8/12)	27% (4/15)	50% (7/14)
æ	40% (6/15)	79% (11/14)	60% (9/15)	73% (11/15)
aɪ	46% (6/13)	71% (10/14)	0% (0/12)	33% (5/15)
ɑ	27% (4/15)	50% (7/14)	29% (4/14)	29% (4/14)
ʌ	0% (0/15)	0% (0/14)	21% (3/14)	42% (5/12)
oʊ	14% (2/14)	27% (4/15)	13% (2/15)	40% (6/15)
u	13% (2/15)	8% (1/13)	14% (2/14)	21% (3/14)
MEAN	**23% (32/144)**	**41% (56/137)**	**24% (35/142)**	**43% (62/144)**

Appendix A: Stimuli used in the experiment

Each block consists of two target items plus three filler. The target items are listed first.

Block 1

she's veeting /viːtiŋ/	"she's bicycling"
he's gooding /guːdiŋ/	"he's cleaning"
she's rogging /ɹɑgiŋ/	"she's teaching"
they're zopping /zɑpiŋ/	"they're laughing"
they're gibing /gɑɪbiŋ/	"they're kissing"

Block 2

he's vitting /vɪtiŋ/	"he's hunting"
she's zadding /zædiŋ/	"she's combing her hair"
she's maping /meɪpiŋ/	"she's washing dishes"
he's gaking /geɪkiŋ/	"he's swinging"
they're veeming /viːmiŋ/	"they're shaking hands"

Block 3

he's chating /tʃeɪtiŋ/	"he's bowling"
he's zodding /zɑdiŋ/	"he's shaving"

she’s nooping /nuːpiŋ/ “she’s painting”
he’s rooking /ɹuːkiŋ/ “he’s tearing”
they’re choping /tʃoʊpiŋ/ “they’re singing”

Block 4

they’re chetting /tʃɛtiŋ/ “they’re dancing”
he’s vudding /vʌdiŋ/ “he’s hammering”
he’s shobing /ʃoʊbiŋ/ “he’s taking a bath”
he’s toging /toʊgiŋ/ “he’s typing”
she’s chigging /tʃɪgiŋ/ “she’s drinking”

Block 5

she’s vatting /vætiŋ/ “she’s diving”
he’s jidding /dʒɪdiŋ/ “he’s writing”
he’s vaping /veɪpiŋ/ “he’s speaking”
he’s chaking /tʃeɪkiŋ/ “he’s kicking”
they’re jebbing /dʒɛbiŋ/ “they’re juggling”

Block 6

she’s chotting /tʃɑtiŋ/ “she’s ironing”
he’s nading /neɪdiŋ/ “he’s crying”
they’re zibbing /zɪbiŋ/ “they’re boxing”
he’s bicking /bɪkiŋ/ “he’s reading”
she’s nassing /næsiŋ/ “she’s playing music”

Block 7

she’s futting /fʌtiŋ/ “she’s knitting”
he’s kedding /kɛdiŋ/ “he’s swimming”
he’s habbing /hæbiŋ/ “he’s hang-gliding”
she’s dagging /dægiŋ/ “she’s brushing her teeth”
he’s foping /foʊpiŋ/ “he’s eating”

Block 8

he’s foting /foʊtiŋ/ “he’s driving”
they’re jiding /dʒaɪdiŋ/ “they’re hugging”

she’s tepping /tɛpiŋ/ “she’s jumping’
he’s gecking /gɛkiŋ/ “he’s raking leaves”
she’s febbing /fɛbiŋ/ “she’s sewing”

Block 9

he’s dooting /duːtiŋ/ “he’s mowing the lawn”
he’s geeding /giːdiŋ/ “he’s running”
they’re chiking /tʃaɪkiŋ/ “they’re canoeing”
she’s kiping /kaɪpiŋ/ “she’s sleeping”
he’s gabing /geɪbiŋ/ “he’s mopping”

Block 10

he’s ziting /zaɪtiŋ/ “he’s climbing”
she’s poding /poʊdiŋ/ “she’s sweeping”
she’s lubbing /lʌbiŋ/ “she’s washing clothes”
he’s gucking /gʌkiŋ/ “he’s skipping rope”
he’s zecking /zɛkiŋ/ “he’s taking a shower”