Perceptual Similarity Modulates Context Effects in Online Compensation for Phonological Variation[*]

Adam J. Chong and Megha Sundara
University of California, Los Angeles

1 Introduction

During the process of spoken word recognition, listeners must cope with a large amount of variation in the speech signal. Some of this variation is conditioned by context (e.g. allophonic variation) and thus, regular and predictable. Faced with this kind of variation, listeners must be able to map two or more acoustically distinct surface forms onto the same lexical representation. In this paper, we investigate listeners' ability to compensate for one such alternation – word-final tapping in American English. Specifically, we present evidence to counter Ranbom, Connine and Yudman's (2009) claim that context is irrelevant in the recovery of tapped forms in American English.

2 Tapping in American English: Environments and frequency

In American English, /t/ and /d/ can surface phonetically as a tap (also called a flap) [ɾ] in a number of environments. In word-medial environments, /t/ and /d/ are almost invariably produced with a tap intervocalically in post-tonic contexts (e.g. *water*). Overall, spontaneous speech corpora (Patterson & Connine 2001) as well as production studies in the lab (e.g. Herd, Jongman & Sereno, 2010; Zue & Laferriere, 1979) show that a tap occurs between 76% of the time for /t/ words and 99% of the time for /d/ words in word-medial contexts, making them the most frequent variant of /t/ and /d/ in this environment.

In contrast, in word-final position, a tap occurs only optionally, and thus, less frequently. This most often happens when the following word begins with an unstressed vowel, as in *again* or *in* (Kahn, 1980; Oshika, Zue, Weeks, Neu & Aurbach, 1975; Turk, 1992). Across analyses of production studies and speech corpora, the frequency of the [ɾ] variant of /t/-final words in non-utterance-final, word-final context ranges from a low 5.6% (Herd et al., 2010; see also Byrd, 1994; de Jong, 1998) all the way up to 70% (Ranbom et al., 2009). The difference in tapping rate between these studies may be attributed to the fact that Ranbom et al. (2009) analyzed a corpus of spontaneous speech (Switchboard corpus; Godfrey & Holliman, 1997) while Herd et al. (2010) asked speakers to produce the target word in a sentence frame, *say X again*. It is unclear whether or not participants in the Herd et al. (2010) study inserted a prosodic break after the target word, a context in which tapping is usually blocked (de Jong, 1998). Interestingly, Ranbom et al. (2009) found that [ɾ]s derived from /t/ words even surfaced in supposedly non-licensing positions (e.g., preceding a consonant) 17.6% of the time! So it seems that the occurrence of a surface [ɾ] is only partially blocked by an inappropriate phonological context. Ranbom et al. (2009) did not investigate taps derived from /d/. Overall, the frequency of tapping of /d/-final words has been less investigated. One study by Herd et al. (2010) found that across a word boundary, /d/s were produced as taps 24.2% of time (compared to 5.6% of time for /t/ words in the same context). This suggests that /d/s might be produced as taps more frequently than /t/s across a word boundary. Given the variability of occurrence of word-final taps, we investigated how well tapped variants in word-final position activate the lexical representation of a target word.

[*] This research was funded by UCLA COR Faculty Research Grant (2012-13) to MS. We would like to especially thank Jesse Harris, Andrea Weber and Kie Zuraw for comments on a previous draft, Alice Twu for helping with the running of experimental participants, and Lauren Winans for recording stimuli. We would also like to thank members of the UCLA Phonetics Lab, audiences at National University of Singapore, Laboratory Phonology 2014 in Tokyo and AMP 2014 at MIT for comments and discussion, and the comments of three anonymous reviewers from AMP 2014.

© 2014 Adam J. Chong and Megha Sundara
Proceedings of *AMP 2014*
Completed February 27, 2015

3 How do listeners encode variation?

In research on spoken word recognition, most researchers draw a distinction between a canonical form, i.e., the citation form, and regular non-canonical variants (e.g., Gaskell & Marslen-Wilson, 1996), which are predictable from context. Converging evidence from priming and lexical decision tasks shows that a canonical form activates the lexical representation of the word even in contexts where the canonical form is not the most frequent variant (Pitt, Dilley & Tat, 2011; Ranbom & Connine, 2007; Sumner & Samuel, 2005). Thus, canonical productions seem to have a privileged status regardless of actual production frequency or context. For example, even word-medially, an environment that favours [ɾ], American English listeners label *butter* with a canonical [t] as a 'word' over 97% of the time (94% for *butter* with [ɾ] realization; Pitt et al., 2011; see Ranbom & Connine, 2007, for similar results on nasal flaps). Similarly, word-final [t]s successfully facilitate performance on a lexical decision task in contexts where [t] is the most common variant (e.g. pre-consonantally, 99% correct) and also where it is not (e.g. pre-vocalically, 100% correct; Ranbom et al., 2009). Taken together, these results provide robust evidence for the privileged position of canonical forms in spoken word recognition.

While context does not seem to affect the degree to which canonical variants activate a target representation, it does affect the degree to which non-canonical variants activate target representations. In the case of variants of medial /t/, research shows that non-canonical [ʔ] and [ɾ] variants are classified as words (e.g. in *witness*) only in contexts that license it. For example, in a context that favors [ʔ] (e.g. in *witness*), the [ʔ] variant was classified as a word 94% of the time in contrast to other non-canonical forms (e.g. [ɾ]: 21%; Pitt et al., 2011). Note that in production data, a word like *witness* is almost never produced with a tap. Contrastively, in a context that favors [t] (e.g. *pistol*), a [ʔ] variant is only classified as a word 18% of the time. The ability of a non-canonical variant to prime a target word thus seems to be closely linked to the frequency with which the variant form of that word appears in a specific context (Connine, Ranbom & Patterson, 2008; Ranbom & Connine, 2007; see also Pitt, 2009 for similar results for variants with deleted medial /t/s).

