Perception of Incompletely Neutralized /d/ and /t/ Flaps in American English[*]

Aaron Braver

Rutgers, The State University of New Jersey

1. Introduction

It has long been known that underlying /d/ and /t/ become flaps in certain post-tonic positions in American English (see, e.g., Kahn 1980). A number of studies have examined the possibility that the underlying voicing status of a flap might be reflected on the surface (Fisher and Hirsh 1976, Fox and Terbeek 1977, Zue and Laferriere 1979, Joos 1942, Port 1976, Huff 1980, Herd et al. 2010, Braver 2011). Of the studies that found /d/ and /t/ flaps to be incompletely neutralized, the surface distinctions tend to be relatively small, and mostly found in terms of preceding vowel duration. While several studies have been conducted to test listeners' ability to categorize /d/ and /t/ flaps, they have had mixed results: some studies find some degree of distinguishability (Sharf 1960, Fisher and Hirsh 1976), while others suggest that listeners perform at near-chance levels (Malécot and Lloyd 1968, Herd et al. 2010). These studies tend to focus on identification tasks using a small number of actual English words, and use measures of distinguishability that do not tease apart listeners' bias from their actual sensitivity to the distinction being tested.

This paper reports three perception experiments on flapping in American English, designed to address some of the questions and concerns raised by previous such studies. Specifically, the experiments consist of both identification and discrimination tasks, and use nonce-words to mitigate effects of lexical frequency on listeners' choices. Additionally, the use of d' as a measure of distinguishability, which unlike the percent-correct measure takes bias into account, paints a more clear picture of listeners' actual ability to distinguish /d/ and /t/ flaps.

1.1 Previous Studies of Incomplete Neutralization in American English Flapping

A number of production studies report that, for some speakers, flaps stemming from underlying /d/ ('/d/ flaps') and those stemming from underlying /t/ ('/t/ flaps') are distinct on the

[*] Thanks are due to Shigeto Kawahara, Bruce Tesar, and Kristen Syrett for their input on various stages of this research.

surface. The most commonly reported distinction between /d/ and /t/ flaps is a difference in the duration of the preceding vowel; however differences have also been reported in degree of intensity dip during flap closure and duration of the closure (Fisher and Hirsh 1976, Fox and Terbeek 1977, Zue and Laferriere 1979, Braver 2011). For example, Herd et al.'s 2010 study found, on average, that vowels preceding /d/ flaps were 6ms longer than vowels preceding /t/ flaps. Huff (1980) found mixed results in New York City speakers, with the distinction between /d/ and /t/ flaps manifesting as F1 and F2 differences in some, but not all, pre-flap vowels. 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 with the vowel [ɪ]).

In describing prior perception studies of incomplete neutralization, and throughout this paper, I maintain the more or less standard distinction between 'identification' and 'discrimination', where 'identification' refers to listeners' ability to label a given segment with its appropriate phonological category, and 'discrimination' refers to listeners' ability to tell the difference between two (or more) segments, regardless of the method used to do so. Listeners can use identification to help them discriminate ('this sound was a /d/, and the next sound was a /t/, so they are different'), but discrimination can also proceed without identification ('these two sounds were different, but I do not know what the sounds were') (see, e.g., Liberman et al. 1957, 1967 on the hypothesis that discrimination is guided by identification). I use 'distinguish' and 'distinguishability' as cover terms, which refer to an ability to either discriminate or identify.

Early perception studies of listeners' abilities to distinguish /d/ flaps from /t/ flaps show mixed results. Sharf (1960) presents an identification task with English minimal pairs constructed from tokens produced by one male and one female speaker. While listeners were accurate 86% of the time in identifying /d/ and /t/ flaps in tokens from the female speaker, their accuracy dropped to 61% correct for tokens from the male speaker. Malécot and Lloyd (1968), who used a similar procedure, 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 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).[1]

A more recent study, Herd et al. (2010), presents a forced-choice identification task based on four pairs of actual English words. By carefully selecting tokens from their production experiment, three conditions were created: a 'mean' condition, with duration differences between pre-/d/ and pre-/t/ vowels of 7–9ms, an 'enhanced' condition, with duration differences between 22–34ms, and an 'opposite' condition, in which pre-/t/ vowels were longer than pre-/d/ vowels by 5–21ms. Listeners heard tokens from each of four word pairs, and were asked to select which member of the pair they heard (e.g., when listeners heard [ˈliɾɚ], they chose between 'leader' and 'liter'). On average, listeners fell near chance, correctly identifying 52% of tokens in the mean and enhanced conditions, and 48% in the opposite condition. Tokens with an underlying /d/ (e.g., 'leader') were more often correctly identified (57%) than words containing underlying /t/ (e.g., 'liter', 44% correct).

