Title
Processing allophonic variants in the visual world paradigm

Permalink
https://escholarship.org/uc/item/5nj1j4qn

Author
Chong, Adam Junxiang

Publication Date
2013

Peer reviewed|Thesis/dissertation
Processing allophonic variants in the visual world paradigm

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Arts in Linguistics

by

Junxiang Adam Chong
This study investigates the ability of adult American English listeners to recognize familiar words produced with an allophonic variant. It reports the results of a visual world paradigm experiment. Subjects were presented with /t/- and /d/-final target words produced with a canonical stop, a tap or a one-feature mispronunciation. For both /t/ and /d/-final words, subjects fixated to the target object more in the canonical stop condition than in the mispronunciation condition. Crucially, subjects fixated equally at the target image in both the canonical stop condition and the non-canonical tap condition across
both word types. Further analysis, however, showed that listeners were faster at shifting their gaze to the target object when the label was produced with a stop than when it was produced with a tap. Our results suggest that both canonical stop and tap variants are viable labels for a given target word, although the canonical variant seems to be faster at activating a target lexical representation. The implications of these findings for models of spoken word recognition are discussed.
The thesis of Adam Chong is approved.

Robert Daland

Kie Ross Zuraw

Megha Sundara, Committee Chair

University of California, Los Angeles

2015
# Table of Contents

1. Introduction ................................................................................................................. 1  
   1.1 Tapping in American English: Environment and Frequency .................. 1  
   1.2 How do listeners encode variants? ............................................................... 4  

2. The study ....................................................................................................................... 8  
   2.1 Subjects ................................................................................................................. 10  
   2.2 Target words ........................................................................................................ 10  
   2.3 Audio stimuli ......................................................................................................... 12  
   2.4 Visual stimuli ......................................................................................................... 15  
   2.5 Apparatus and procedure ..................................................................................... 15  
   2.6 Results ..................................................................................................................... 18  
      2.6.1 Stops vs. Novel ................................................................................................. 21  
      2.6.2 Stops vs. MPs vs. Taps ................................................................................... 21  
      2.6.3 Timecourse of fixations ................................................................................. 23  

3. Discussion and Conclusion ........................................................................................... 26  

Appendix A: List of target and distractor items ............................................................ 33  

References ....................................................................................................................... 34
Acknowledgements

This work would not have come to fruition were it not for the help of a number of people. Firstly, I would like to thank my chair, Megha Sundara, for all the help, advice and thoughtful guidance she provided in the last year and a half. I am also indebted to Robert Daland whose Experimental Phonology course was the seed from which the idea for this study grew. I would also like to thank Kie Zuraw for her insightful comments and helpful discussion during various stages of this project.

This work has also benefited from comments and discussion from various members of the Phonetics Lab. For this I am grateful to Marc Garellek, Jamie White, Chad Vicenik, Robyn Orfitelli and the other members of the Phonetics Seminar group who listened and commented on the results of this work.

Starting a new experimental methodology is always a daunting task and this project would not have got off the ground at all without the technical help of Henry Tehrani who helped in the set-up of the new Eye-tracking lab space and Brian Nguyen. Brian, especially, was always willing to offer support in relation to any Eye-tracker related woes. The support staff members at SR Research also deserve thanks for being so forthcoming with help over email.
Lastly, I am grateful to other members of the UCLA Linguistics Department and my family and friends for their support through this endeavor.
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, i.e., allophonic, 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 the ability of listeners to recognize familiar words that are produced with an allophonic variant in a visual world paradigm.

1.1 Tapping in American English: Environment and Frequency

We are interested in the case of American English /t/ and /d/. In American English, both /t/ and /d/ can surface phonetically as a tap [ɾ] in a number of environments. Tapping most commonly occurs in word-medial environments when /t/ and /d/ are intervocalic and when the following vowel is unstressed (Oshika, Zue, Weeks, Neu & Aurbach, 1975; Zue & Laferriere, 1979; Kahn, 1980; Turk, 1992). For instance, in a word like water - where /t/ is intervocalic between a stressed vowel and unstressed vowel - a tap is considered obligatory (Turk, 1992). However, when medial /t/ occurs between two unstressed vowels, e.g., in the word provocat[ɾ]ive~provoca[r]ive, this change is optional and depends on various phonological and morphological factors (see Withgott, 1982).
Besides phonological factors, tapping in word-medial position is governed by morphology as well as word frequency. Lower frequency words are generally tapped less frequently than high frequency words and morphologically complex words are tapped less frequently than morphologically simple words, where the addition of a suffix places a word-final /t/ in a tapping environment. There is also an interaction between both factors, with low frequency morphologically complex words tending to be tapped less frequently than high frequency morphologically simple words (Patterson & Connine, 2001). In sum, spontaneous speech corpora (Patterson & Connine 2001) as well as production studies in the lab (e.g. Zue & Laferriere, 1979; Herd, Jongman & Sereno, 2010) show that a tap occurs between 76% and 99% of the time in word-medial contexts, making it the most frequent variant of /t/ 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 (Oshika et al., 1975; Kahn, 1980; Turk, 1992). However, tapping can also occur when the following word is stressed, suggesting that the stress conditions are more relaxed across lexical boundaries (Oshika et al., 1975). The optional occurrence of tapping in word-final /t/ is additionally variable because it is tied closely to prosodic factors. For instance, a pitch accent on the word as well as a prosodic break between it and the next word seem to disrupt the occurrence of a tap in word-final contexts (de Jong, 1998).
In keeping with this optional distribution, across analyses of production studies and speech corpora, the frequency of the [r] 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, Connine & Yudman, 2009). The difference in tapping rate between these two studies can 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) analyzed production frequencies from a production task in which speakers had 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. Interestingly, Ranbom et al. (2009) found that [r]s even surfaced in supposedly non-licensing position (e.g., preceding a consonant) 13.5% of the time! So it seems the occurrence of a surface [r] is only partially blocked by an inappropriate phonological context.

