Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch

Natasha Warner[a,b,*], Allard Jongman[c], Joan Sereno[c], Rachèl Kemps[b,d]

[a] *Department of Linguistics, University of Arizona, PO Box 210028, Tucson, AZ 85721-0028, USA*
[b] *Max Planck Institute for Psycholinguistics, Postbus 310, NL-6500 AH Nijmegen, Netherlands*
[c] *Linguistics Department, University of Kansas, Blake Hall, Lawrence, KS 66045, USA*
[d] *Interfaculty Research Unit for Language and Speech, University of Nijmegen, Nijmegen, Netherlands*

Received 4 June 2001; received in revised form 3 March 2003; accepted 7 March 2003

Abstract

Words which are expected to contain the same surface string of segments may, under identical prosodic circumstances, sometimes be realized with slight differences in duration. Some researchers have attributed such effects to differences in the words' underlying forms (incomplete neutralization), while others have suggested orthographic influence and extremely careful speech as the cause. In this paper, we demonstrate such sub-phonemic durational differences in Dutch, a language which some past research has found not to have such effects. Past literature has also shown that listeners can often make use of incomplete neutralization to distinguish apparent homophones. We extend perceptual investigations of this topic, and show that listeners can perceive even durational differences which are not consistently observed in production. We further show that a difference which is primarily orthographic rather than underlying can also create such durational differences. We conclude that a wide variety of factors, in addition to underlying form, can induce speakers to produce slight durational differences which listeners can also use in perception.
© 2003 Elsevier Ltd. All rights reserved.

1. Introduction

Past research has indicated that under some conditions, speakers produce small differences in duration of words which, based on processes of positional neutralization in the phonology of the language, would be expected to be homophones. In this paper, we take a novel approach to the

---

*Corresponding author. Department of Linguistics, University of Arizona, PO Box 210028, Tucson, AZ 85721-0028 USA. Tel.: +1-520-626-5591; fax: +1-520-626-9014.

E-mail address: nwarner@u.arizona.edu (N. Warner).

investigation of task effects in such durational differences, and also further the study of the perception of such differences. We extend the study of such differences to Dutch, which has not been thoroughly investigated on this point. We relate the present results to sources of small, non-contrastive acoustic differences other than neutralized underlying differences, and argue that underlying forms are only one of many sources of sub-phonemic differences.

The most studied case of sub-phonemic durational differences involves final devoicing, and this is also the main focus of this study. Final devoicing is a phonological process found in a variety of languages, which neutralizes voicing contrasts that are maintained in other positions. In German, Polish, and Catalan, word-final obstruents do not contrast in voicing. Nevertheless, researchers have, at least in some cases, shown small differences in duration between underlying voiced and voiceless final obstruents, or their preceding vowels. Thus, German *Rad* 'wheel' may differ slightly from *Rat* 'advice' (both normally transcribed [ɹɑt]). For example, the vowel or duration of closure voicing may be longer, or stop closure or burst shorter, in *Rad* than in *Rat*. This phenomenon of small durational differences in supposedly neutralizing environments is usually termed 'incomplete neutralization.' However, we prefer the term 'sub-phonemic durational differences,' because we will discuss effects other than positional neutralization which behave similarly.

The striking pattern about such differences is that although they are much smaller than differences in the same acoustic intervals where voicing is not neutralized (intervocalically), the duration differences are in the same direction as in non-neutralized forms. That is, for both neutralized and non-neutralized forms, vowel duration is longer before voiced than voiceless obstruents, as is duration of voicing during stop closure, while closures and bursts are both shorter for voiced stops. This finding is rather disturbing in its implications for phonological theory because it runs contrary to the idea of categorical distinctions among segments.

Over the past 20 years, many studies have investigated the acoustics and/or perception of final devoicing in neutralizing contexts in a variety of languages. For example, Slowiaczek and Dinnsen (1985) and Port and O'Dell (1985) found reliable effects of underlying voicing in Polish and German, respectively. The Polish study (Slowiaczek & Dinnsen, 1985) used five speakers and 15 minimal pairs differing only in the underlying voicing of their final obstruent, and embedded the target words in frame sentences. This study showed an effect of underlying voicing on vowel duration (a difference of approximately 10 ms) which was significant across speakers, and consistent across types of final obstruent. The authors also find effects of underlying voicing on closure duration and closure voicing duration, but these were limited to certain speakers, environments, or final obstruents. In a study on German using a larger number of speakers (10) and words read in isolation, Port and O'Dell (1985) found effects on vowel duration (approximately 15 ms), closure voicing (approximately 5 ms), and burst duration (approximately 15 ms), all significant across speakers. Furthermore, these authors tested listeners' ability to identify the productions, and found that listeners can tell which member of the minimal pair was intended, with significantly greater than chance accuracy (59% correct overall). In her study of German, Charles-Luce (1985) embedded words in frame sentences, with position within the clause (final vs. medial) and initial phoneme of following word (vowel vs. consonant) as well as voicing manipulated. This study used both stop-final and fricative-final words. The effect of underlying voicing was significant across speakers only for vowel duration in clause-final fricatives and for closure voicing in clause-final prevocalic stops. Thus, in this study as well, significant effects of

