# TRACKING PERCEPTION OF PRONUNCIATION VARIATION BY TRACKING LOOKS TO PRINTED WORDS: THE CASE OF WORD-FINAL /t/

*Holger Mitterer & James M. McQueen*

Max Planck Institute for Psycholinguistics
holger.mitterer@mpi.nl james.mcQueen@mpi.nl

## ABSTRACT

We investigated perception of words with reduced word-final /t/ using an adapted eye-tracking paradigm. Dutch listeners followed spoken instructions to click on printed words which were accompanied on a computer screen by simple shapes (e.g., a circle). Targets were either above or next to their shapes, and the shapes uniquely identified the targets when the spoken forms were ambiguous between words with or without final /t/ (e.g., *bult*, bump, vs. *bul*, diploma). Analysis of listeners' eye-movements revealed, in contrast to earlier results, that listeners use the following segmental context when compensating for /t/-reduction. Reflecting that /t/-reduction is more likely to occur before bilabials, listeners were more likely to look at the /t/-final words if the next word's first segment was bilabial. This result supports models of speech perception in which prelexical phonological processes use segmental context to modulate word recognition.

**Keywords:** perception, variation, /t/ reduction, phonological processing, visual-world paradigm.

## 1. INTRODUCTION

A continuous-speech phenomenon such as reduction of word-final /t/ makes word recognition harder, especially in cases where it would lead to a different word, as when the English word *mist* sounds like *miss*. We investigate here how listeners recover the intended meaning of reduced forms. Specifically, we ask whether listeners make use of following context when compensating for /t/ reduction. If so, this would suggest that word recognition cannot be based simply on matching the current input to stored lexical knowledge, but that it must also involve phonological knowledge, so that words can be recognized as a function of the contexts in which they appear.

Previous investigations of /t/ deletion in Dutch speech production have shown that word-final /t/ is most likely to be reduced after /s/ and before bilabial consonants [10]. In perception, however, an effect only of previous context was found [10]. Listeners were more inclined to restore a reduced word-final /t/ if the consonant preceding the to-be-restored /t/ was an /s/ (mirroring the production results). But there was no parallel between production and perception for following context effects. Listeners were just as likely to restore a reduced word-final /t/ when the following consonant was bilabial as when it was alveolar.

If this data pattern were to hold up, then one could explain recognition of words with reduced final /t/ in terms of lexical storage alone. Since the preceding context is part of the word, knowledge about reduction could be stored in the lexicon. For example, /t/-reduced variants of words with final /st/ in their citation form could have their own lexical entries. But if an effect of following context could be found, then this would suggest that phonological knowledge other than that about the word itself is brought to bear in recognising /t/-reduced forms.

We therefore tested for effects of following context using a task different from that used previously. Mitterer and Ernestus [10] used a simple two-alternative forced-choice (2AFC) task, which may have unduly focussed listeners' attention on the phonetic properties of the stimuli, thereby masking effects of phonological processing. We used a paradigm which instead focussed listeners' attention away from the phonetic detail. A new eye-tracking paradigm was developed in which listeners followed spoken instructions to click on printed words on a computer screen. Instructions were disambiguated semantically (see Method), so that listeners did not need to attend to stimulus details to perform the task – unlike in the 2AFC case. Participants' eye-movements were recorded as they followed these instructions. Eye-movements may be better able to reflect on-line phonological processing than 2AFC categorisation judgements, and thus could show an influence of following context on the perception of reduced word-final /t/'s.

## 2. METHOD

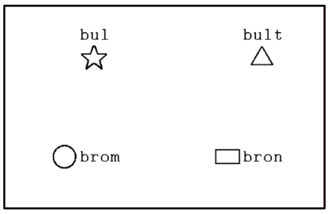
In the standard version of the visual-world paradigm, participants view a display of several objects and hear a sentence – often containing an instruction to click on one of the depicted objects [12]. Eye-movements reflect on-line processing because participants fixate on objects before the objects' names have been fully disambiguated in the instructions. For example, participants hearing the syllable *ham* will immediately fixate pictures of a *ham* and of a *hamster,* instead of waiting until the instruction has specified the target unambiguously [1, 3, 11].