[1] Fisher and Hirsh (1976) note that one judge could not confidently determine whether a flap was voiced or voiceless, and so he labeled all flaps as 'flap, but I can't say whether voiceless or voiced' (p. 187).

1.2 Motivation for the Current Study

In their perception study, Herd et al. (2010) find two kinds of lexical frequency effects: a main effect of frequency and an interaction between lexical frequency and underlying voicing. First, words with high lexical frequency were correctly identified more often (59%) than words with low lexical frequency (42%). Second, /d/ words were less susceptible to this effect than /t/ words (62% correct for high-frequency /d/ words, 51% correct for low-frequency /d/ words; 55% correct for high frequency /t/ words, 33% correct for low-frequency /t/ words). These results are likely due to the general bias for listeners to choose more frequent responses (Connine et al. 1993). In order to mitigate possible effects of lexical frequency, the tasks in this experiment use nonce words—which have a lexical frequency of zero—rather than actual English words. Using nonce words also allows for a greater number of different stimuli, since English has only a limited number of minimal pairs that differ just by /d/ or /t/ in a flapping environment.

In addition to lexical frequency effects, Herd et al. (2010) found a strong bias for /d/: /d/ words were accurately perceived 57% of the time, while /t/ words were only perceived accurately 44% of the time. The perception results reported here are reported in terms of d', which unlike the percent-correct measure teases apart sensitivity from bias (Macmillan and Creelman 2005). Bias of this sort can 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, in this scenario, 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/.

Additionally, most prior perception studies of incomplete neutralization in flapping have focused on identification tasks. A question of interest, then, is whether the results from tasks like the one presented in Herd et al. (2010) are due to the difficulty of the identification task itself, or from the imperceptibility of the contrast being tested. In order to distinguish between these two possibilities, three perceptual tasks are reported here: an identification task, an ABX task, and a 2-Alternative Forced Choice (2AFC) task. In order to afford every possible benefit to listeners, a practice phase with both actual English words and nonce words was included at the start of each task, and feedback was provided after each trial.

2. Experiment 1: Identification

The vowel duration differences found in production experiments of incomplete neutralization in flapping are quite small—generally less than 10ms (Herd et al. 2010, Braver 2011, 2012). Herd et al. (2010) showed that on a basic identification task, listeners perform poorly. The task 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. Given the difficulty of identification tasks, two additional tasks are reported in the sections that follow—an ABX task and a 2AFC task—to examine the possibility that earlier results from identification studies are due to the difficulty of such tasks, rather than an inability on the part of listeners to distinguish /d/ flaps from /t/ flaps.

The identification task was designed to examine whether listeners can accurately categorize /d/ flaps from /t/ flaps in nonce word tokens taken from a related production task not reported here for reasons of space (see Braver 2012). The experiment 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.

2.1 Participants and Equipment

21 undergraduates participated in this experiment, all of whom were native speakers of English. The experiment took place at the Rutgers Phonetics Laboratory, with stimuli displayed and responses recorded by SuperLab 4.5 (Cedrus Corporation 2010), using Sennheiser HD 280 Professional headphones.

2.2 Stimuli

The tokens for all tasks reported in this paper came from a related production experiment (see Braver 2012). In the production experiment, speakers participated in one of two tasks—a minimal pair reading task and a 'wug'-type fill-in-the-blank task (Berko 1958, Fourakis and Iverson 1984). No significant differences were found between the two tasks. The stimuli in the production experiment were comprised of minimal pairs of disyllabic nonce words which ended in either /d/ or /t/. The '-ing' suffix was then added to these stimuli, placing the alveolar stop in a post-tonic intervocalic context where flapping occurs. 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 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/. The stimuli were written in English orthography, with the second syllable capitalized to indicate stress, and with schwas indicated by 'uh' (for example: puhPAT(-ing)∼puhPAD(-ing), and tuh-KEET(-ing)∼tuhKEED(-ing)).