underlying voicing were highly restricted by environment. The significant effects represent differences of approximately 10 ms for vowel duration and 5 ms for closure voicing.

Some authors suggest that effects of underlying voicing in neutralization environments are due to orthographic differences or speaking style. In general, research on the effects of orthography suggests that the larger durational effects are found for languages that represent the underlying voicing contrast in the orthography (e.g., German and Polish). In addition, for these languages, the less the experimental design emphasizes the role of orthography, the smaller the duration effects. For example, Fourakis and Iverson (1984), in a study on German, attempt to circumvent the potential orthographic effect by using two tasks, one a traditional reading task, and one in which the experimenter prompted the speaker to produce morphologically related forms of verbs, so that the speaker did not read the target words. Fourakis and Iverson find some significant effects of underlying voicing in the reading task, but none in the verb conjugation task, and conclude that incomplete neutralization occurs only when speakers try to distinguish between words with differing orthography when reading. Similarly, Jassem and Richter (1989) find no significant differences between underlying voiced and voiceless final segments in Polish when minimizing the role of orthography by prompting subjects with questions to which the target words formed obvious one-word answers. Finally, results are mixed for languages in which the underlying voicing distinction is not maintained in the orthography. Dinnsen and Charles-Luce (1984) report incomplete neutralization for Catalan (although none of the observed duration differences hold across all 5 speakers). In contrast, Kopkalli (1993) finds no significant duration differences for word-final stops in Turkish.

In perhaps the most extensive investigation of the influence of speech style on final devoicing, Port and Crawford (1989) report on the effects of underlying voicing in German using five speakers and three minimal pairs. They elicited the target words from speakers under a variety of stylistic conditions in which words were embedded in different semantically plausible sentences. Port and Crawford (1989) found that discriminant analysis to classify productions by underlying final voicing was least successful (approximately 55% correct) when target words were embedded in sentences that did not draw attention to the minimal pairs (whether read or repeated orally). It was most successful (78% correct) when speakers dictated the words to the experimenter. Port and Crawford (1989) conclude that although there are effects of underlying final voicing that are not limited to careful speech, speakers can make more or less clear differences in the pairs depending on speech style. They also point out that, perceptually, listeners can make use of even the differences produced in less careful speech. They suggest that many very small acoustic differences, rather than a few large ones, are involved in making the distinction.

Charles-Luce (1993), in a study on Catalan, introduces a new aspect to the question of speech style. Stimuli were embedded either in a semantically disambiguating or neutral context. With semantic information present, few durational differences were significant, and these were limited to just one minimal pair and to a particular phonological environment. However, in the sentences lacking semantically biasing information, there was a significant effect on vowel duration (average 15 ms difference) regardless of environment for three of the five minimal pairs. Charles-Luce (1993) concludes that only when speakers read sentences lacking disambiguating semantic information are there durational differences between the members of a minimal pair.

Although we will limit the discussion here to final devoicing, the possibility of incomplete neutralization has been investigated for several other cases as well, with equally mixed results.

Lahiri, Schriefers, and Kuijpers (1987) find complete neutralization of vowel length in Dutch, Kim and Jongman (1996) find complete neutralization of manner in Korean. Dinnsen (1985) discusses evidence for incomplete neutralization of English flapped /t/ and /d/.