McQueen and Viebahn [8] recently showed that if printed words are presented on the display, similar effects are obtained to those found with pictures [1]. This broadens the applicability of the visual-world paradigm because one is no longer constrained to use only objects or pictures of objects as stimuli. We used the printed-word version of the visual-world paradigm here, but adapted it further so that the target was designated by a combination of its printed form and an accompanying geometrical shape.

Consider the example display in Figure 1, coupled with the instruction "Click on the word [byl] above the star". This instruction is ambiguous up to the final word, because [byl] may refer either to *bul* (diploma) or to *bult* (bump) – if the final /t/ has been reduced. The instruction is disambiguated on the final word, however, so listeners do not have to make an explicit decision, based on the phonetic detail in the critical syllable, about whether they have heard *bul* or *bult*. In Dutch, the instructions are well-suited to our needs, because the preposition "above", *boven,* starts with a bilabial consonant, while the preposition "next to", *naast*, starts with an alveolar consonant. If listeners are sensitive to the production fact that reduction of word-final /t/ is more likely if the following consonant is bilabial, we should find more looks to the word with final /t/ (*bult* in this case) when the following preposition is *boven* than when it is *naast*. Critically, this effect should occur before the ambiguity is resolved by the shape name.

We thus examined recognition of minimal pairs such as *bult/bul* where the targets were followed by *boven* or *naast*. We also manipulated the phonetic detail in the critical ambiguous syllables, using forms biased towards either a /t/ interpretation or a no-/t/ interpretation.

**Figure 1:** An experimental display.



### 2.1. Participants

40 native speakers of Dutch were paid to take part. None reported hearing problems.

### 2.2. Visual Stimuli

32 minimal /Ct#/-/C#/ pairs, such as *bult-bul*, were used. There were a further 16 filler minimal pairs, ending with nasal consonants (e.g., *brom-bron* in Figure 1). These stimuli were combined in 96 trials, such that all pairs appeared four times. Every /Ct#/-/C#/ pair thus appeared on the screen twice as distractors (i.e., when each member of the filler pair was the target), once with the /Ct#/ word (e.g., *bult*) as target, and once with the /C#/ word (e.g., *bul*) as target. In all of these trials both members of each pair appeared in the same spatial relation to their associated shape (see Figure 1). This meant that the critical instructions could be fully disambiguated only by the shape names.

On an additional 48 trials, two pairs of *identical* words, drawn from the set of /Ct#/-/C#/ and filler minimal pairs, appeared on the screen. Targets in these trials were designated either by their position in relation to the same shape (e.g., one was above a star, and the other was next to a star), or by the shape itself (e.g., one was above a star, and the other was above a circle). These *semantic* trials allow us to test whether participants immediately use the available information to guide their eye-movements, or just wait for the sentence to finish. Given that the positional cue occurs earlier in the sentence than the shape cue, we expect eye-movements to the target to be faster in the position condition. The semantic trials thus allow us to assess the validity of this new version of the visual-world paradigm.

### 2.3. Auditory Stimuli

A female native speaker of Dutch was recorded producing the instruction sentences for all experimental target pairs derived from the combinations in Table 1. Similar frames were used for the instructions for the filler trials.

**Table 1:** The sentence frames for the instructions.

| | Preceding Context | Target | Following Context | Shape |
|---|---|---|---|---|
| Dutch | klik op het woordje | *TARGET* | boven de<br>naast de | ster<br>rechthoek<br>cirkel<br>driehoek |
| IPA | klɪk<br>ɔp ət<br>wᴏʀdjə | C(C)VC(t) | bovə də<br>nast də | stɛʀ<br>ʀɛxthuk<br>sɪʀkɛl<br>dʀihuk |
| Gloss | click on the word | *TARGET* | above the<br>next to the | star<br>rectangle<br>circle<br>triangle |

All materials were cross-spliced, so that, critically, the following contexts *boven de* and *naast de* were identical in all instructions. Furthermore, for each experimental pair two ambiguous versions were created, both without a /t/-burst: one with a short (pen-)ultimate consonant and a 50 ms closure, the other with a (pen-)ultimate consonant that was 25 ms longer and a 25 ms closure. The first ("+/t/-bias") is more likely to occur if the speaker intends a word-final /t/; the second ("-/t/-bias") if the speaker does not [10].