Providing an account for context effects in the recognition of non-canonical variants is a challenge for models of word recognition. In the class of abstractionist models, there is typically one lexical representation for a word and variant pronunciations are linked to the underlying lexical representation via some sort of a generalization (see Ernestus, 2014 for an overview). One well-known single representation abstractionist model is the inference model of spoken word recognition (Gaskell & Marslen-Wilson, 1996; 1998; Gaskell, 2003). In this model, listeners use their language experience to effectively undo phonological processes like assimilation using their knowledge of the contexts tied to specific variant forms. A similar context-dependent, but probabilistic model proposed by Mitterer (2011) posits that listeners calculate the likelihood that a word has been uttered given a certain input and context, i.e. *p*(word|evidence). In the case of place assimilation, this optimal perception account predicts unequal activation of a lexical item like "lean" given different inputs, *p*("*lean*"|"lean bacon") > *p*("*lean*"|"leam bacon") > *p*("*lean*"|"leam gammon"), with the activation of "*lean*" being higher when the variant form "leam" is produced before a viable ("bacon") versus unviable ("gammon") context.

Against the large body of evidence for context effects in word recognition, one particular finding has proven problematic for context-dependent abstractionist models in which there is only one lexical representation for a given word. It comes from a study investigating the role of context in the recovery of tapped word-final coda /t/s (Ranbom et al., 2009). Ranbom et al. auditorily presented listeners /t/-final target words in a sentence context that resulted in either a [t] (followed by a prosodic break; e.g. *For those of you who would like to* ***eat, early*** *lunch will be served*) or a tap (followed by a vowel onset; e.g. *For those of you who would like to* ***eat early****, lunch will be served*). At the offset of the target word, subjects were shown the printed target word on a screen (e.g. *eat*) and asked to make a lexical decision. Ranbom et al., found that overall, listeners were faster at recognizing a target word when presented with a canonical [t] than with a non-canonical [ɾ], replicating the advantage of the canonical form. Crucially, subjects were nearly at ceiling in recognizing tap variants in the appropriate (97%) and inappropriate context (98%). There were also no differences in reaction time between word-final [ɾ] productions in the licensing (*M* = 684 ms) versus non-licensing contexts (*M* = 682 ms). Therefore, context did not seem to play a role in the recognition of words produced with non-canonical [ɾ] variants.

Based on the absence of context-effects and comparable efficacy of canonical and non-canonical [ɾ] variants in activating target lexical representations, Connine and colleagues propose that taps, like canonical /t/s, are directly represented in the lexicon (Connine, 2004; Patterson & Connine, 2001; Ranbom & Connine, 2007; Ranbom et al., 2009; see also McLennan, Luce & Charles-Luce 2003, 2005). Connine and colleagues argue that because taps occur even in contexts where they are not licensed (17.6% pre-consonantally), the statistical relationship between the context and the tap variant is only weakly encoded, and thus, context plays no role in the recognition fo words with taps. Ranbom et al. further suggest that the general advantage of the canonical variant is due to its stronger representation in the lexicon compared to the [ɾ] variant by virtue of its greater overall frequency of occurrence. Ranbom et al. therefore argue for a *multiple* abstract representational account. Crucially, unlike in an account with just one lexical representation where phonological processes are "undone" in light of a phonological context, Ranbom et al. argue that the surface [ɾ] variant is mapped directly to a stored lexical representation with a [ɾ] variant.

Ranbom et al.'s failure to detect differences between tap variants presented in appropriate and inappropriate contexts is problematic for abstractionist models of word recognition with single representations, like the inference account (Gaskell & Marslen-Wilson, 1996) and the optimal perception account (Mitterer, 2011; also see Norris & McQueen, 2008). Although tap forms occur in ostensibly inappropriate contexts (17.6%), they are nonetheless much more frequent in an appropriate context (70%). Given input frequency in different context, both of these accounts predict that recognition of target words produced with tap forms should be worse in a non-licensing context compared to a licensing one.

Additionally, the comparable efficacy of the tap form and the canonical form is unexpected given the pervading evidence that word recognition is a gradient process. Previous work on the impact of within-category differences in Voice Onset Time (VOT) on word recognition has shown that listeners demonstrate gradient sensitivity even to small acoustic differences (Andruski, Blumstein & Burton, 1994; McMurray, Tanenhaus & Aslin, 2002). More recently even young children have been shown to be sensitive to the degree of phonological mismatch between a surface form and a target lexical representation, showing a proportional decrease in lexical activation as the degree of mismatch increases (White & Morgan, 2008). Given the gradient nature of word recognition, the fact that a perceptually-distant variant like a tap is not disruptive to word recognition, particularly when presented in an unviable context is surprising.

Before conceding that the processing of perceptually-distant variants like the tap in word-final position is indeed different from all other non-canonical variants investigated thus far, we consider a methodological explanation for Ranbom et al.'s findings. While the sentence context preceding the target word was neutral in Ranbom et al.'s design, the context following the target word was semantically biasing (e.g. *For those of you who would like to eat early, lunch will be served*). Thus, it is possible that listeners responded based on the following semantic context alone (e.g. *lunch*), accounting for the near ceiling performance on tap variants (>98%), as well as the failure to find context effects. In view of this methodological confound in Ranbom et al.'s experiment, we presented listeners with a non-biasing semantic context to revisit the importance of phonological context in the processing of tapped variants. We also extend the investigation to the effect of context on the recovery of tap variants of /d/.

4 Experiment 1

As in Ranbom et al., we investigated adult native English speakers' recognition of /t/-final words. However, unlike Ranbom et al. (2009), we did not present targets in a semantically-biasing context. This was done to ensure that subjects were not at ceiling in their responses. Instead, listeners were presented with variant forms in two semantically-neutral sentence frames: *Click the word X now!* or *Click the word X again!*

Specifically, listeners heard target words with stops (e.g. [bæt]) or taps (e.g. [bæɾ]). All words were presented in the 'now' and 'again' context. Listeners were expected to choose between two written labels on a computer screen – the printed text of the target word (*bat* in all the cases above), and a distractor - a series of "XXX" matched for length to the target word. Like in Ranbom et al., we cross-spliced target words to create a mismatch between the conditioning environment and the variant form. Additionally, we included 1-feature place mispronunciations (e.g. [bæp]) of the target words as well. We were specifically interested in how tap labels presented in mismatch contexts compared with mispronunciations. Lastly, phonologically dissimilar labels (e.g. *fish*) for target words were also included to encourage listeners to choose the distractor "XXX". These labels were a maximal mismatch to the surface form of the target. The choice to use a series of XXXs as a distractor was motivated by two factors. Firstly, in experiments that use eyetracking in the visual world with text (McQueen & Viebahn 2007), the usual visual set-up includes the target word and a number of real word competitors that are phonological *and* orthographic neighbors. In the present case, however, it was difficult to find competitor items for our target words that fit both this criteria (e.g. *back* is a phonological neighbor of *bat* but not an orthographic neighbor). Including such items might introduce extraneous effects based on differences in the orthographic form.