1.1. Goals of this paper

While past research has shown that underlying voicing can affect duration in a neutralization environment in ways that are relatively consistent across speakers and minimal pairs, the literature has also shown that such effects are usually small, and may be dependent on speech style or context. The present research extends the study of such effects along several lines. First, results from a large-scale acoustic investigation of final devoicing in Dutch are presented (Section 2). Second, several perception experiments are then conducted. Past studies of perception of durational differences in final devoicing have been limited to presenting listeners with actual productions of the minimal pairs and asking them to identify the intended word. We perform such a test, but also investigate the relationship between production and perception and the influence of speaker-specific production differences (Section 3). Specifically, stimuli were selected from four of the original speakers who varied in the extent to which vowel and consonant duration were affected by the voicing of the word-final consonant. This allows us to explore the extent to which listeners distinguish the voicing of the word-final consonant as a function of which acoustic cues to the distinction are present and consistent in the speech signal. We further investigate these durational differences by performing a perception experiment to specifically test the most likely acoustic cue to the underlying voicing distinction as determined by the production data, allowing stimuli to differ only in vowel duration (Section 4). In a further experiment, we allow stimuli to vary only in consonant closure duration, an acoustic characteristic that should not, according to the production data, serve as a cue (Section 5).

Finally, we investigate orthographic influence in a new way. Past research has attempted to determine whether durational differences could be due to orthographic influence by prompting speakers to produce speech without reading. However, highly literate speakers may activate orthographic representations of words when they are prompted to say them, even if they are not reading them at the time. Therefore, we take the opposite approach to the possible influence of orthography by investigating words with the same string of phonemes that differ only in terms of spelling, but not in terms of underlying form (Section 6). In sum, we take a novel approach to the question of durational differences in final devoicing, and we relate durational differences caused by underlying form to other sub-phonemic durational differences in support of our argument that underlying form is only one source of such differences.

2. Experiment 1: production of final devoicing in Dutch

The present experiment investigates final devoicing in Dutch. Dutch has a final devoicing process similar to that in the languages discussed above. For example, /met/ 'measures (sg.)' and /med/ 'avoided (sg.)' would both be transcribed [meⁱt], although voicing is distinguished intervocalically in /metən/ [meⁱtən] 'to measure' and /medən/ [meⁱdən] 'avoided (pl.).' Unlike German, Polish, Catalan, and Turkish, there has been little thorough investigation of final

devoicing and duration in Dutch. Jongman, Sereno, Raaijmakers, and Lahiri (1992) mention that pilot results show no effect of underlying voicing of final obstruents in Dutch minimal pairs. Baumann (1995) presents acoustic measurements and a perceptual test in which listeners are asked to identify members of devoicing minimal pairs. She finds that listeners' identification is at chance accuracy. Her acoustic results showed only one significant effect on duration (closure duration), and that for only one syntactic condition, with no corresponding effect on perception. Baumann concludes that Dutch lacks an incomplete neutralization effect. Ernestus and Baayen (in press) find longer burst duration for non-words spelled with final 'p, t' than those with 'b, d,' and find evidence that listeners can hear acoustic differences in final voicing and use these to hypothesize which past tense allomorph non-words would take. As with the literature on German and other languages, the results for Dutch final devoicing appear rather inconsistent.

Jongman et al. (1992) report only pilot results on neutralization as part of a different study, and Baumann's (1995) primary interest is in the effect of a following clitic rather than incomplete neutralization. As with studies on other languages, the number of speakers in all of the past studies on Dutch was small, which may explain the inconsistency of the effects found. We first wished to perform a thorough acoustic study to establish whether the underlying voicing of final stops has any effect in a neutralization environment in Dutch. Since past work found that Dutch might lack this effect (Jongman et al., 1992; Baumann, 1995), and since the effects even for other languages are often small, we used a larger number of speakers and items than previous studies in other languages in order to assure that any reliable effects would indeed be detectable.

2.1. Methods

Using the CELEX database (Baayen, Piepenbrock, & van Rijn, 1993), we selected 10 minimal pairs with phonemically short vowels and 10 with phonemically long vowels, with each pair consisting of one Dutch word ending in underlying /t/ and one in /d/ (Table 1). There are very few minimal pairs in Dutch for final /p/ vs. /b/, and there is no voiced velar stop in Dutch except in a few recent loanwords. Therefore, the materials for the studies reported here use exclusively alveolars for the relevant consonant. Although there is no reason to predict a difference in neutralization for phonemically long and short vowels, it is possible that the effect on vowel duration would be larger for one length or the other, so we included this as a factor. The selected words also appear in suffixed forms which make the underlying voicing of the stop clear (e.g., *rat, ratten* 'rat, rats'; *wed, wedden* 'bet (1st sg.), bet (pl.)'), with the exception of *wat, Ad, KID*. We also selected 16 minimal pairs in which the /t/ or /d/ occurs intervocalically, that is where the voicing distinction is maintained, in order to compare any durational effects in the neutralization environment to the corresponding effects in the distinction environment. Ten of these had phonemically long vowels and six (as many pairs as could be found) had short vowels.