While the frequency of tapping in /d/-final words has not been as thoroughly investigated, in the only study that reports the frequency of both /t/-tap and /d/-tap alternations, /d/ has been reported to tap more frequently at 24.2% (Herd et al., 2010) where /t/ was only tapped 5.6% of the time. Thus, overall across all contexts, the canonical /d/ form is still the most frequent variant, as is the canonical [t] form (Ranbom et al., 2009) 73% compared to 27% [r]
variants). However, singleton [t] codas followed by a vowel-initial word, are produced a majority of the time (70%) with a [ɾ] (Ranbom et al., 2009).

Given the variability of occurrence, we use a visual world paradigm in this study to investigate whether the tapped variants in word-final position activate the lexical representation of a target word as well as canonical variants.

### 1.2 How do listeners encode variants?

Faced with multiple surface variants in the signal, previous research shows that adult listeners are able to recover the appropriate target word. Most researchers draw a distinction between a canonical form, i.e. the citation form, and regular non-canonical variants. 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; Sumner & Samuel, 2005; Ranbom & Connine, 2007). In fact, canonical productions seem to have a privileged status regardless of actual production frequency. For example, even word-medially, an environment that favours [ɾ], 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). Canonical forms are also better at priming lexical representations of a target word than non-canonical variants. In a series of experiments, Ranbom & Connine (2007)
found that canonical [nt] productions had an advantage over the nasal flap in facilitating lexical access, despite the higher production frequency of nasal flaps than [nt] in their corpus analysis (81.8% nasal flap vs. 15% [nt]).

While a low frequency of occurrence does not inhibit the ability of canonical forms to activate a given target word, it does affect the degree to which non-canonical forms are classified as words. Research shows that both [ʔ] and [ɾ] variants are classified as words (e.g. in witness); however, 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 a word appears with that variant form.

Similarly, a non-canonical nasal flap facilitates recognition of a word better when the nasal flap production is the more common form of the word (Ranbom & Connine, 2007). That is, production frequency plays a role in the processing of non-canonical variants (see also Connine, Ranbom & Patterson, 2008). Based on the efficacy with which non-canonical variants activate a target lexical representation, some researchers have argued that non-canonical forms are directly represented in the lexicon (Patterson & Connine, 2001; Connine, 2004; Ranbom &
Connine, 2007; however, see McLennan, Luce & Charles-Luce 2003, 2005 for an alternative in which surface taps are mapped onto underlying /t/.

These two factors, the advantage of the canonical form and the importance of production frequency for non-canonical variants, have also been demonstrated in relation to word-final /t/s where the alternation is more optional than in word-medial position (Ranbom et al., 2009). First, Ranbom et al. found that [t] variants prime target words consistent with previous research showing an advantage for the canonical form. Second, even in their respective licensing contexts, the canonical form primes a target word more than [ɾ] variants. A processing account like the inference model (Gaskell & Marslen-Wilson, 1996 1998) cannot explain why the canonical form is better at activating underlying forms, even when the tap variant is in the appropriate licensing context.

Third, Ranbom et al. (2009) found significant priming effects for word-final [ɾ] productions even in inappropriate licensing contexts (e.g. preceding a pause). Again, this result cannot be explained by an inference account, because an inference account predicts that word recognition fails when a variant is presented in an inappropriate context.

In order to explain the priming effects found for the [ɾ] variant even in contexts in which they are not licensed, they propose firstly that both variants are directly represented in the lexicon, with word recognition being insensitive to phonological context. Consistent with
this idea, in their corpus study, they found that [r] productions occurred even in contexts where they are not licensed. Because of this, they assert that the statistical relationship between the context and a particular variant is only weakly encoded since the listener, through language experience, has learnt that the following context poorly predicts the variant produced. Ranbom et al. (2009) further suggest that the general advantage of the canonical variant is due to its stronger representation in the lexicon compared to the [r] variant by virtue of its greater overall frequency of occurrence.

This account has immediate predictions for a visual world paradigm study (VWP; e.g. Allopenna, Magnuson & Tanenhaus, 1998). Previous work using this methodology has highlighted the gradient nature of lexical access (e.g. McMurray, Tanenhaus & Aslin, 2002). That is, listeners notice subtle acoustic differences of produced words (e.g. VOT) and this sensitivity is reflected in the degree to which listeners fixate to a target item. Therefore, this is a useful paradigm in investigating listeners’ behavior in response to different allophonic variants. Following Ranbom et al.’s (2009) predictions, it is expected that by virtue of its greater overall frequency of occurrence, a canonical forms is expected to activate a target word to a greater extent than a non-canonical variant even when that non-canonical variant is presented in its licensing context. We test this prediction using the visual world paradigm.
2. **The study**

In this study, we tested adults’ ability to recognize familiar words when produced with canonical [t] and [d], non-canonical [ɾ], and one-feature mispronunciations in a word recognition task using a visual world paradigm. Specifically, we were interested in words with word-final [t] and [d] and their alternation with [ɾ]. Unlike in the word-medial context where underlying /t/ and /d/ almost invariably surface as [ɾ] in a given word of a particular phonological shape, word-final /t/s and /d/s surface variably in the same word as [t] and [d] or [ɾ] depending on conditioning context. Additionally, Herd et al. (2010) report that /d/ words are tapped more often than /t/ words. Thus, including [t] and [d] final words, allows us to investigate the impact of allophonic variation and its frequency on word recognition.