Moreover, we were primarily concerned with whether or not a mismatch context has an effect on the recognition of a tapped /t/-word. While one can test this using lexical competitors, it is unclear what the competitor should be with this sort of allophonic variation. Previous investigations using eyetracking have focused primarily on alternations between different phonemes in a given language (e.g [m] for /n/ in *lean bacon*). In the present case, the tap is an allophone and is not contrastive. Therefore, in this task, listeners were not expected to make a choice between two words, but rather decide whether or not the word they heard was the word on the screen. They were instructed to click on the XXXs if what they heard was not the word shown on the screen. We rationalized that the probability with which they selected the target word would be a measure of how *good* a fit the audio stimuli was for a given target lexical representation. If the audio stimulus is a good match, listeners should click on the target word, if not, they should reject the target word and click on the XXXs (i.e. indicating that what they heard was not the word on the screen).

4.1 *Participants* Twenty participants (16 F; Mean age = 20.3) were recruited via the University of California, Los Angeles Psychology Subject Pool. Subjects received course credit for participation. All identified themselves as native speakers of American English.

4.2 *Audio stimuli* 30 monosyllabic /t/-final words and 30 monosyllabic /d/-final words were used as auditory primes. For each word type, 12 were produced with a stop, 12 with a tap and 6 with a 1-feature place mispronunciation. Half the stops and taps were produced in the match, and the other half in the mismatch context (6 in each). All the target words were of the type (C)CVC where the vowel could be rhotacized, since coda /t/s undergo tapping only when preceded by a vowel. Additionally, we selected 12 monosyllabic non-/t/-final words that were phonologically dissimilar to the target words (e.g. *thief* for *mat*). Trials with phonologically dissimilar words were included to encourage participants to choose the distractor. Finally, we included 20 filler monosyllabic words which did not end with a /t/. This was done to ensure that listeners did not generate expectations regarding the stimuli. In sum, there were 92 test items.

The audio stimuli were digitally recorded by a female native speaker of American English. Target words were recorded in two sentence frames: (1) *Click the word X now!* or (2) *Click the word X again!* These two sentence frames were used since we wanted to provide an appropriate context for the phonological alternation between [t]/[d] and [ɾ]. Sentence (1) prompted the production of canonical [t] and [d]. In the phonological literature it is generally accepted that /t/ and /d/ are produced as [ɾ] following a [-consonantal] segment (i.e. vowels and glides) and preceding an unstressed vowel (e.g. Turk, 1992; Kahn, 1980; de Jong, 1998). Using sentence frame (2), therefore, provided the appropriate phonological context for tapping to occur. We further made sure that there was no prosodic break following the target word since this has been known to disrupt the occurrence of taps.

The closure duration of the target words were measured in Praat (Boersma & Weenink, 2011) and were graphed to confirm that the durations of the canonical [t], and the tap formed a bimodal distribution, following the method of classification used in Herd *et al.* (2010). Mispronunciation of the target words always involved one-feature coda mispronunciations and always contained the same vowel as in the target [t]/[d] or tap-final word (e.g. *bap* for *bat*). Both mispronunciations and phonologically dissimilar novel items were produced in both sentential frames and counterbalanced across subject groups. All sentences were recorded at 44,100 Hz using PCQuirer (Scicon R&D, Los Angeles, CA) and a Shure SM10A microphone. These were played back at a comfortable listening level of 75 dB over 3M Peltor noise-cancelling headphones.

In addition to canonical [t]/[d] and non-canonical [ɾ] presented in the correct licensing context, we cross-spliced target words and embedded them in the inappropriate phonological context, thereby producing a mismatch in the variant and the context. Target words were excised from their original context. The right boundary for segmentation was after the offset of aspiration noise and onset of the vowel for stop-final and tap-final words respectively. Target words were then spliced into a mismatch context and the intonation of the resulting stimuli item was resynthesized from the original stimuli item using the PSOLA algorithm in Praat (Boersma & Weenink, 2013).

4.3 *Visual stimuli* The visual targets were printed words presented on a 21.5 inch display monitor. These were created in Adobe Photoshop in the font Times New Roman with a point size of 80. These were then saved as portable network graphics (.png) files. These images were positioned with their center points being the center of each half of the screen. The distractor texts displayed were printed 'X's the number of which matched the number of characters of the printed target word. Visual targets were always real words. In the stop, tap, mispronunciation and novel conditions, the visual target was always a /t/-final word or /d/-final word. On the filler trials, there was always a match between the audio stimuli (a non-/t/ or non-/d/-final word) and printed target.

4.4 *Procedure* Participants sat on a chair in a 4-sided booth facing a 21.5-inch Asus display screen with a 2-ms refresh rate and an SR Eyelink 1000 (SR Research, *Mississauga, Canada)* sampling at 1000 Hz. They were seated between 500 and 600 mm away from the eyetracking camera (*M* = 562, *s.d.*=21). Speech stimuli were presented over 3M Peltor noise-cancelling headphones. In this experiment, participants' fixations were recorded by the Eyelink system but these data were not analyzed. Each trial ended when the participant made a response and therefore the variability in response time made it impossible to find a uniform time window of analysis for the eyetracking data. Thus they were not analyzed.