Fifteen native Dutch speakers participated in the experiment. Most were students at the University of Nijmegen. All spoke at least some English, but all aspects of the recording session were conducted in Dutch by a native Dutch-speaking experimenter (the fourth author). Some past literature on incomplete neutralization has suggested that it is preferable to use speakers who do not have substantial exposure to a language which lacks final devoicing, such as English. The subjects for all experiments in this paper were college students, and thus presumably had reasonable competence in English, as is typical for young educated Dutch speakers. Even though

Table 1
Items for Experiment 1

Neutralization environment, short vowels

at /ɑt/	'ate sg.'	Ad /ɑd/	proper name
bet /bɛt/	'dab sg.'	bed /bɛd/	'bed'
bit /bɪt/	'bit' (horse or computer)	bid /bɪd/	'pray 1st-sg.'
bot /bɔt/	'bone'	bod /bɔd/	'offer'
kit /kɪt/	'sealant'	KID /kɪd/	'artificial insemination with a donor' (abbrev.)
pat /pɑt/	'stalemate'	pad /pɑd/	'path'
rat /rɑt/	'rat'	rad /rɑd/	'wheel'
schut /sxʏt/	'lock'	schud /sxʏd/	'shake 1st-sg.'
wat /wɑt/	'what'	wad /wɑd/	'mudflat'
wet /wɛt/	'law'	wed /wɛd/	'bet 1st-sg.'

Neutralization environment, long vowels

baat /bat/	'benefit'	baad /bad/	'bathe 1st-sg.'
biet /bit/	'beet'	bied /bid/	'offer 1st-sg.'
boot /bot/	'boat'	bood /bod/	'offered 1st-sg.'
eet /et/	'eat sg.'	eed /ed/	'oath'
meet /met/	'measure sg.'	meed /med/	'avoided 1st-sg.'
moet /mut/	'must sg.'	moed /mud/	'courage'
noot /not/	'nut'	nood /nod/	'necessity'
smeet /smet/	'threw sg.'	smeed /smed/	'forge 1st-sg.'
voet /vut/	'foot'	voed /vud/	'feed 1st-sg.'
zweet /zwet/	'sweat'	Zweed /zwed/	'Swede'

Distinction environment, short vowels

bitten /bɪtən/	'bits'	bidden /bɪdən/	'to pray'
kutten /kʏtən/	vulgar term	kudden /kʏdən/	'herds'
patten /pɑtən/	'stalemates'	padden /pɑdən/	'toads'
schutten /sxʏtən/	'to pass through a lock'	schudden /sxʏdən/	'to shake'
watten /wɑtən/	'cotton batting'	wadden /wɑdən/	'mudflats'
wetten /wɛtən/	'laws'	wedden /wɛdən/	'to bet'

Distinction environment, long vowels

baatte /batə/	'availed sg.'	baadden /badən/	'bathed pl.'
boten /botən/	'boats'	boden /bodən/	'offered pl.'
goten /xotən/	'drainpipes'	goden /xodən/	'gods'
heten /hetən/	'to be named'	heden /hedən/	'present day'
maten /matən/	'buddies'	maden /madən/	'maggots'
meten /metən/	'to measure'	meden /medən/	'avoided pl.'
smeten /smetən/	'threw pl.'	smeden /smedən/	'to forge'
voeten /vutən/	'feet'	voedden /vudən/	'fed pl.'
zaten /zatən/	'sat pl.'	zaden /zadən/	'seeds'
zweetten /zwetən/	'sweated pl.'	Zweden /zwedən/	'Sweden'

Infinitive forms of some words, and singular past tense forms of others, such as *baten* and *baden*, could have been used instead of these orthographically more complex forms. *Baden* and *baadden* are probably the same string of phonemes, as discussed with regard to Experiment 5. The durational differences in pairs such as *baden* and *baadden* (Experiment 5) are small relative to differences between *baten* and *baden*, and thus the choice to have speakers produce *baden* vs. *baadden* in Experiment 1 should not affect the comparison between /t/ and /d/. The reason for use of the orthographically more complex forms is that these same items were embedded in texts in a related experiment not reported here, and it was sometimes necessary to use a particular person or number in order to embed the word in the text.

Dutch speakers have some competence in English, they may not have a good command of the English final voicing distinction. Nevertheless, all experiments were conducted entirely in Dutch, by a native Dutch speaker, and in the Netherlands, to minimize English influence. Dialect was not controlled, since there are no known effects of dialect on final devoicing in Dutch.