Adults were tested in a visual world paradigm. On each trial, subjects first saw two images on a display screen in a silent baseline phase. This allowed us to gauge any baseline preferences for the images shown. In the test phase, they then heard a sentence labeling one of the images in which the target word was produced in one of four conditions:

1. With a stop (e.g. *Look for the [bæt] Sam lost!*)

2. With a tap, in the appropriate licensing context, in this case before a vowel-initial word (e.g. *Look for the [bæɾ] again!*)

3. With a one-feature place mispronunciation (e.g. *Look for the [bæp] again!*).
4. With a phonological dissimilar novel label (e.g. *Look for the [lɪf] again!*)

The dissimilar labels on the novel trials were used to encourage subjects to look at the distractor item (e.g. [lɪf] for a *ball*). Thus, in novel trials, we expected subjects to look less at the target in the test phase compared to the baseline phase.

When /t/- and /d/-final target words are pronounced with canonical stops, adults were expected to look more to the target in the test phase compared to the silent baseline phase. Previous studies (e.g. McMurray *et al.*, 2002) have shown that lexical access in adults is gradient. Therefore we expected that one-feature mispronunciations should prompt fewer looks to the target object in comparison to the correct pronunciation condition (Swingley, 2009).

The crucial comparison here is between canonical forms and [ɾ] variants of a word. Here, there are two predictions. If there is an advantage for the canonical form, as Ranbom *et al.* (2009) suggest, we expect that the canonical forms of labels ([t]/[d]) should elicit a greater fixation proportion than the forms produced with a final [ɾ]. Alternatively, an inference account (Gaskell & Marslen-Wilson, 1996, 1998) predicts that because both canonical and [ɾ] variants are produced in the appropriate licensing context, they should be equally effective at activating the lexical representation of target words, since the [ɾ] production is most common when words have a singleton [t] coda and are followed by a vowel. That is, we expect to see comparable
target fixation proportions when a label is produced with either a canonical variant or a \([r]\) variant.

2.1 Subjects

38 participants were recruited via the University of California, Los Angeles Psychology Subject Pool and via advertisements. Subjects received course credit for participation. All identified themselves as native speakers of American English. Subjects were randomly assigned into one of 4 counterbalancing groups.

2.2 Target words

Since our final goal is to test children using the same paradigm, we chose target words that were known to 18-month-old infants. This allows us to easily compare the results of the subsequent infant study with the present adult one. The target words were a set of 6 [t]-final and 6 [d]-final nouns reported to be familiar to toddlers at 18-months of age based on the lexical norms of the MacArthur Bates Communicative Development Index (CDI; Dale & Fenson 1996). A list of target words is shown in Table 1 with their CDI comprehension at 16-months (the oldest age for which these norms are available) and CDI production frequency 18-months respectively.
Table 1. Comprehension frequency of target [t]- and [d]-words for 16-month-olds and production frequency for 18-month-olds.

<table>
<thead>
<tr>
<th>Target word</th>
<th>CDI comprehension freq. 16 months (%)</th>
<th>CDI production freq. 18 months (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[t]-words</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>hat</td>
<td>53.7</td>
</tr>
<tr>
<td>2</td>
<td>cat</td>
<td>51.2</td>
</tr>
<tr>
<td>3</td>
<td>boat</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>foot</td>
<td>41.2</td>
</tr>
<tr>
<td>5</td>
<td>bat</td>
<td>17.5</td>
</tr>
<tr>
<td>6</td>
<td>plate</td>
<td>13.8</td>
</tr>
<tr>
<td></td>
<td>[d]-words</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>bird</td>
<td>67.5</td>
</tr>
<tr>
<td>2</td>
<td>bed</td>
<td>28.8</td>
</tr>
<tr>
<td>3</td>
<td>bread</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>food</td>
<td>16.3</td>
</tr>
<tr>
<td>5</td>
<td>cloud</td>
<td>12.5</td>
</tr>
<tr>
<td>6</td>
<td>slide</td>
<td>12.5</td>
</tr>
</tbody>
</table>

Unfamiliar items selected for the study were real objects which were similar in category status (e.g. objects, animals) and visual complexity to the familiar items. The names of the unfamiliar items (see Appendix A) were not known by toddlers based on the MacArthur CDI (Dale & Fenson 1996). Additionally, we included 6 nonce words that were phonotactically legal in English but phonologically dissimilar to the names of the items they label: tesh [teʃ], wiss [wis], bize [baɪz], dape [deɪp] (previously used in Albright & Hayes, 2003), lif [lɪf] and neem [nim] (previously used in Stager & Werker, 1997).

2.3 Audio stimuli
The audio stimuli were digitally recorded by a phonetically-trained female native speaker of American English in an infant-directed register. Target words were recorded in two sentence frames: (1) Look for the [target] Sam lost! or (2) Look for the [target] 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 [d], and the tap formed a bimodal distribution, following the method of classification in Herd et al. (2010) where closure duration was the key measure. The average closure duration of [t] productions and [d] productions were 87.35 ms (s.d. = 11.82) and 71.98 ms (s.d. = 10.57) respectively and the average closure duration for their corresponding tapped productions were 35.46 ms (s.d. = 9.07) and 33.98 ms (s.d. = 9.75) respectively. The tap durations reported are in the range
reported by Zue & Laferriere (1979), Ranbom et al. (2009) and Herd et al. (2010). The closure duration of stops reported here are little longer than those reported by Byrd (1993) though this is expected given that the register used here was infant-directed. The overall distribution of taps vs. stops is illustrated in Fig. 1. The important point to note is that closure durations between taps and stops form a bimodal distribution with stops being longer than taps, and even more crucially, there was no overlap in their distributions.