### 2.4. Procedure

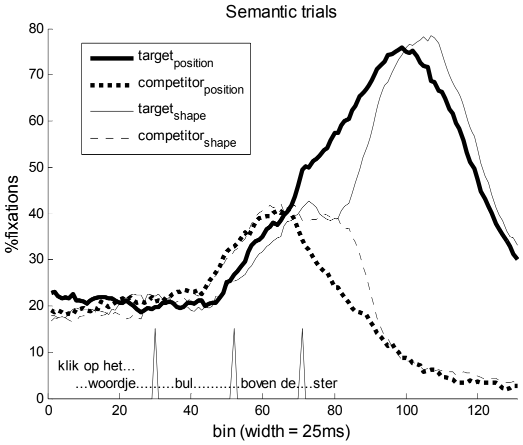
Each participant completed six practice trials. The eye-tracking equipment was then mounted and the main experiment began. The 144 trials were randomized for each participant individually, with semantic and /t/-word trials mixed.

## 3. RESULTS

### 3.1. Semantic trials

Figure 2 shows the fixation proportions to targets in bins, after normalization for sentence-part duration, of approximately 25 ms. Fixations to the targets rise at the expense of the competitor and reach their maximum earlier in the position condition than in the shape condition. This difference between conditions is significant in bins 72 to 91. This reflects the temporal unfolding of the instructions, in which the position information precedes the shape information (see Table 1).

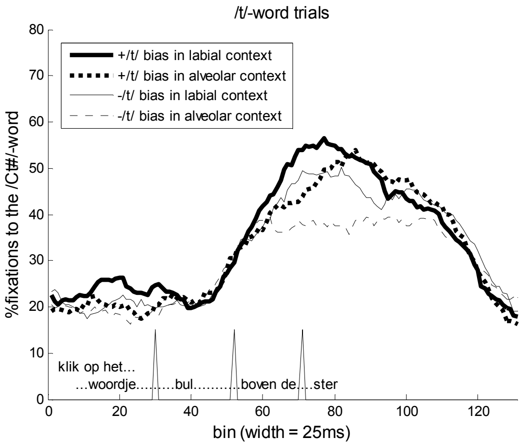
**Figure 2:** Fixations to targets and competitors in the semantic trials. Thick lines represent the data from the position condition, thin lines the data for the shape condition. The continuous lines represent the fixation proportions to the target and the dashed lines the fixation proportions to the competitor. The vertical lines indicate the splicing points in the instructions.



### 3.2. /t/-word trials

Figure 3 shows the proportion of fixations to the words with final /t/, averaged over cases where the /t/-final word is the target and cases where it is the competitor of the target without a /t/. There were significantly more fixations of the /t/-words if the acoustic form of the coda was more in line with the presence of an underlying /t/ and if the following context was bilabial. There was no interaction between these two factors.

**Figure 3:** Fixations to words with final /t/. Thick lines indicate fixation proportions for stimuli in which the acoustic forms had a +/t/ bias; thin lines for stimuli in which the forms had a -/t/ bias. Continuous lines show fixation proportions in the bilabial-context condition (*"boven"*) and dotted lines show fixation proportions in the alveolar-context condition (*"naast"*). The vertical lines indicate the splicing points in the instructions.



## 4. DISCUSSION

The results from the semantic trials indicate that the extended visual-world paradigm is well-suited to investigate the on-line processing of speech perception, as it reveals immediate effects of incoming information on the interpretation of the sentence. Moreover, the use of printed words as targets in conjunction with associated geometrical shapes allows one to control well the phonological context in which critical words appear, while using sentences that are not artificial in the task setting.