At the beginning of each trial, participants saw a crosshair in the center of the screen, after which the audio stimulus started playing. The visual target and distractor was then presented at the onset of the coda consonant of the target word, and the mouse cursor always appeared at the center of the screen at the onset of visual stimulus presentation. Subjects were instructed to keep their hand on the mouse throughout the entire experiment and click on the word they heard. They were also instructed to ignore the meaning of the following context (*now* or *again*). Each experimental session started with three practice trials in order to familiarize participants with the task. Practice trials consisted of a correctly pronounced target word (*place*), a dissimilar label for a printed target word (*groom* when the auditory prime was *farm*) and a mispronounced auditory label (*cug* for *cub*, with the visual target *cub*). After the practice trials, the test phase commenced. Participants were assigned to one of 5 experimental groups. The assignment of target words to the stop, tap, or mispronunciation condition was counterbalanced across each experimental group; the novel and filler trials were constant across subjects. Moreover, sentential context was counterbalanced for mispronounced trials across subjects. A complete list of stimuli and conditions is presented in Appendix A.

The order of presentation of trials and the side of presentation was randomized between trials in Experiment Builder (SR Research, Mississauga, Canada), such that the target word appeared equally on each side. Each experimental session lasted about 8 minutes.

4.5 *Results: Accuracy* As is typical, trials with response times less than 200 ms or greater than 2000 ms were removed from the data (68 trials in total, ~4% of the data). Figure 1 shows percent word responses were calculated for each condition (L) and by context for stops and taps (R). As expected, with both word types, listeners selected the target word much less often when they heard a mispronunciation (/d/: *M* = 49.8%; /t/: *M* = 31.7%) than when they heard a stop or tap label. While listeners selected the target equally often in the stop condition for both /d/ (*M* = 99.5%) and /t/ words (*M* = 97.5%), they selected the target more frequently for tapped /d/ (*M* = 97.8%) than tapped /t/ forms (*M* = 86.8%). Finally, as expected, when listeners heard a phonologically-dissimilar label for a /t/-final target word, they never selected the target word. Since listeners were at floor in their responses in the novel condition, we performed an empirical logit transformation on the response data following (Barr, 2008) and conducted the analysis using mixed effects logistic regression using the *lme4* package (Bates, Maechler & Dai, 2008) in R (R Core Development Team, 2008). The significance of factors was calculated using the Anova() function of the *car* package (Fox & Weisberg, 2011) in R (R Core Development Team, 2008).

The analysis revealed a significant effect of Condition ($\chi^2(3) = 1042.97$, $p < 0.001$). There was also a significant effect of Word type (/t/ vs. /d/: $\chi^2(1) = 23.56$, $p < 0.001$) with more word responses for /d/-final words overall, and a significant two-way interaction of Condition and Word type ($\chi^2(3) = 18.00$, $p < 0.001$).

Subset comparisons were conducted for /t/ and /d/ words separately. In each case, there was a significant effect of Condition (/t/: $\chi^2(3) = 1296.3$, $p < 0.001$; /d/: $\chi^2(3) = 274.18$, $p < 0.001$). Post-hoc pair-wise comparisons using the glht() function of the *multcomp* package (Horton, Bretz & Westfall, 2008) in R revealed that for /t/ words all conditions differed significantly from each other (Stop vs. Tap: $p < 0.001$; Stop vs. MP: $p < 0.001$; Stop vs. Novel: $p < 0.001$; Tap vs. MP: $p < 0.001$; Tap vs. Novel: $p < 0.001$; MP vs. Novel: $p < 0.001$). On the other hand, for /d/ words, all conditions differed significantly from each other (all $p < 0.001$), except for the stop and tap condition ($p = 0.84$).

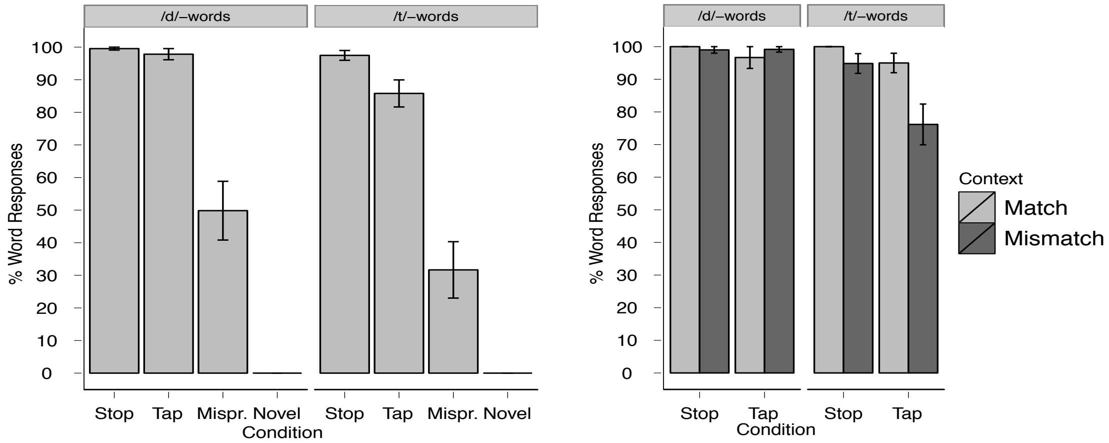

Figure 1. Proportion of target word responses. (L): Overall responses by word type. (R): Responses for stop and tap condition by context by word type

In sum, listeners were more likely to pick the word response in the stop condition than in the tap condition, providing evidence for the advantage of the canonical form, although they only showed this pattern with /t/ words but not /d/ words. Additionally, percent word responses in the mispronunciation condition were lower than that in the stop condition but higher than that for the novel condition. Thus, listeners showed a graded sensitivity to phonetic detail in word recognition replicating previous findings on listeners' sensitivity to mispronunciations in word recognition (e.g. Swingley, 2009). This is important for task validation as we chose to go with XXXs rather than a real word competitor in our visual stimuli.

Next we analyzed the effects of context on percent word responses for the stop and tap conditions by word-type (Figure 1 R). For /t/-words, percent word responses for stops were over 90% in both contexts (Match: $M = 100\%$; Mismatch: $M = 94.8\%$). In contrast, percent word responses for taps were over 90% in the matching *again* context ($M = 95\%$) but not in the mismatched *now* context ($M = 76.2\%$). For /d/-words, however, percent word responses was similar across condition and context (~ 99%).