![Closure duration](image)

**Fig. 1.** Distribution of stop/tap closure duration (Experiment 1)

The mean durations of target items in each condition were as follows: /t/-words: stop ($M = 429$ ms, $s.d. = 43$), tap ($M = 351$ ms, $s.d. = 59$), mispronunciation ($M = 385$ ms, $s.d. = 35$); /d/-words: stop ($M = 472$ ms, $s.d. = 80$), tap ($M = 375$ ms, $s.d. = 84$), mispronuncia-
tion ($M = 448$ ms, $s.d. = 90$); novel words ($M = 414$ ms, $s.d. = 62$). As expected, the duration of the target words in the tap condition were shorter than their counterparts in the canonical conditions, due to the shorter duration of [ɾ] compared to [t] or [d].

Mispronunciation of the target words always involved one-feature coda mispronunciations and always contained the same vowel as in the canonical [t]/[d] or tapped [t]/[d] conditions. Mispronunciations are presented in Table 2. Additionally, two more sentences were recorded, *Do you see it?* and *Can you find it?* These sentences played following the offset of first sentence to maintain participants’ interest in the procedure. To make the task more natural, we played a video of a puppet monkey, Sam, and recorded a short passage to introduce him. All sentences were recorded at 44,100 Hz were scaled to 65 dB in Praat (Boersma & Weenink, 2011) and played at a comfortable listening level of 65 dB.

<table>
<thead>
<tr>
<th></th>
<th>/t/-words</th>
<th>/d/-words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Pron.</td>
<td>Mispronunciation</td>
<td>Correct Pron.</td>
</tr>
</tbody>
</table>

Table 2. One-feature mispronunciation of test items
2.4 Visual stimuli

Visual stimuli were digitized color images, presented side by side on a 21.5 inch display monitor. Images were 350 by 350 pixels in size. Each familiar image was yoked with a unique image that was unfamiliar to infants in every trial. An example stimulus pair is depicted in Fig. 2. A list of stimulus items is given in Appendix A.

![Sample pair of display images (familiar target = cat, distractor = stingray). Yellow boxes represent interest areas.](image)

2.5 Apparatus and Procedure

We follow the experimental procedure set out in White & Morgan (2008). The ability of this paradigm to show graded sensitivity effects to mispronunciations should allow us to see if there is an effect of regularly alternating segments on word recognition. 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). They were seated between 500 and 600 mm away from the eyetracking camera ($M = 553$, $s.d. = 16.81$). Speech stimuli were presented through concealed a set of Cyber Acoustics CA-3602 loudspeakers hidden behind curtains under the display. Participants’ fixations were recorded by the Eyelink system, using the arm-mount remote configuration. Participants’ gaze was calibrated and validated using a five-point calibration. Validations of Fair and Good as determined by the eyetracking software were accepted. The average degree of error was 0.48°.

Before the trials started, subjects were shown a short video with a puppet monkey that was introduced as Sam. This made the semantic context of the stimuli sentences felicitous. Each trial consisted of two phases: a baseline phase and a test phase. Each trial began with a looming object displayed at the center of the display with an accompanying baby giggle. When the subject fixated at the center of the screen, the experimenter initiated the baseline phase. In the silent baseline phase, subjects were shown two images side-by-side 960 pixels apart, a familiar target image and an unfamiliar distractor image, simultaneously. Images were left on screen for 4 seconds. The baseline phase was included to allow subjects time to look at both objects and to allow us to gauge any baseline preferences for a particular image. Following this, the subject’s gaze was re-centered and the experimenter initiated the test phase. In the test
phase, subjects were shown the same two images again. At the onset of images, subjects heard the audio stimulus *Look for the X Sam lost! Or Look for the X again!* synchronized with the onset of image presentation. Subjects were presented with a follow-up sentence *Do you see it/Can you find it?,* 750 ms after the offset of the first sentence. Images were left on screen for approximately another 2 seconds. Thus, each test phase lasted for approximately 6 seconds.

The experiment consisted of 24 trials. There were 6 trials in the novel (nonce word) condition. The remaining 18 trials were divided evenly between the 3 remaining test conditions, such that there were 6 trials per condition. 4 counterbalancing groups were created. No participant heard the same target word in both canonical and tap conditions. Novel trials were used to encourage participants to look at the distractor item. All subjects were presented the same 6 novel trials; the assignment of stimulus pairs to canonical, tap or mispronunciation condition was counterbalanced across subjects. The 6 [t] target words and 6 [d] target words were divided into two groups of 3 words each. MP trials were derived from a subset of the canonical and tap trials. 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 target images appeared equally on the left and on the right. Target image location was consistent across baseline and test phases. The dependent measure was the change in participants’ proportion looking to
the target object between the baseline phase and the test phase. Each experimental session lasted about 3 minutes.

2.6 Results

Two interest areas (Target and Distractor) were set at 500 by 500 pixels around the target images (see Fig. 2). The interest areas were larger than the images themselves so as not to penalize fixations which, while not on the images themselves, were nonetheless in the right area of the screen. Looking behavior was sampled at 2-ms intervals in two phases. For the baseline phase, looking behavior was extracted for the entire 4-second duration of the phase. We eliminated trials in which subjects did not look at both images during the baseline phase (14 trials, 1.9%). Subjects were excluded entirely if more than 50% of their trials were not usable due to the above criteria (3 of 38). Another three subjects were excluded due to experimenter error. This yielded a final pool of 32 subjects divided evenly across the 4 counterbalancing groups, with 8 subjects in each group.