The speakers read, two times each, word lists containing the experimental items in a pseudo-random order. Filler items were included at the beginning and end of lists to avoid list intonation effects. The members of a minimal pair never followed each other directly in a list, to prevent subjects from producing an emphatic difference between words as they would if dictating them in pairs. Speakers were familiarized with the words in natural contexts before being recorded. All recordings were made using a high quality DAT recorder, with the speaker seated in a sound-treated booth. A total of 1080 productions of minimal pairs were recorded: 300 with short vowels and 300 with long vowels in the neutralization context, and 180 with short vowels and 300 with long vowels in the distinction context.

The speech was digitized at a sampling rate of 16,000 Hz, and all measurements were made using the XWaves software. The following intervals were measured: duration of the (first) vowel up to the closure for the stop, duration of voicing during the closure, duration of closure, and duration of burst. For /d/ in the distinction (intervocalic) environment, closure voicing duration is equivalent to closure duration.

Onset of the vowel was defined as onset of voicing if the end of the preceding segment was voiceless, or the end of the burst if the vowel followed a fully voiced stop. For vowels following nasals, the sudden discontinuity in the spectrogram was taken as the boundary. For vowels after /w/, onset of the second formant was identified as onset of the vowel, and for vowels after /r/, the end of the final tap of the trill was chosen. The end of the vowel and beginning of closure was defined as the end of the second formant of the vowel, which usually coincides with a sudden drop in amplitude of voicing. The end of closure voicing was defined as the end of periodic vibration in the waveform. The end of closure was located at the onset of the sudden discontinuity in the waveform for the burst. If no burst was visible in the waveform, as was sometimes the case for intervocalic /d/, the spectrogram was examined for a sudden wide-band noise, which was defined as the burst. In a few cases, intervocalic /d/ was produced as a glide with no burst. Changes in amplitude and formant frequency were then used to define the duration of the consonant, as no 'closure' occurred, and the burst duration was considered to be 0 ms. Although these cases cannot be measured as accurately as those with clear closures and bursts, they are restricted to the intervocalic (distinction) environment, so they will not affect the accuracy of the data for the final devoicing conditions. Aside from these cases, the end of the burst was taken to be the onset of voicing for the following schwa for intervocalic /t/, the end of the broadband burst noise in the spectrogram for intervocalic /d/, and the end of visible noise in the spectrogram for final stops.

2.2. Results

Measurements were averaged across the two productions from each speaker, as well as across items for the by subjects statistical analyses, and across speakers for the by items analyses. Average durations appear in Table 2. ANOVAs were conducted separately for neutralization and distinction environments, since the effect of neutralization vs. distinction is expected to be far

Table 2
Average durations for all measurements of Experiment 1 (ms). Durations are shown for neutralization and distinction environments, by phonemic vowel length and underlying voicing. Closure voicing duration is omitted for voiced stops in the distinction environment, because voicing continues throughout the closure in these stops

Measurement	Underlying voicing	Neutralization environment		Distinction environment	
		Short V	Long V	Short V	Long V
Vowel duration	Voiceless	120	175	83	162
	Voiced	124	178	103	184
Closure duration	Voiceless	82	73	61	60
	Voiced	80	72	32	35
Burst duration	Voiceless	139	131	46	48
	Voiced	136	122	16	16
Closure voicing duration	Voiceless	22	28	22	24
	Voiced	23	27		

larger than any effect of underlying voicing in the neutralization environment. Vowel duration, closure voicing duration, closure duration (including closure voicing), and burst duration were the dependent variables. Each ANOVA had the two independent variables Underlying Voicing and Phonemic Vowel Length, with subjects ($F1$) or items ($F2$) as the repeated measure.

Underlying Voicing of the stop significantly affected vowel duration in both neutralization and distinction environments, with vowels before /d/ 3.5 ms longer than before /t/ in the neutralization environment ($F1(1,14) = 9.39$, $p < 0.01$, $F2(1,18) = 29.54$, $p < 0.001$), and more than 20 ms longer in the distinction environment ($F1(1,14) = 145.02$, $p < 0.001$, $F2(1,14) = 683.17$, $p < 0.001$). Not surprisingly, Phonemic Vowel Length also significantly affected vowel duration in both environments (neutralization: $F1(1,14) = 390.27$, $p < 0.001$, $F2(1,18) = 12.18$, $p < 0.005$; distinction: $F1(1,14) = 507.28$, $p < 0.001$, $F2(1,14) = 29.90$, $p < 0.001$), with an effect much larger than that of Underlying Voicing. The interaction of Underlying Voicing and Phonemic Vowel Length was not significant for either environment.