This paradigm was used to investigate the perception of words with reduced final /t/. We found that listeners use the context following a reduced /t/ to compensate for that reduction. Participants looked more at the /t/-final words, as they were hearing a following word, when that following word began with a bilabial consonant than when it began with an alveolar consonant. These perceptual results thus mirror the context-sensitivity previously found in production, where /t/ is most likely to be reduced if a bilabial consonant follows [10]. We suggest that the failure to observe this context effect using a 2AFC task [10] may reflect a task difference. Whereas the 2AFC task may encourage listeners to focus on phonetic detail and hence to disregard phonological context, the present task focussed listeners' attention on meaning, allowing us to see the phonological processes that we would argue operate in normal speech recognition.

These results suggest that reduced forms cannot be recognised simply by mapping the current input onto stored lexical knowledge. Since the following consonant is not itself part of a reduced word, knowledge involving that consonant (i.e., about the relative likelihood of that reduced word in that context) cannot readily be stored in that word's lexical entry. These data thus suggest that there is some kind of contextual inference process in speech perception, whereby recognition of reduced words is modulated by their phonological context.

One possible account is Gaskell's model of phonological inference [4, 5]. This model has been used to explain contextual effects in compensation for assimilation, but could also be applied to other continuous-speech processes such as /t/ reduction. In this model, an early phonetic stage proceeds independently of context, and then a phonological processor takes segmental context into account. The lack of a contextual effect in the 2AFC task [10] could be explained as the result of attention being focussed on the output of the first stage.

Phonological inferences in this model, though occurring after early phonetic analysis, still occur prelexically. If a reduced /t/ can be restored by a prelexical contextual inference process, then only citation forms would need to be stored in the lexicon. It is thus unnecessary to store lexically both reduced and unreduced variants of words, as assumed in models in which the lexicon consists of word-sized exemplars [2, 6]. Furthermore, such models have limited scope for phonological generalization [7, 9]. If pronunciation variation were stored lexically, the effect of following context would thus have to be learned word for word. Such models therefore offer a less plausible account of the perception of reduced words than models with abstract lexical representations and prelexical phonological inferences.

## 5. REFERENCES

[1] Allopenna, P.D., Magnuson, J.S., Tanenhaus, M.K. 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *J Mem Lang* 38,419-439.

[2] Bybee, J. 2001. *Phonology and Language Use.* Cambridge University Press, Cambridge.

[3] Dahan, D., Tanenhaus, M.K. 2004. Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *J Exp Psychol Learn* 30, 498–513.

[4] Gaskell, G.M. 2003. Modelling regressive and progressive effects of assimilation in speech perception. *J Phonetics* 31, 447-463

[5] Gaskell, G.M., Hare, M., Marslen-Wilson, W.D. 1995. A connectionist model of phonological representation in speech perception. *Cognitive Sci* 19, 407-439.

[6] Hawkins, S. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. *J Phonetics* 31, 373-405

[7] McQueen J.M., Cutler A., Norris, D. 2006. Phonological abstraction in the mental lexicon. *Cognitive Sci* 30, 1113-1126

[8] McQueen J.M., Viebahn, M. 2007. Tracking recognition of spoken words by tracking looks to printed words. *Q J Exp Psychol* 60, 661-671.

[9] Mitterer, H. 2006. Is vowel normalization independent of lexical processing? *Phonetica* 63, 209-229.

[10] Mitterer, H,, Ernestus, M. 2006. Listeners recover /t/s that speakers lenite: Evidence from /t/-lenition in Dutch. *J Phonetics* 34, 73-103.

[11] Salverda A.P., Dahan, D., McQueen, J.M. 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition* 90, 51-89.

[12] Tanenhaus, M,K,, Spivey-Knowlton M.J., Eberhard K.M., Sedivy, J.C. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632-1634.