There was an overall advantage for the stop forms over the tap variants confirming the advantage for the canonical form ($\chi^2(1) = 6.13$, $p = 0.01$). There was also a significant effect of Context ($\chi^2(1) = 9.20$, $p = 0.002$), with more word responses in the match than mismatch context and a significant effect of Word type ($\chi^2(1) = 7.98$, $p = 0.005$) with overall more word responses with /d/ words that /t/ words. The two-way interaction of Condition and Word type was significant ($\chi^2(1) = 6.25$, $p = 0.01$) as was context by word type ($\chi^2(1) = 17.40$, $p < 0.001$). The interaction of Condition and Context was not significant ($\chi^2(1) = 0.23$, $p = 0.63$). Importantly, however, there was a significant three-way interaction of Condition and Context and Word type ($\chi^2(1) = 4.92$, $p = 0.02$). By-item analyses showed the same pattern of results, although the two-way interaction of condition by word-type was not significant ($\chi^2(1) = 2.52$, $p = 0.11$). Given the significant three-way interaction, we did a subset analysis for /t/ and /d/ words separately.

For /t/ words, there was a significant main effect Condition ($\chi^2(1) = 12.65$, $p < 0.001$) with more word responses in the stop condition than tap condition. There was also overall, significantly more word responses in the match context than mismatch ($\chi^2(1) = 10.71$, $p = 0.001$). Crucially, there was also a significant two-way interaction of Condition and Context ($\chi^2(1) = 5.09$, $p = 0.02$). In planned comparisons of /t/ trials, performance on stop trials did not differ by context ($\chi^2(1) = 2.50$, $p = 0.12$), although performance on tap trials did ($\chi^2(1) = 14.85$, $p < 0.001$) with significantly more word responses in the match than mismatch context. Performance in the match context did not differ by condition ($\chi^2(1) = 2.17$, $p = 0.14$) although performance in the mismatch context did ($\chi^2(1) = 26.88$, $p < 0.001$), with more word responses in the stop condition than tap condition. For /d/ words, on the other hand, no factors were significant (Condition: $\chi^2(1) = 0.20$, $p = 0.65$; Context: $\chi^2(1) = 0.0006$, $p = 0.98$; Condition and Context: $\chi^2(1) = 1.67$, $p = 0.20$).

The results of Experiment 1 provide evidence for the importance of context in the processing of tap variants of /t/ in American English. Subjects made more word responses when tap variants of a /t/ word were presented in the match context compared to the mismatch context. Additionally, the target response rate for tap /t/ variants in a mismatch context was much higher than the target response rate for mispronounced words (74.2% vs. 32.5%). Taken together these findings indicate that there is a cost to processing tap variants of /t/ in a mismatch context, although listeners do not treat these as mispronunciations. In contrast, a mismatch context did not matter for tap variants of /d/, suggesting that phonological context might not be entirely necessary for the recovery of some phonological variants.

Recall that subjects' percent word responses in the stop condition were at ceiling (100%) in the match context in this experiment. In Experiment 2, we used eye-tracking to confirm the role of context in the recognition of taps, and also to probe into the effects of context on the recognition of canonical stops. In addition to yielding a time-course of recognition for words with variant forms, the dense sampling of eye-tracking data makes ceiling effects unlikely. Moreover, in Experiment 1, only the mismatch context (*now* for taps and *again* for stops) was cross-spliced. Stimuli in all other conditions were presented *in situ.* Recall that this was done to be consistent with the previous study by Ranbom et al. However, it is possible that splicing the stimuli degraded the quality of the mismatch stimuli such that there was a cost resulting in lower word recognition rates in the mismatch contexts. In Experiment 2, we cross-spliced stimuli in both match and mismatch contexts. If the pattern of results in Experiments 1 are truly the result of context, then we expected to replicate the same pattern of results in Experiment 2 using a different methodology, viz., eyetracking.

5 Experiment 2

In Experiment 2, we investigated the extent and time course of activation of the target word when listeners are presented with the canonical and tap variants in a match compared to a mismatch context. To circumvent the ceiling effects observed in the accuracy data obtained in Experiment 1, in Experiment 2, we implemented the cross-modal priming paradigm using eye-tracking. Unlike in Experiment 1, stimuli in all conditions were cross-spliced in Experiment 2. Like in Experiment 1, we implemented the task using printed words (McQueen & Viebahn, 2007). Unlike in Experiment 1, all sentences with the auditory primes in Experiment 2, ended with a filler sentence (*'Can you find it?'*/ *'Have you found it?'*). These were included to help maintain participants' gaze on the screen. Additionally, we did not require subjects to make click responses.

5.1 *Participants* Another 20 native speakers of American English (18 F; mean age = 20.3) were recruited via the UCLA Psychology Subject Pool. Subjects received course credit for participation. Four additional subjects were tested but were excluded because they were not native speakers of English (2), poor calibration (1) and computer failure (1).

5.2 *Audio and visual stimuli* In Experiment 1, only stimuli for the mismatch context were cross-spliced. In Experiment 2, we spliced in the context following the target word in all conditions. The words "now" and "again" were recorded in isolation by the same speaker who produced our target sentences. These tokens were then spliced into each target sentence, replacing the original following context. The intonation contour was then smoothed manually in Praat (Boersma & Weenink, 2011) to make each stimuli item sound more natural. In order to ensure that subjects maintained their gaze on the screen, we added an additional sentence following each test sentence. Two sentences were recorded by the same speaker who recorded the target sentences: (1) *Have you found it?* and (2) *Can you find it?* After the offset of the target sentence, there was a 350 ms period of silence after which one of these two sentences played. Half of the test trials contained sentence (1) and half contained sentence (2). Thus, each trial was longer in duration.

These modifications also allowed us to have a fixed time-window for analysis. The same visual stimuli used in Experiment 1 were used in Experiment 2.