Closure duration was significantly longer for words with phonemically short (81 ms) than long (72 ms) vowels in the neutralization environment ($F1(1,14) = 24.05$, $p < 0.001$, $F2(1,18) = 21.08$, $p < 0.001$), but there was no main effect of Underlying Voicing, nor an interaction. In the distinction environment, there was a significant interaction of Underlying Voicing and Length in the by subjects analysis ($F1(1,14) = 5.91$, $p < 0.03$), with closures for /t/ (60 ms) considerably longer than for /d/ (33 ms), but slightly more so for short vowels ($F1(1,14) = 144.67$, $p < 0.001$) than long vowels ($F1(1,14) = 135.59$, $p < 0.001$). In the by items analysis, only the main effect of Underlying Voicing was significant ($F2(1,14) = 431.75$, $p < 0.001$).

Burst duration in the neutralization environment showed a significant interaction ($F1(1, 14) = 4.71$, $p < 0.05$, $F2(1, 18) = 7.85$, $p < 0.02$), and tests of simple effects revealed a significant effect of Underlying Voicing for long vowels, with /t/ bursts (131 ms) 9 ms longer than /d/ bursts (122 ms) ($F1(1, 14) = 6.44$, $p < 0.03$, $F2(1, 9) = 74.42$, $p < 0.001$), but no significant effect for short vowels. In the distinction environment, only Underlying Voicing significantly affected burst duration ($F1(1,14) = 150.52$, $p < 0.001$, $F2(1,14) = 384.31$, $p < 0.001$), with voiceless bursts (47 ms) more than 30 ms longer than voiced ones (16 ms).

Closure voicing duration is only meaningful for the neutralization environment, since /d/ is fully voiced in the distinction environment. In the neutralization environment, there is a significant effect only of Vowel Length ($F1(1,14) = 13.35$, $p < 0.005$, $F2(1,18) = 46.00$, $p < 0.001$), with more closure voicing after long vowels.

2.3. Discussion

Vowel duration (3.5 ms longer for /d/) and burst duration for the phonemically long vowels only (9 ms longer for /t/) show significant effects of underlying voicing of the stop in the neutralization environment, but no other measures do. Both of these effects are in the expected direction, the same direction as in the distinction environment. In the neutralization environment, effects of phonemic vowel length are often much larger than effects of underlying voicing, even where there is no strong reason to predict an effect of phonemic vowel length (i.e., closure voicing duration). Also, since there is no reason to predict that the effect on burst duration would be limited to words with phonemically long vowels, it seems that the effect on vowel duration is the only one which is consistent. Still, it is clear that there are reliable, if small, effects of underlying voicing on duration in the neutralization environment.

The effect on vowel duration is considerably smaller than that found for other languages. While past studies find significant effects of 10–15 ms on vowel duration (Port & O'Dell, 1985; Slowiaczek & Dinnsen, 1985), we find a difference in Dutch of just 3.5 ms. The fact that such a small difference is significant both by subjects and items suggests that the large number of speakers and items makes the study very sensitive, and thus that reliable patterns in the data are unlikely to have been missed. Our finding of durational effects of underlying voicing in neutralization environment in Dutch contradicts some past studies (Jongman et al., 1992; Baumann, 1995). Both of these were parts of studies on other topics. Our results suggest that quite a large study, in both number of subjects and items, is necessary in order to confirm the presence of such effects in a language.

Whalen (1991, 1992) has found that low-frequency words may be pronounced more slowly, with longer durations, than high-frequency words. We could not control for word frequency in our materials because nearly all suitable minimal pairs of the language were used. It is possible, therefore, that the durational differences we find could be word frequency effects in disguise, rather than effects of underlying final voicing. However, in that case, one would expect lower frequency words to have greater durations for both vowels and consonants. The opposing direction of the effects on vowel duration and burst duration argues against such an explanation.

3. Experiment 2: perceptibility of final voicing

Most previous studies of the perception of final voicing indicate that listeners can make use of the small durational differences to distinguish final voicing in neutralization position (Port & O'Dell, 1985; Port & Crawford, 1989). This is clearly not a distinction in the usual linguistic sense of contrasting segments, since listeners' accuracy is not very good, but is significantly better than chance. We performed a simple perception experiment to test whether listeners could use the acoustic differences found in our Dutch data. Listeners heard productions from the neutralization

environment, and were asked to identify which member of the pair (e.g., *eet* 'eat sg.' or *eed* 'oath') they had heard. Productions from a range of speakers were chosen to examine the relationship of perception to production.