Because the canonical forms and nonce words differed in their onset, for a comparison between canonical forms and nonce words, the test phase consisted of a 2-second window from the onset of the target words. For a comparison between target words produced with stops, taps, and one-feature mispronunciations the test phase was defined as a 2-second window from the
onset of the coda of the target word where the words become disambiguated. The first 367ms of test phase was excluded from analysis as previous research shows that eye movements that occur before this point are unlikely to be responses to the target word, given the time necessary to initiate eye movement (Swingley & Aslin 2000; 2002).\(^1\) The choice of a 2000 ms window follows from previous studies using this paradigm (e.g. Swingley 2009).

The dependent measure was the difference in participants’ looking to the familiar object between the silent baseline phase and the test phase (%Test - %Baseline). This measure indicates therefore how much subjects looked to the target object after it was named, taking into account any baseline preferences for that object. A difference score was calculated for each trial. The difference scores were then subject to analyses using linear mixed-effect models with the \textit{lme4} package (Bates, Maechler & Dai 2008) in R (R Development Core Team, 2008). The \texttt{anova()} function was used to compare models which were in a subset relationship. The results of this likelihood ratio are reported as chi-squares.

### 2.6.1 \textit{Stops vs. Novel}

In the first analysis (Fig. 3), we were interested in comparing adult listeners’ behavior in the stop and novel conditions as a means of validating the task. Here we compared looking

\(^1\) This is the time-window usually used with infants (Swingley & Aslin, 2000, 2002). Results do not differ significantly using this time-window as compared to using a later time-window used with adults. Therefore, the 367-ms start point of the analysis window was maintained for uniformity with a planned infant experiment.
behavior starting from the start of the first consonant of the target word as the target and novel word differed in their initial segment. Test-baseline difference scores are shown in Fig. 3. As expected, adult listeners looked significantly more to the target object in the test phase compared to baseline in the canonical stop condition than in the novel condition ($\chi^2(1) = 51.47, p < 0.001$). In fact, when adult listeners heard a novel label for the familiar target on the screen, they looked more at the distractor object. This result validates the basic paradigm.

Fig. 3. Difference score between stops vs. novel, Experiment 1. Error bars represent $\pm 1\ SE$. 
2.6.2 StOPS VS. MPs VS. TAPS

Next, we compared looks to the target image when the word was produced in three conditions: canonical stop, tap and one-feature mispronunciation (MP). In the second analysis, we compared participants’ behavior when they heard target words produced with a stop, a tap and a one-feature mispronunciation (Figure 4). Looking behavior was analyzed from a 2-s window starting from the start of the coda consonant (following the offset of the vowel). Recall that the stop, tap and mispronunciation trials differ in their coda consonants. From Figure 3 we can see that listeners looked more to the target image in the test compared to baseline in all three conditions, suggesting that they were all treated as potential labels for the target objects.

**Fig. 4.** Difference score between stop, tap and mispronunciation condition. Error bars represent $\pm 1 \ SE$. 
We included two factors in our model, Word type (/t/ vs. /d/) and Condition (Stop vs. Tap vs. Mispronunciation), as well as, their interaction. Condition was sum-coded with the Tap condition as reference. The interaction of Word Type by Condition was not significant ($\chi^2(2) = 3.35, p = 0.19$). This interaction term was excluded in subsequent model comparisons. Test of the main effects of Word Type and Condition were conducted against a superset model with both variables but without the interaction term. Participants’ looking behavior did not differ by Word Type ($\chi^2(1) = 1.18, p = 0.28$), indicating that they treated /t/ and /d/ words similarly across conditions. Importantly, there was a significant effect of Condition ($\chi^2(2) = 14.54, p < 0.001$).

Post-hoc pairwise comparisons were conducted using the \textit{glht()} function of the \textit{multcomp} package (Hothorn, Bretz & Westfall 2008). All comparisons were adjusted using Tukey’s method. The analysis revealed that listeners looked significantly more at the target object in the stop condition and tap condition than in the mispronunciation condition (Stop vs. Mispronunciation: $p < 0.001$; Tap vs. Mispronunciation: $p < 0.001$). Importantly, listeners did not differ in behavior in the stop and tap conditions ($p = 0.99$), looking equally at the target image at test.

2.6.3. Timecourse of fixations

We were also interested in comparing how quickly listeners oriented their gaze to the target object in the stop and tap conditions (Figures 5 and 6). To do so, target fixations were
analyzed using Growth Curve Analysis (GCA) with orthogonal power polynomials to (a) take into account the lack of independence between successive time bins and (b) to capture the shape of the relation between fixation proportions and time (Mirman, Dixon & Magnuson, 2008). In GCA, visual world eyetracking data are analyzed through two hierarchical submodels. The first sub-model, the base model, captures the effect of time through the use of third-order orthogonal polynomials \((t, t^2, t^3)\). These three parameters are considered necessary to fit the typical sigmoidal form of the curve associated with target fixation proportions over time. The use of orthogonal polynomials allows us to transform the time vectors to make them independent since natural polynomials are highly correlated (see Mirman et al., 2008). What this means is that \(t, t^2,\) and \(t^3\) are not actually interdependent as natural polynomials are, but are rather independent from each other.