5.3 *Procedure* The experimental procedure was largely identical to that used in Experiment 1 except in one way. Instead of clicking on the word on the screen, participants were instructed to just look at the word that they heard. They were seated between 500 and 600 mm away from the eyetracking camera (*M*=564, *s.d.*=16.89) and their eye movements were recorded using the arm-mount remote configuration of the SR Eyelink 1000 (SR Research, Mississauga, Canada). After an initial calibration, drift correction was carried out at the beginning of each trial and participants were re-calibrated as needed.

As in Experiment 1, subjects were centered at the beginning of each trial. Auditory stimuli with the primes then played and the visual target was displayed on the computer screen at the onset of the coda consonant of the target word. The visual target remained on the screen until the end of each trial. Each trial lasted approximately 4 seconds. The experiment lasted approximately 10 minutes.

5.4 *Results* Two interest areas (Target and Distractor) were set *a priori* at 350 x 240 pixels around the printed text on the screen. The interest areas were larger than the text images themselves so as not to penalize fixations which, although not on the text itself, were nonetheless in the right area of the screen. Looking behavior was sampled at 2ms intervals by the Eyelink system. Taking into account the time it takes for initiating a saccade in response to auditory stimuli, the initial window started 200ms (Matin et al. 1993) after the onset of the coda consonant and continued to 2000ms after the onset of the coda consonant (i.e. window of 1800ms).

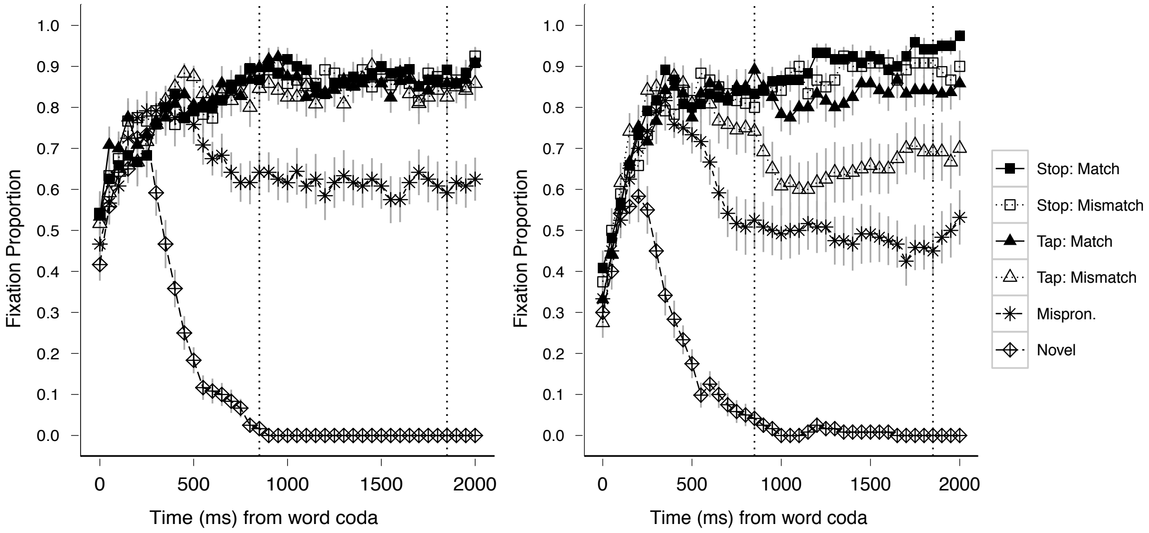

Figure 2. Time course of fixations to target word by condition and context. (L) /d/ words; (R) /t/ words. Dotted vertical lines delineate time window of analysis (850ms-1850ms)

We first ran non-parametric statistics to establish a time window for analysis (Maris & Oostenveld, 2007, Von Holzen & Mani, 2012; Tsuji, Mazuka, Cristia & Fikkert, 2013). Once a time window was established, fixation proportions were submitted to a Growth Curve Analysis (GCA: Mirman, Dixon & Magnuson, 2008; Mirman, 2014). Here, we only report on the comparison between the stop and tap conditions for both /t/ and /d/ words. Given, the significant effect of context for tap variants of /t/ words and not /d/ words in Experiment 1, we established the time window in which fixations deviated reliably in the match and mismatch context in the tap condition for /t/ words. Target fixations in the match and mismatch context deviated significantly from each other between 850 ms and 1850 ms (cluster *t* statistic = 59.64, Monte Carlo p = 0.006), with more looks to the target in the match than mismatch context. This time window was then used for both /t/ and /d/ word comparisons.

The independent variables of Context, Condition and Word Type were sum coded. Here we only report the effects of each factor on the intercept, which indicates the average fixations over the entire time-window. There was a significant effect of Condition ($\chi^2(1) = 14.63$, $p = 0.001$), Context ($\chi^2(1) = 9.58$, $p = 0.002$), and Word type ($\chi^2(1) = 8.81$, $p = 0.003$). Moreover, there were significant interactions between Condition and Context ($\chi^2(1) = 5.30$, $p = 0.004$), Condition and Word Type ($\chi^2(1) = 9.28$, $p = 0.002$), and Context and Word Type ($\chi^2(1) = 4.80$, $p = 0.03$). The three-way interaction of Condition by Context by Word Type was marginally significant ($\chi^2(1) = 3.73$, $p = 0.053$).

Due to the marginal three-way interaction, we conducted separate analyses for /t/ and /d/ words. For /t/ words, there was a significant effect of both Condition ($\chi^2(1) = 14.15$, $p = 0.002$), and Context ($\chi^2(1) = 9.03$, $p = 0.003$) on target fixations. There was also a significant interaction ($\chi^2(1) = 1.01$, $p = 0.31$). For /d/ words, no effects were significant (Condition: $\chi^2(1) = 14.15$, $p = 0.002$; Context: $\chi^2(1) = 0.82$, $p = 0.36$; Condition by Context: $\chi^2(1) = 0.14$ $p = 0.70$). The eyetracking results paralleled the accuracy results from Experiment 1. Within /t/ words, the effect of context did not matter for trials in the stop condition ($\chi^2(1) = 1.27$, $p = 0.26$). In the tap condition, however, listeners looked significantly more at the target word in match vs. mismatch context ($\chi^2(1) = 11.32$, $p < 0.001$). In the match context, listeners looked significantly more to the target word when produced with a stop than a tap ($\chi^2(1) = 5.03$, $p = 0.025$). This was similarly the case in the mismatch context ($\chi^2(1) = 12.61$, $p < 0.001$).