3.1. Methods

In order to limit the size of the experiment, four of the 15 speakers from Experiment 1 were chosen, representing a variety of results in their production data. Speaker 3 was chosen for having a relatively large difference in vowel duration, and the largest difference of any speaker for burst duration, with both differences in the same direction as the average results. Speaker 5 had the largest effect on vowel duration of any speaker, but a relatively small effect on burst duration (not significant when tested for that speaker individually, $F(1, 18) = 2.11$, $p = 0.164$). Speaker 6 was selected for having no difference at all in vowel duration, and furthermore had a small effect in the opposite direction from the average for burst duration. Speaker 14 had effects in the opposite direction from the average for both vowel and burst duration.

All productions of words in a neutralization environment (i.e., long and short vowels, both productions of each word) from these four speakers were presented to listeners as stimuli. Stimuli were blocked by speaker to give listeners maximal chance to make use of speaker-specific cues. Within each speaker block, stimuli were presented in a pseudo-random order, with neither the two productions of a word nor members of a minimal pair immediately following each other. Thirty subjects participated in the perception experiment, 14 of the subjects who had been speakers in Experiment 1, including all four whose productions constituted the stimuli, and 16 additional subjects (also native Dutch speakers). Subjects were instructed that they would hear four speakers producing words ending in 't' or 'd,' and that they should respond with which word they thought they had heard. Those subjects who had also participated in Experiment 1 were informed that their own voice might occur.

Subjects were given a list of the experimental items with which to familiarize themselves. The experimenter explained low-frequency items and pointed out that some items are inflected forms. Subjects were tested individually in sound-treated booths, and stimuli were presented over headphones. Stimulus presentation and collection of results was controlled using the NESU (Nijmegen Experimental Setup) software (Wittenburg, Nagengast, & Baumann, 1998). As each stimulus was presented, the two words of the minimal pair appeared in Dutch orthography on a computer screen to remind the subject what the two possible words were. Subjects responded by pressing buttons labeled 't' and 'd.' Half of the subjects had 't' as the left button and half had 'd,' and the order of the two words on the screen was matched to the order of buttons on the button box. This experiment took place approximately 2–3 months after Experiment 1.

3.2. Results

Responses were averaged across the two productions of each, and across items for the by subjects analysis and across subjects for the by items analysis. ANOVAs were then conducted on the proportion of 't' responses, with the factors Intended Voicing of the final stop, Speaker (which speaker produced the stimuli), Phonemic Vowel Length, and Experience (whether or not the

Table 3
Percent "t" responses, by intended underlying voicing and speaker

Speaker	Intended final /t/	Intended final /d/
3	62.6	50.1
5	61.8	50.3
6	54.1	55.3
14	56.9	60.8

listener had also participated in the production experiment).[1] All are repeated measures factors except the Vowel Length factor in the by items analysis and the Experience factor in the by subjects analysis. The main effect of Intended Voicing was significant ($F1(1,28) = 22.06$, $p < 0.001$, $F2(1,18) = 16.31$, $p < 0.005$), as was the interaction of Intended Voicing and Speaker ($F1(3,84) = 16.72$, $p < 0.001$, $F2(3,54) = 9.93$, $p < 0.001$). Results appear in Table 3. The Speaker by Vowel Length interaction was also significant ($F1(3,84) = 10.30$, $p < 0.001$, $F2(3,54) = 9.08$, $p < 0.001$). The interaction of Intended Voicing by Vowel Length did not quite reach significance either by subjects or by items ($F1(1,28) = 3.92$, $p = 0.058$, $F2(1,18) = 3.12$, $p = 0.094$). No other main effects or interactions reached significance either by subjects or by items.

Examination of the means by Speaker and Vowel Length showed that the significant interaction of these factors stems from productions by speakers 5 and 14 receiving more 't' responses to short (58% and 63%, respectively) than long vowels (54% and 55%), while speaker 6 received more 't' responses to long (58%) than short (51%) vowels, and speaker 3 showed little effect of Vowel Length. Means for all combinations of Intended Voicing and Vowel Length show that the near-significant interaction of these factors reflects a difference in the size of the Intended Voicing effect: Both long and short vowel words receive more 't' responses if they were intended to be /t/-final words, but this effect is larger for long vowels (59% 't' responses to /t/ and 52% to /d/) than short ones (58% and 56%, respectively).