The dependent variable was mean target fixation proportion. Fixation data was binned into 50 ms intervals, resulting in 80 bins for the baseline phase and 33 bins for the test phase. The base model simply included all three orthogonal polynomials in each phase as fixed predictors. The three orthogonal polynomials were required to capture the shape of the curve. The independent variables of Condition (Stop vs. Tap) and Word Type (/t/ vs. /d/). In this paper we were primarily interested in whether the fixed effects predicted the average overall target fixation proportion. This can be seen in the fixed effects of experimental variables on the intercept.
Interactions of fixed effects of experimental variables on the linear term are related to differences in the overall angle of the curve as a function of condition. The initial full model consisted of all three polynomial terms, the two independent variables and their interactions and the interaction of the independent variables and the linear term. All models also included by-subject random effects with random slopes for all time variable terms and subject-by-condition random effects with random effects for all time variable terms.

All analyses were carried out in R (R Core Development Team, 2008), using the lmer() function in the lme4 package (Bates, Maechler, & Dai, 2008). Models were compared using the anova() function, with comparisons being made between models in a subset relationship. There was neither a significant main effect of Condition ($\chi^2(2) = 4.82, p = 0.09$) nor a significant effect of Word Type ($\chi^2(1) < 0.001, p = 0.99$). There was also no significant interaction ($\chi^2(1) = 2.17, p = 0.14$). This confirms the result from the analysis above. In terms of the effect of the independent variables on the linear term, there was a significant effect of Condition on linear term ($\chi^2(1) = 4.73, p = 0.03$). Neither the effect of Word Type on the linear term ($\chi^2(1) = 0.07, p = 0.79$) nor the effect of the interaction of Condition and Word Type was significant ($\chi^2(1) = 3.43, p = 0.06$). This suggests that while the overall fixations were similar across both conditions, listeners were quicker at orienting towards the target image when they heard a stop label than when they heard a tap label.
Adult American English listeners therefore recognize words equally well when these words are produced with either a canonical stop variant or tap variant, although they were quicker at orienting their gaze to the target when they heard a stop variant. Furthermore, mispronounced words incur a cost to word recognition when compared to both the stop and tap conditions although they are nonetheless accepted as labels for the target image.

![Time-course of adults’ fixation to the target in the test phase for /t/-words. The dashed red line represents the start of analysis window, 367 ms after the coda of the target word. Error bars represent ±1 SE.](image)

**Fig. 5.** Time-course of adults’ fixation to the target in the test phase for /t/-words. The dashed red line represents the start of analysis window, 367 ms after the coda of the target word. Error bars represent ±1 SE.
Fig. 6. Time-course of adults’ fixation to the target in the test phase for /d/-words. The dashed red line represents the start of analysis window, 367 ms after the coda of the target word. Error bars represent ±1 SE.

3. Discussion and Conclusion

In this study, we investigated the ability of listeners to recognize target words when these words are produced with both a canonical form and an allophonic variant. Listeners were presented with labels for target items with word-final /t/ and /d/ that were produced with a canonical stop, a tap or a mispronunciation. Additionally, listeners also heard phonologically dissimilar labels for target items. We had the following 4 predictions. (1) When presented with a novel label, it was predicted that adults would look more at the distractor item in the test phase than in the baseline phase. (2) Subjects were expected to look more at the target object in the test phase when presented with canonical stop variants. (3) In the mispronunciation condition,
subjects were predicted to look more at the target in the test phase, but not as much as when
the label is presented with a canonical stop. (4) We were crucially interested in whether the [r]
variant patterned with the canonical stops. (4a) If listeners truly ignore phonological context
during processing of this particular variant and have the canonical forms more strongly repre-
sented due to their higher overall production frequency as Ranbom et al. (2009) assert, the [r]
variant should not be as good as the canonical forms in activating the lexical representation of
the target word. (4b) If, on the other hand, the efficacy with which a variant activates a lexical
representation is tied to its frequency in its given context, then we expected both the canonical
form and the [r] variant to be equally effective at activating the lexical representation of the
target word. That is, we expected equal proportion looks to the target in both conditions.

In our experiment, when presented with a novel label for a known target image, subjects
looked less to the target image and more at the distractor image in the test phase, replicating
the result of White & Morgan (2008). This tells us that adults succeeded on the task and that
they interpreted the novel label as a potential label for the less familiar distractor item. In the
canonical stop condition, adults looked more at the target object in the test phase than in the
baseline phase. Additionally, consistent with previous mispronunciation studies (e.g. Swingley
2009; also see McMurray et al., 2002; see also Connine et al., 1993), adults accepted mispro-
nounced words as labels for target objects, as exhibited by an increase in looks to the target
image in the test phase compared to the baseline. This increase, however, was less than the increase seen when subjects heard correct labels for target images produced with either a stop or a tap, indicating more uncertainty in the mispronunciation cases than in the stop or tap cases. This replicates findings of previous word recognition studies (McMurray et al., 2002; Swingley, 2009), confirming that word recognition proceeds in a gradient manner.

Crucially, participants treated labels produced with both stops and taps as equivalently in terms of overall target fixation proportions when presented in their respective licensing contexts. This was true of both /t/- and /d/-words. The average target fixation proportion data support the inference hypothesis (Gaskell & Marslen-Wilson, 1996; 1998), which suggests that both variants presented in the appropriate context are equally effective at activating a target lexical representation.