To summarize, we largely replicated the crucial findings from Experiment 1. For /d/ words, there was no cost of a mismatch context for tap variants. In contrast, listeners looked less to the target word when tap variants of /t/ words were presented in a mismatch context. Context did not matter for the processing of the canonical stop variant. Finally, although we found that listeners did not differ in responses when they heard a stop or tap variant in a match context in Experiment 1, there was a cost to word recognition for the tap even in the match context in Experiment 2. We think this may have arisen because cross-splicing effects are more deleterious for the recognition of taps than canonical stops.

6 Discussion and Conclusion

The main goal of the study was to investigate the importance of context in the recognition of words produced with a phonological variant, namely word-final tap variants of /t/ and /d/ in American English. In two experiments we investigated the ability of listeners to recognize target words produced with a canonical stop (/t/ or /d/) and a regular tap variant. In Experiment 1, listeners were tested on a word identification task. Then in Experiment 2, listeners were tested using an eye-tracking task implemented using a text variant of the visual world paradigm.

Firstly, we found that context does not matter when processing canonical variants. We found no difference in lexical activation when listeners were presented with the canonical form in the match or the mismatch context. We saw this in accuracy data (Experiment 1) as well as target fixation proportions (Experiments 2). This replicates the robust finding that canonical forms, despite differences in actual production frequency across contexts, are privileged (McLennan et al, 2003; 2005; Pitt, 2009; Pitt et al, 2011; Ranbom et al., 2009; Ranbom & Connine, 2007; Sumner & Samuel, 2005; Tucker, 2011).

The crucial finding was that context matters for the processing of tap variants of /t/ but not for tap variants of /d/. Contra Ranbom et al. (2009), a tap variant of /t/ presented in a mismatch context has a disruptive effect on word recognition. Subjects were less accurate at recognizing words with tapped /t/s (Experiment 1) and looked less to the target word when the tapped /t/s was presented in the mismatch context compared to a match context (Experiment 2). This finding has important implications for a model of phonological representation and processing. There is clear evidence that, in the absence of semantically biasing cues in our study, phonological context is important in the processing of [ɾ] variants of word-final /t/ in American English. This calls into question the central motivation for Ranbom et al.'s (2009) multiple abstract representational model. Ranbom et al. (2009) proposed a multiple representational model precisely to account for the lack of context effects found in the processing of word-final tap variants. They argued that because listeners hear taps in environments that ostensibly do not license it, they learn not to rely on context when processing tap variants. This is certainly consistent with Ranbom et al's results since they found that tap variants facilitate lexical access of a /t/ target word to the same extent as stops, even when presented in an inappropriate environment (e.g. preceding a consonant or preceding a prosodic break). These results led them to suggest that both the canonical stop variant and the non-canonical tap variant are encoded directly in the lexicon, with the stop variant being more strongly encoded due to its overall higher frequency of occurrence. In fact, the necessity of context in the processing of tap variants obviates the need for a multiple representation model; the negative impact of a mismatch context on the processing of non-canonical variants is already predicted by models with one lexical representation.

With /d/ words, however, subjects were equally accurate at recognizing the target word in Experiment 1 and looked equally to the target in Experiment 2 regardless of whether a tap variant of a /d/ word was produced in an appropriate phonological context. Importantly, American English listeners have been shown to be able to somewhat discriminate the difference between [d] and tap despite the close perceptually similarity (Boomershine, Hall, Hume & Johnson, 2008; see also de Jong, 1998). Therefore, the current result cannot be explained by an inability to discriminate between the two phones. Moreover, this finding is difficult to accommodate in models with single abstractionist representations like a traditional inference account (Gaskell & Marslen-Wilson, 1996). In the inference account, phonological variants presented out of context should not facilitate lexical access. Recall that despite being worse at recognizing a target word when presented with a tap variant of /t/ in a non-licensing context, listeners did not treat these forms as mispronunciations. That is, a tap variant of /t/ in a mismatch context continued to facilitate lexical access, just not to the same degree as in a match context. Moreover, there was no cost at all of a mismatch context for tap variants of /d/. Our results then provide evidence that a mismatch context does not have the same disruptive impact for all types of phonological variants. Instead, it is likely that the degree of perceptual deviation from the canonical form accounts for the gradient nature of word recognition, with taps and /t/s being more distant than taps and /d/s (Herd et al., 2010),

Taken together, our findings suggest that contextual information is less important when there is a closer perceptual match between a surface form and the lexical representation. Given a word recognition model that is sensitive to the perceptual match, it is thus possible to account for both the advantage of canonical forms regardless of context and the lack of context effects for tap variants of /d/ words. As argued previously by Gaskell (2003), perceptual similarity to the lexical representation is likely to mitigate processing costs associated with an unviable mismatch context (see also Norris & McQueen, 2008; Connine et al., 1997). In the case of canonical forms, a mismatch context is not likely to override the perceptual evidence for a given target word provided by bottom-up cues. Similarly, the close perceptual similarity between tap and /d/ might mitigate any processing cost a mismatch context might have. Consistent with this idea, a connectionist model trained to compensate for various strengths (incomplete to complete) of assimilation demonstrated negligible effect of an unviable context in the absence of assimilation (Gaskell, 2003). Incorporating the degree of perceptual deviation between the variant and the lexical representation into a model of spoken word recognition also allows for some testable predictions. Specifically, the magnitude of context effects should vary as a function of the perceptual distance between the variant and the underlying lexical representation.