Tokens were then selected from this production task for use in the perception experiments reported here. From among the tokens of all 12 speakers in the acoustic experiment, tokens 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 tokens. Tokens were chosen from each speaker to maximize the pre-flap vowel duration difference between members of a minimal pair, while at the same time balancing onset and vowel of the target syllable, as well as for the voicing of the target segment (/d/ or /t/). The Identification experiment, as with the perception experiments that follow, was blocked such that a given block had only tokens from a single speaker.

2.3 Procedure

Listeners read instructions for the task and practiced with both English and nonce words. On each experimental trial, listeners heard a single token and were directed to press one of two buttons indicating whether the sound immediately preceding the '-ing' was a /t/ or a /d/. For example, upon hearing 'buhKEED-ing', listeners would press the /d/ button. Visual feedback was provided on each trial.

The task was divided into three blocks (one for each of the three speakers from whom tokens were used). Each block consisted of 36 trials (half /d/ and half /t/), randomized, with three repetitions each (=108 total trials). Block order was balanced (Latin Square) across all listeners.

Listeners' sensitivity was assessed through the d′ measure, which unlike percent-correct, teases apart listeners' actual sensitivity from bias (Macmillan and Creelman 2005). d′ is calculated using hit rate (H)—the proportion of target trials to which a listener responds 'target'—and false alarm rate (F)—the proportion of non-target trials to which a listener responds 'target'. d′ for this task was calculated as $d' = z(H) - z(F)$. A d′ score of zero indicates an inability to distinguish; as d′ scores increase, they indicate an improvement in distinguishability.

2.4 Results

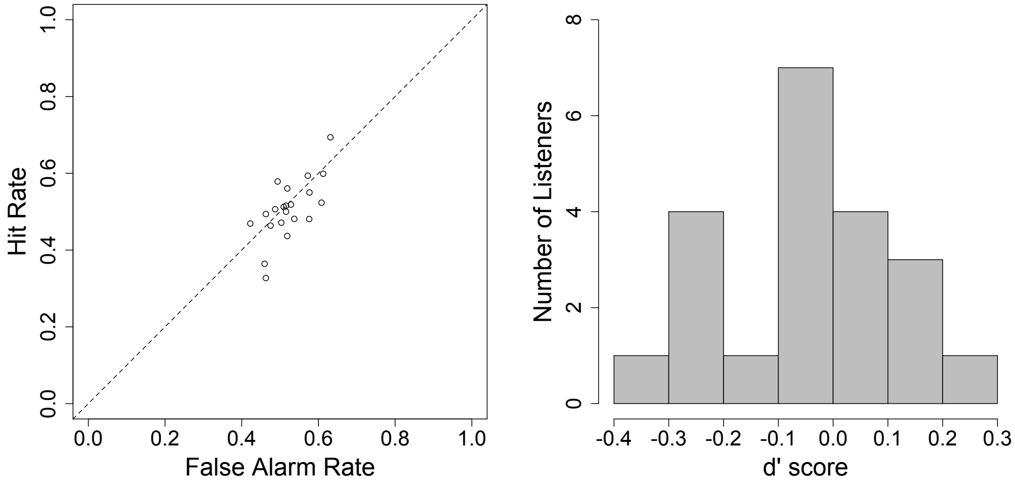
The mean d' score across all listeners for the Identification task was $d' = -0.04$, which is not significantly different from zero (Wilcoxon test: $V = 76, n.s.$). In other words, listeners (correctly) pressed the /d/ button when they had heard a /d/ just as frequently as they had (incorrectly) pressed the /d/ button when they had actually heard a /t/. Figure 1(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 false alarms (saying '/d/' when they had actually heard a /t/). This is reflected in the frequency distribution of d′ scores shown in Figure 1(b), which center around zero.

2.5 Section summary and discussion

Listeners in this task were unable to correctly categorize /d/ flaps and /t/ flaps. The results are in line with those found by Herd et al. (2010), which used actual English words, even though this study used nonce words. This experiment, however, leaves open the question of whether listeners will show better performance on easier tasks—a question addressed by the ABX and 2AFC tasks described in the following sections.

3. Experiment 2: ABX Task

In an ABX task, listeners hear three stimuli on a given trial, and so are able to make comparisons to aid in their decision-making. In order to investigate whether American English



(a) Hits vs. False Alarms: ID task

(b) Frequency plot of d′ scores: ID task

Figure 1: Results of the Identification task

listeners truly are unable to distinguish between /d/ and /t/ flaps—or whether results from identification tasks are due to the difficulty of the task itself—an ABX task was conducted.