However, when we examined the time-course of fixation proportions, the canonical condition in both /t/- and /d/-words were characterized by a faster increase in fixation proportions, suggesting that participants were quicker at shifting their gaze towards the target object when the label was produced with a canonical stop. The difference in slope of increase in fixation proportion for both /t/ and /d/ words support Ranbom et al.’s (2009) account, which predicts an advantage for the canonical form over the non-canonical variant. In order for us to re-
ally disentangle these two accounts, we would need to examine whether or not word recognition is inhibited when the tap variant is presented in an inappropriate context.

Taken together, the results of the present study suggest a number of implications for models of word recognition. Spoken word recognition models largely fall into two distinct types: inference-based accounts and representational accounts. An inference type account argues for some inferential process that occurs to “undo” a phonological process to recover the underlying form (Lahiri & Marslen-Wilson, 1991; Gaskell & Marslen-Wilson, 1996; 1998). For such an account, the phonological context in which these processes occur is crucial. In the appropriate context, canonical and non-canonical variants are predicted to activate a target lexical representation to the same extent. In the current study, the fact that there was no difference in average target fixation proportions between both the canonical and tap variants supports an inference account. On the other hand, an inference account is unable explain the fact that listeners were faster at shifting their gaze to the target object in the canonical condition than in the tap condition across both /t/- and /d/-words. This suggests that canonical forms were faster at activating the lexical representation of a target word than tap variants in word-final position, replicating the findings in Ranbom et al. (2009).

We will now discuss how representational accounts might account for us results. Exemplar accounts are one variety of representational accounts in which listeners have highly de-
tailed lexical representations of variant forms that are entirely dependent on a listener’s previous experience (e.g. Pierrehumbert, 2002). In this exemplar-based approach, details such as speaker indexical information and other surface phonetic details are maintained in the lexical representation. Importantly, the context of production is also encoded. The strength of the representation is determined by the frequency with which a form is heard in a given context. This model predicts then that the variant forms associated with a particular context should be activated most strongly, which would suggest in our case that we should not see any difference between the tap and stop conditions since they are both in licensing contexts.

Another variant of a representational account as put forth by Ranbom et al. (2009) asserts that multiple abstract variants are directly encoded into the lexicon (see also Connine, 2004; Ranbom & Connine, 2007). Specifically, they argue for frequency-based representations in which the strength with which a variant is encoded in the lexicon is determined by the overall frequency of occurrence of a particular variant in conversational speech. In such an account, the higher frequency of [t] forms across all contexts compared to [ɾ] forms results in a stronger encoding of [t] forms than [ɾ] forms in a lexical representation. This predicts that the more frequent canonical form would be better at activating a lexical representation than the less frequent tap variant. In our current study, the stronger encoding of the canonical form would ex-
plain the faster shift in fixations to the target in the canonical condition compared with the tap, although this would not explain the absence of differences in proportion fixation data.

A further claim of Ranbom et al.’s (2009) multiple representation account is that phonological context is ignored for coda [ɾ] due the poor role context plays in predicting this variant’s occurrence since they found priming effects for [ɾ] variants even when the following context was inappropriate (e.g. preceding a consonant). This would predict that in an inappropriate context, [ɾ] variants should still be able to activate a target lexical target representation. An inference-based model predicts that this should not be the case. Our current experiment does not speak to this prediction since listeners were only presented with variants in licensing contexts. For a more stringent test of Ranbom et al.’s (2009) hypothesis, we would need to examine whether or not the [ɾ] variants are still able to activate a lexical representation when spliced into an inappropriate context. In fact, in a visual world paradigm study, they predict that listeners will accept [ɾ] variants as labels for a target object and not as a mispronunciation when the [ɾ] variant is produced in an inappropriate context (e.g. before a consonant).

More broadly, an account of word recognition that has multiple variants of a lexical item stored in the lexicon raises the question of how these get encoded in the first place and how one decides which form is canonical if it is not just the most frequent form. More simply put, do children directly represent multiple surface forms tied to specific phonological contexts
either in an exemplar-like fashion (Pierrehumbert, 2002) or as abstract representations (Connine, 2004; Ranbom & Connine, 2007; Ranbom et al., 2009)? If Ranbom et al. (2009) are correct in asserting that multiple variants are stored, then one would need an account of how this would occur during development. Such an account would also have to explain how learners generalize the alternation to newly learnt words from a learnt canonical form. This is a question that requires further investigation.

Through a word recognition experiment in a visual world paradigm, we examined the interaction between allophonic variation and spoken word recognition. Both canonical and tap variants were equally good labels for target items. Both of these variants produced greater proportion fixations to a target image compared to a silent baseline than mispronounced labels, suggesting that they were both viable labels for a target item. However, we saw differences in the rate at which subjects shifted their gaze to the target object with canonical variants eliciting a steeper slope in fixation proportions than tap variants, suggesting that the canonical variant was better at activating a target lexical representation than a non-canonical variant. Future work will seek to test the ability of [ɾ] variants to activate a lexical representation when presented in a non-licensing context. This will allow us to investigate the role of context plays in the processing of non-canonical variants.
## Appendix A: List of target and distractor items