In sum, through a series of word recognition experiments, we have shown evidence for context effects in the processing of word-final tapping of /t/ words but not /d/ words. The finding of context effects for /t/ words provides evidence against Ranbom et al. (2009), obviating the need to posit a multiple abstract representational account. Instead, our results are largely consistent with a model in which only one abstract lexical representation is present. Importantly, any such model must take into account both frequency of a variant in context *and* the perceptual distance between a variant and its lexical representation. Sensitivity to the perceptual similarity between a surface variant and the lexical representation allows us to account for both the lack of context effects for canonical forms (perfect match) and also tap variants of /d/ (highly perceptually-similar).

References

Andruski, Jean. E., Sheila E. Blumstein & Martha Burton (1994). The effect of subphonetic differences on lexical access. *Cognition, 52,* 163-187.

Barr, Dale (2008). Analyzing ‘visual world’ eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457-474.

Bates, Douglas, Martin Maechler & Bin Dai (2008). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-28. http://lme4.rforge.r-project.org/

Boersma, Paul & David Weenink (2011). Praat: doing phonetics by computer [Computer program]. Version 5.2.45, retrieved 20 July 2011 from http://www.praat.org/

Boomershine, Amanda, Kathleen Currie Hall, Beth Hume & Keith Johnson (2008). The influence of allophony vs. contrast on perception: The case of Spanish and English. In Peter Avery, Elan Dresher and Keren Rice (eds.), *Phonological contrast* (pp. 143-172). New York: Mouton de Gruyter.

Byrd, Dani (1994). Relations of sex and dialect to reduction. *Speech and Communication, 15,* 39-54.

Ernestus, Mirjam (2014). Acoustic reduction and the roles of abstraction and exemplars in speech processing. *Lingua*, 142, 27-41.

de Jong, Kenneth J. (1998). Stress-related variation in the articulation of coda alveolar stops: Flapping revisited. *Journal of Phonetics, 26*, 283-310.

Fox, John & Sanford Weisberg (2011). *An {R}companion to applied regression*. Thousand Oaks CA: Sage.

Gaskell, M. G. (2003). Modelling regressive and progressive effects of assimilation in speech perception. *Journal of Phonetics, 31,* 447-463.

Gaskell, M. Gareth & William D. Marslen-Wilson (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 144-158.

Gaskell, M. Gareth & William D. Marslen-Wilson (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 380-396.

Godfrey, John J., & Edward Holliman (1997). *Switchboard-1 Release 2.* Philadelphia: Linguistic Data Consortium.

Horton, Torsten, Frank Bretz & Peter Westfall (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*, 346-363.

Jaeger, T. Florian (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language, 59,* 434-446.

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

Maris, Eric & Robert Oostenveld (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177-190.

Matin, Ethel, K. C. Shao, & Kenneth Boff (1993). Saccadic overhead: information processing time with and without saccades. *Perception & Psychophysics, 53,* 372–380.

McLennan, Connor T., Paul A. Luce & Jan Charles-Luce (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29*, 539-553.

McLennan, Connor T., Paul A. Luce & Jan Charles-Luce (2005). Representation of lexical form: Evidence from studies of sublexical ambiguity. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 1308-1314.

McMurray, Bob, Michael K. Tanenhaus & Richard N. Aslin (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86,* B33-B42.

McQueen, James M. & Malte Viebahn (2007). Tracking recognition of spoken words by tracking looks to printed words. *Quarterly Journal of Experimental Psychology, 60*, 661-671.

Mirman, Daniel, James A. Dixon & James S. Magnuson (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59*, 475-494.

Mirman, Daniel (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: Chapman Hall/CRC.

Mitterer, Holger (2011). The mental lexicon is fully specified: Evidence from eyetracking. *Journal of Experimental Psychology: Human Perception and Performance* , *37*, 496-513.

Norris, Dennis & James M. McQueen. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*, 357-395.

Oshika, Beatrice T., Victor W. Zue, Rollin V. Weeks, Helene Neu, & Joseph Aurbach (1975). The role of phonological rules in speech understanding research. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-23*, 104-112.

Patterson, David & Cynthia M. Connine (2001). Variant frequency in flap production: A corpus analysis of variant frequency in American English flap production. *Phonetica, 58,* 254-278.

Pitt, Mark. (2009). How are pronunciation variants of spoken words recognized? A test of generalization to newly learned words. *Journal of Memory and Language, 61*, 19-36.

Pitt, Mark, Laura Dilley & Michael Tat (2011). Exploring the role of variant frequency in recognizing pronunciation variants. *Journal of Phonetics, 39*, 304-311.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org>.

Ranbom, Larrisa J. & Cynthia M. Connine (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language, 57*, 273-298.

Ranbom, Larissa J., Cynthia M. Connine & Elana M. Yudman (2009). Is phonological context always used to recognize variant forms in spoken word recognition? The role of variant frequency and context distribution. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1205-1220.

Sumner, Meghan & Arthur G. Samuel (2005). Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language, 52*, 322-338.

Swingley, Daniel (2009). Onsets and codas in 1.5-year-olds' word recognition. *Journal of Memory and Language, 60,* 252-269.

Tsuji, S., Mazuka, R., Cristia, A., & Fikkert, P. (2013). *A cross-linguistic study of the labial-coronal perceptual asymmetry: evidence from Dutch and Japanese infants*. Oral presentation at the 38th Boston University Conference on Language Development, Boston, USA.

Tucker, Benjamin V. (2011). The effect of reduction on the processing of flaps and /g/ in isolated words. *Journal of Phonetics. 39*, 312-318.

Turk, Alice (1992). The American English flapping rule and effect of stress on stop consonant duration. *Working Papers of the Cornell Phonetics Laboratory, 7,* 103-133.

Von Holzen, Katie & Nivedita Mani (2012). Language nonselective lexical access in bilingual toddlers. *Journal of Experimental Child Psychology, 113,* 569-586.

White, Katherine S. & James L. Morgan (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language, 59*, 114-132.

Withgott, Mary M. (1982). *Segmental Evidence for Phonological Constituents* (Unpublished doctoral dissertation). The University of Texas, Austin.

Zue, Victor W. & Martha Laferriere (1979). Acoustic study of medial /t,d/ in American English. *Journal of Acoustical Society of America, 66*, 1039-1050.