3.1 Participants and Equipment

21 native English speaking undergraduates, none of whom had participated in the Identification task, participated in this experiment. The experiment was administered in the Rutgers Phonetics Laboratory, with stimuli displayed and responses recorded by SuperLab 4.5 (Cedrus Corporation 2010), using Sennheiser HD Professional headphones.

3.2 Stimuli

On each trial, listeners heard a total of three nonce words, drawn from the same set of tokens used in the Identification task. On each trial, two tokens were physically identical, and the remaining token formed a minimal pair with the identical tokens, differing only in whether it had a /d/ or a /t/ immediately preceding the '-ing'.

3.3 Procedure

As with the prior task, participants read instructions and practiced with both English and nonce words. On each trial, listeners heard three stimuli (A, then B, then X), and were asked to press a button indicating whether the third word (X) was the same as the first word (A) or the second word (B). For example, if listeners heard 'puhPEED-ing - puhPEET-ing - puhPEET-ing', they would press the button indicating that the second word, B, was the same as X.

The A and B stimuli were separated by 250ms of silence, while the B and X stimuli were separated by 500ms. This longer B-X inter-stimulus interval was used in order to allow listeners time to make an attempt at category labeling of the first two stimuli before hearing the third, and to induce listeners to employ a more categorical, rather than auditory, mode of perception (Pisoni 1973, 1975, Gerrits and Schouten 2004).

As with the Identification task, feedback was provided on each trial. Each block was randomized, and consisted of 18 trials of each of the following 4 configurations (72 trials total per block), such that /d/ words and /t/ words each appeared equally as stimulus A and stimulus B, and such that the target stimulus (X) matched A and matched B an equal number of times: (i) A=/d/, B=/t/, X=/t/, (ii) A=/d/, B=/t/, X=/d/, (iii) A=/t/, B=/d/, X=/d/, (iv) A=/t/, B=/d/, X=/t/. Block order was balanced (Latin Square) across all listeners.

Listeners' sensitivity was again measured using d$'$. Since this task employed a roving ABX design, I assume a differencing strategy (Macmillan and Creelman 2005, pp. 221-225, 233), and d$'$ values were calculated using the `psyphy` package (Knoblauch 2011) with the `method="diff"` option in R (R Development Core Team 2009).

3.4 Results

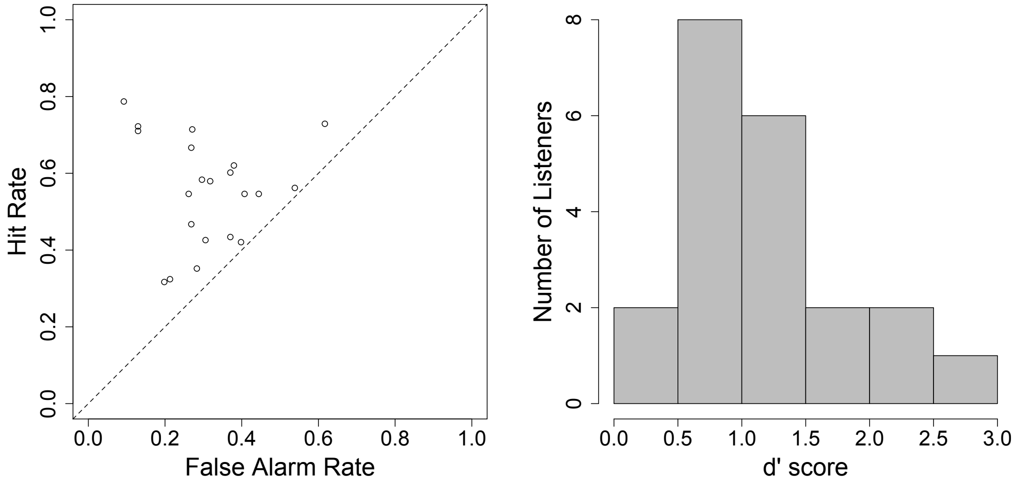
The mean d' score across all listeners for the ABX task was $d' = 1.24$, which is significantly different from zero (Wilcoxon test: $V = 231, p < 0.001.$). In other words, listeners said that A was the same as X when that was the case (a hit) more frequently than they had said A was the same as X when it had actually been B that was the same as X (a false alarm). Figure 2(a) shows a plot of the hit rate vs. false alarm rate for participants in the ABX task. All listeners had higher hit rates than false alarm rates—they had more hits (saying 'A is the same as X' when that was the case) than false alarms (saying 'A is the same as X' when that was not the case). This is reflected in the frequency distribution of d$'$ scores shown in Figure 2(b), which are all above zero.