<table>
<thead>
<tr>
<th>Display</th>
<th>Label</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>Distractor</td>
<td>/t/-words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hat</td>
<td>piston</td>
<td>hat(^S)</td>
<td>hat(^T)</td>
<td>hat(^S)</td>
<td>hat(^T)</td>
</tr>
<tr>
<td>boat</td>
<td>accordion</td>
<td>boat(^S)</td>
<td>boat(^T)</td>
<td>boat(^S)</td>
<td>boat(^T)</td>
</tr>
<tr>
<td>bat</td>
<td>floppy disk</td>
<td>bat(^S)</td>
<td>bat(^T)</td>
<td>bat(^S)</td>
<td>bat(^T)</td>
</tr>
<tr>
<td>cat</td>
<td>seahorse</td>
<td>cat(^T)</td>
<td>cat(^S)</td>
<td>cat(^T)</td>
<td>cat(^S)</td>
</tr>
<tr>
<td>foot</td>
<td>barrel</td>
<td>foot(^T)</td>
<td>foot(^S)</td>
<td>foot(^T)</td>
<td>foot(^S)</td>
</tr>
<tr>
<td>plate</td>
<td>scales</td>
<td>plate(^T)</td>
<td>plate(^S)</td>
<td>plate(^T)</td>
<td>plate(^S)</td>
</tr>
<tr>
<td>cat</td>
<td>stingray</td>
<td>cak(^{MP})</td>
<td>-</td>
<td>cak(^{MP})</td>
<td>-</td>
</tr>
<tr>
<td>foot</td>
<td>padlock</td>
<td>foop(^{MP})</td>
<td>-</td>
<td>foop(^{MP})</td>
<td>-</td>
</tr>
<tr>
<td>plate</td>
<td>cone</td>
<td>plake(^{MP})</td>
<td>-</td>
<td>plake(^{MP})</td>
<td>-</td>
</tr>
<tr>
<td>hat</td>
<td>bullhorn</td>
<td>-</td>
<td>hap(^{MP})</td>
<td>-</td>
<td>hap(^{MP})</td>
</tr>
<tr>
<td>boat</td>
<td>paint roller</td>
<td>-</td>
<td>boak(^{MP})</td>
<td>-</td>
<td>boak(^{MP})</td>
</tr>
<tr>
<td>bat</td>
<td>hourglass</td>
<td>-</td>
<td>bap(^{MP})</td>
<td>-</td>
<td>bap(^{MP})</td>
</tr>
<tr>
<td>/d/-words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slide</td>
<td>iron</td>
<td>slide(^S)</td>
<td>slide(^T)</td>
<td>slide(^T)</td>
<td>slide(^S)</td>
</tr>
<tr>
<td>cloud</td>
<td>magnifying glass</td>
<td>cloud(^S)</td>
<td>cloud(^T)</td>
<td>cloud(^T)</td>
<td>cloud(^S)</td>
</tr>
<tr>
<td>bread</td>
<td>wench</td>
<td>bread(^S)</td>
<td>bread(^T)</td>
<td>bread(^T)</td>
<td>bread(^S)</td>
</tr>
<tr>
<td>bed</td>
<td>garlic</td>
<td>bed(^T)</td>
<td>bed(^S)</td>
<td>bed(^S)</td>
<td>bed(^T)</td>
</tr>
<tr>
<td>bird</td>
<td>camel</td>
<td>bird(^T)</td>
<td>bird(^S)</td>
<td>bird(^S)</td>
<td>bird(^T)</td>
</tr>
<tr>
<td>food</td>
<td>joystick</td>
<td>food(^T)</td>
<td>food(^S)</td>
<td>food(^S)</td>
<td>food(^T)</td>
</tr>
<tr>
<td>bed</td>
<td>beehive</td>
<td>beb(^{MP})</td>
<td>-</td>
<td>-</td>
<td>beb(^{MP})</td>
</tr>
<tr>
<td>bird</td>
<td>french horn</td>
<td>birb(^{MP})</td>
<td>-</td>
<td>-</td>
<td>birb(^{MP})</td>
</tr>
<tr>
<td>food</td>
<td>cassette</td>
<td>foog(^{MP})</td>
<td>-</td>
<td>-</td>
<td>foog(^{MP})</td>
</tr>
<tr>
<td>slide</td>
<td>anchor</td>
<td>-</td>
<td>slige(^{MP})</td>
<td>slige(^{MP})</td>
<td>-</td>
</tr>
<tr>
<td>cloud</td>
<td>microscope</td>
<td>-</td>
<td>cloub(^{MP})</td>
<td>cloub(^{MP})</td>
<td>-</td>
</tr>
<tr>
<td>bread</td>
<td>pipe</td>
<td>-</td>
<td>breag(^{MP})</td>
<td>breag(^{MP})</td>
<td>-</td>
</tr>
<tr>
<td>Novel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ball</td>
<td>pogo-stick</td>
<td>lif(^N)</td>
<td>lif(^N)</td>
<td>lif(^N)</td>
<td>lif(^N)</td>
</tr>
<tr>
<td>shoe</td>
<td>rake</td>
<td>neem(^N)</td>
<td>neem(^N)</td>
<td>neem(^N)</td>
<td>neem(^N)</td>
</tr>
<tr>
<td>car</td>
<td>reel</td>
<td>tesh(^N)</td>
<td>tesh(^N)</td>
<td>tesh(^N)</td>
<td>tesh(^N)</td>
</tr>
<tr>
<td>balloon</td>
<td>tambourine</td>
<td>wiss(^N)</td>
<td>wiss(^N)</td>
<td>wiss(^N)</td>
<td>wiss(^N)</td>
</tr>
<tr>
<td>bear</td>
<td>fishingrod</td>
<td>bize(^N)</td>
<td>bize(^N)</td>
<td>bize(^N)</td>
<td>bize(^N)</td>
</tr>
<tr>
<td>cookie</td>
<td>chainsaw</td>
<td>dape(^N)</td>
<td>dape(^N)</td>
<td>dape(^N)</td>
<td>dape(^N)</td>
</tr>
</tbody>
</table>

Superscripts indicate trial condition: \(S\) = stop, \(T\) = tap, \(MP\) = mispronunciation and \(N\) = novel.
References