After completing the task, participants were informally asked what strategy they had used. Many listeners indicated that they had focused on cues that do not seem to correlate with voicing, or to physical differences between individual tokens such as variations in pitch contour.

3.5 Section summary and discussion

While listeners' d$'$ scores were relatively good on this task as compared with the Identification task, the informal reports of their use of cues unrelated to the underlying voicing distinction suggests that the results may not reflect listeners' actual ability to distinguish /t/ flaps from /d/ flaps. Rather, since two tokens on each trial were physically identical, it is possible that listeners listened for any deviation between the tokens—regardless of whether it was related to the target voicing distinction.

To support the hypothesis that listeners focused on physical differences between the stimuli rather than the target distinctions, the best five and worst five performing minimal pairs of stimuli were analyzed. Among the best-performing pairs, a difference unrelated to the voicing distinction between the two tokens was always evident. For example, in



(a) Hits vs. False Alarms: ABX task

(b) Frequency plot of d′ scores: ABX task

Figure 2: Results of the ABX task

the pair 'tuhKAD-ing'∼'tuhKAT-ing' (from speaker 12 in the related production experiment), the '-ing' syllable of 'tuhKAD-ing' had creaky voice, while the matching syllable of 'tuhKAT-ing' did not. While this difference is presumably unrelated to the t/d distinction at hand, listeners in the ABX task could have used the creaky voice to identify which tokens were identical when choosing their responses. Similarly, in the pair 'duhKEED-ing'∼'duhKEET-ing' (from the same speaker), the mean pitch of the stressed syllable differed by 16.13 Hz (243.93Hz for 'duhKEED-ing', 260.06Hz for 'duhKEET-ing'). Listeners may have based their responses on this pitch difference, which seems unrelated to the t/d distinction.

At the same time, one might expect the worst-performing pairs to be those with the smallest vowel duration differences, since preceding vowel duration is hypothesized to be the main cue to the underlying voicing status of a flap. This, however, is not the case. In fact, all five of the worst-performing pairs had larger vowel duration differences (mean difference: 7.49ms) than the 'duhKEED-ing'∼'duhKEET-ing' pair (vowel duration difference: 1.63ms), which was among the best-performing pairs. In other words, a large difference in preceding vowel duration did not necessarily mean that listeners would correctly distinguish a given pair. This suggests that listeners in this task must have used some other sort of cue—most likely the physical differences between tokens that were unrelated to the distinction being investigated.

The results of this analysis also suggest that, despite the expectation that the extended ISI employed in this experiment would induce speakers to use a categorical mode of perception, speakers in fact used a more auditory mode of perception. Had listeners used a categorical mode of perception, one would expect them to pay less attention to the physical differences between tokens than they did. One possible explanation for this tendency is that listeners were simply unable to call upon the categorical mode of perception to help them in the task at hand. If the cues necessary for categorizing a flap as underlyingly voiced or voiceless are not present, or are too muted, categorization simply cannot happen. Given the paucity of cues for categorization, listeners may have grasped for any distinctions possible—whether relevant for phonetic categorization or not.

The listeners' relatively good performance on this task suggests an ability to discriminate between the particular /d/ and /t/ tokens presented to them—but, taken together with the analysis of the best- and worst-performing tokens, does not necessarily indicate a more general ability to distinguish flapped segments on the basis of voicing cues or to relate them to voiced or voiceless categories. In order to further support this conclusion, a 2AFC task was run in which listeners cannot rely on comparisons involving physically identical stimuli.

4. Experiment 3: 2AFC Task

In order to examine whether listeners' relatively strong performance on the ABX task was due to the use of comparisons involving physically identical stimuli and distinctions unrelated to voicing, a 2-Alternative Forced Choice task (2AFC) was run. While in the ABX task listeners hear three stimuli on a given trial—two of which are physically identical—listeners in a 2AFC task hear only two stimuli per trial, and they are always different from one another. This design, while preventing the 'physical similarities' strategy, is still easier than an identification task, since listeners are able to compare two stimuli (Macmillan and Creelman 2005, pp. 167-170).

4.1 Participants and Equipment

24 undergraduates participated in this experiment, none of whom participated in either of the previous experiments. Again, the experiment was carried out at the Rutgers Phonetics Laboratory, with stimuli displayed and responses recorded by SuperLab 4.5 (Cedrus Corporation 2010), using Sennheiser HD 280 Professional headphones.

4.2 Stimuli

All tokens were taken from the same set of stimuli used in the previous tasks.

4.3 Procedure

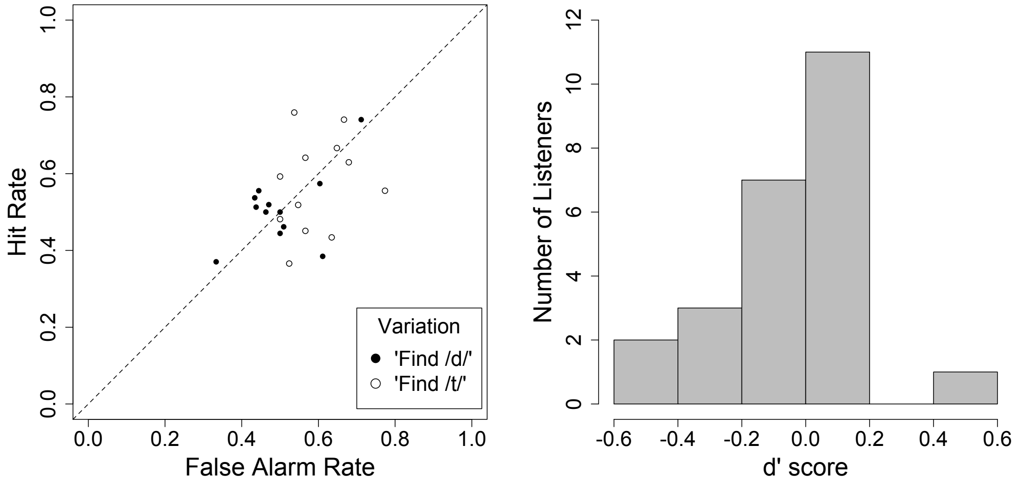
On each trial, listeners heard two tokens—members of a minimal pair—separated by 250ms, and were asked to focus on the sound immediately preceding the '-ing' in each word. Half of the listeners were asked whether the /d/ word came first or second; the other half were asked whether the /t/ word came first or second. For example, a listener who was told to 'find /d/' and heard 'puhPEED-ing - puhPEET-ing' would respond that /d/ had come in the first word.

The experiment was divided into three blocks, with each block containing tokens from only one speaker. Each block consisted of 36 randomized trials (half /d/ and half /t/). Block order was balanced (Latin Square) across all listeners. As with the previous tasks, feedback was provided on each trial.

d′ was again used as a measure of sensitivity, but because 2AFC tasks are easier than identification tasks, d′ was calculated as $d' = \frac{z(H) - z(F)}{\sqrt{2}}$ (Macmillan and Creelman 2005, p. 167-170).

4.4 Results

The mean d′ score for all listeners in 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 when that was the case just as often as when that had not been the case. Figure 3(a) shows a plot of the hit rate vs. false alarm rate for participants in the 2AFC task. As with the identification task, and unlike the ABX task, listeners are clustered around the hit rate = false alarm rate line. Figure 3(b) shows the frequency distribution of d′ scores, which are centered around zero.



(a) Hits vs. False Alarms: 2AFC task

(b) Frequency plot of d′ scores: 2AFC task

Figure 3: Results of the 2AFC task

4.5 Section Summary and Discussion

Listeners in the 2AFC task were unable to accurately distinguish between /d/ flaps and /t/ flaps. Listeners' performance was worse in this task—where a 'physical similarities' strategy was unavailable—than in the ABX task, where such a strategy was possible. Taken together with the post-hoc analysis of the best- and worst-performing pairs in the ABX task, as well as the results from the Identification task, the results of this task suggest that listeners are neither able to categorize nor discriminate /d/ flaps from /t/ flaps.

5. Conclusion

Taken together, the experiments reported here suggest that incompletely neutralized /t/ and /d/ flaps are not distinguishable by American English listeners. Prior studies, which have found mixed results, have generally focused on identification tasks using actual words of English. The experiments presented here made use of nonce words, mitigating potential influence from lexical frequency effects.

The Identification task in §2 is in line with the results of earlier studies, such as Herd et al. (2010), which suggest that listeners cannot categorize /d/ and /t/ flaps in tasks of this type. It further suggests that these results hold both when nonce words are used rather than actual English words, and also when measures like d′, which tease apart listeners' bias from sensitivity, are used as a metric of distinguishability.

In addition to the identification study, an ABX task and a 2AFC task were run to investigate the possibility that listeners' poor performance in earlier identification tasks was due, in part, to the difficulty of such tasks. While listeners in the ABX task showed relatively strong performance, this was most likely due to the use of a 'physical similarities' strategy, rather than an actual ability to discriminate between /d/ and /t/ flaps on the basis of cues related to voicing. The 2AFC task lends further support to this hypothesis. Listeners showed little to no ability to discriminate in the 2AFC task, which is easier than identification tasks, but in which the 'physical similarities' strategy is impossible.

References

Berko, Jean. 1958. The child's learning of English morphology. *Word* 14:150–177.

Braver, Aaron. 2011. Incomplete neutralization in American English flapping: A production study. In *Proceedings of the 34th Annual Penn Linguistics Colloquium*, volume 17 of *University of Pennsylvania Working Papers in Linguistics*. Penn Linguistics Club. http://repository.upenn.edu/pwpl/vol17/iss1/5/.

Braver, Aaron. 2012. Imperceptible incomplete neutralization: Production, identification, and discrimination of /d/ and /t/ flaps in American English. ms. Rutgers.

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

Connine, Cynthia M., Debra Titone, and Jian Wang. 1993. Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology* 19:81–94.

Fisher, William M., and Ira J. Hirsh. 1976. Intervocalic flapping in English. In *Papers from the Twelfth Regional Meeting of the Chicago Linguistic Society*, 183–198. Chicago Linguistic Society.

Fourakis, Marios, and Gregory Iverson. 1984. On the 'incomplete neutralization' of German final obstruents. *Phonetica* 41:140–149.

Fox, Robert A., and Dale Terbeek. 1977. Dental flaps, vowel duration, and rule ordering in

American English. *Journal of Phonetics* 5:27–34.

Gerrits, Ellen, and M.E.H. Schouten. 2004. Categorical perception depends on the discrimination task. *Perception and Psychophysics* 66:363–376.

Herd, Wendy, Allard Jongman, and Joan Sereno. 2010. An acoustic and perceptual analysis of /t/ and /d/ flaps in american english. *Journal of Phonetics* 38:504–516.

Huff, Charles T. 1980. Voicing and flap neutralization in New York City English. *Research in Phonetics* 1:233–256.

Joos, Martin. 1942. A phonological dilemma in Canadian English. *Language* 18:141–144.

Kahn, Daniel. 1980. *Syllable-based generalizations in English phonology*. New York: Garland.

Knoblauch, Kenneth. 2011. *Psyphy: Functions for analyzing psychophysical data in R*.

Liberman, Alvin M., Franklin S. Cooper, Donald P. Shankweiler, and Michael Studdert-Kennedy. 1967. Perception of the speech code. *Psychological Review* 74:431–461.

Liberman, Alvin M., Katherine Safford Harris, Howard S. Hoffman, and Belver C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54:358–368.

Macmillan, Neil A., and C. Douglas Creelman. 2005. *Detection Theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates Inc., 2nd edition.

Malécot, André, and Paul M. Lloyd. 1968. The /t/:/d/ distinction in american alveolar flaps. *Lingua* 19:264–272.

Pisoni, David. 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics* 13:253–260.

Pisoni, David. 1975. Auditory short-term memory and vowel perception. *Memory and Cognition* 3:7–8.

Port, Robert. 1976. The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words. Doctoral Dissertation, University of Connecticut.

R Development Core Team. 2009. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Sharf, Donald J. 1960. Distinctiveness of 'voiced T' words. *American Speech* 35:105–109.

Zue, Victor W., and Martha Laferriere. 1979. Acoustic study of medial /t, d/ in American English. *Journal of the Acoustical Society of America* 66:1039–1050.

Department of Linguistics
Rutgers, The State University of New Jersey
18 Seminary Place
New Brunswick, NJ 08901

abraver@rutgers.edu