Since the interaction of Intended Voicing by Vowel Length was not quite significant and the interaction of Speaker by Vowel Length does not involve the Intended Voicing factor, we focus in further analyses on the interaction of Speaker by Intended Voicing (Table 3). We collapsed across the factors of Vowel Length and Experience by treating all items and all subjects as equivalent, regardless of whether the item had a long or a short vowel and of whether the subject had participated in Experiment 1 or not (i.e., the subsequent tests of simple effects had 20 items in the by items analyses and 30 subjects in the by subjects analyses). We then performed tests of the simple effect of Intended Voicing for each of the four speakers separately. We also calculated d' for each speaker, in order to give a direct measure of perceptibility of Intended Voicing. For speakers 3 (who had both vowel and burst duration differences) and 5 (who had a vowel duration

---

[1] Although signal detection analyses are often used for this type of data, we prefer the use of ANOVA with proportion of 't' responses as the dependent variable. By analyzing proportion of 't' responses, rather than proportion correct, both hit rate and false positive rate are included in the analysis, and therefore bias is accommodated, just as in a signal detection analysis. The choice of dependent variable can prevent a problem with bias. We have included d' for the overall results for each speaker, however.

difference), there were significantly more 't' responses to the words which were intended to have final /t/ (63% and 62%, respectively) than to those intended to have final /d/ (50% for each of these speakers) (speaker 3: $F1(1,29) = 18.62$, $p < 0.001$, $F2(1,19) = 18.44$, $p < 0.001$, $d' = 0.33$; speaker 5: $F1(1,29) = 44.76$, $p < 0.001$, $F2(1,19) = 13.44$, $p < 0.005$, $d' = 0.31$). For speaker 6 (no vowel duration difference), the effect of Intended Voicing was not significant ($F1$ and $F2 < 1$, $d' = -0.03$). For speaker 14 (vowel duration difference in opposite of average direction), there were more 't' responses to words which were intended to have final 'd' (61%) than to those intended to have final 't' (57%) (i.e., listeners responded in the opposite direction from the speaker's intention). This effect was significant by subjects but not by items ($F1(1,29) = 6.86$, $p < 0.02$, $F2(1,19) = 2.16$, $p > 0.05$, $d' = -0.10$).

Among those subjects who participated in Experiment 1, it is also interesting to determine whether the pattern in the speaker's own productions affects that individual's perceptual judgments. For example, are speakers who themselves produce longer vowels before /d/ than /t/ better able to distinguish their own and other speaker's productions than speakers who produce little difference in vowel duration, or even a difference in the opposite direction? To determine this, we conducted a separate analysis of just the 14 subjects who had participated in Experiment 1. We divided these subjects into three groups based on size and consistency of effect on vowel duration in Experiment 1, by testing the significance of the effect on vowel duration in the production data for each speaker separately. Five speakers showed significant effects on vowel duration (all with vowels longer before /d/ than /t/), four speakers showed non-significant effects in the same direction, and five speakers had either no difference in vowel duration at all or a non-significant difference in the opposite direction.

To test the relationship of production and perception, we used a three factor ANOVA, with the factors Production Difference (vowels significantly longer before /d/, non-significantly so, or not so), Intended Voicing, and Speaker (which speaker produced the stimuli). All are repeated measures factors except the Production Difference factor in the by subjects analysis. The main effect of Intended Voicing was significant ($F1(1,11) = 19.68$, $p < 0.005$, $F2(1,19) = 11.78$, $p < 0.005$), as was the interaction of Intended Voicing by Speaker ($F1(3,33) = 10.12$, $p < 0.001$, $F2(3,57) = 7.53$, $p < 0.001$). The main effect of Production Difference reached significance only by items ($F1(2,11) = 1.83$, $p = 0.206$, $F2(2,38) = 3.99$, $p < 0.03$). No interactions involving Production Difference reached significance either by subjects or by items.

The means by Production Difference indicate that listeners who in their own speech produce significantly longer vowels before /d/ than /t/ are, overall, the most likely to respond with 't' in the perception experiment. Listeners who themselves produce longer vowels before /t/ than /d/ are the least likely to respond with 't.' The means by Production Difference and Intended Voicing suggest that listeners who themselves produce larger or more consistent differences in vowel duration may be more sensitive to the distinction as produced by others. Listeners who produce significantly longer vowels before /d/ than /t/ respond 't' to 64% of /t/ stimuli and 57% of /d/ stimuli, whereas the results for those who have a non-significant vowel duration difference in the same direction are 59% and 56%, and the results for those with vowels longer before /t/ are 57% and 53%. However, since neither the interaction of the Production Difference factor with the Intended Voicing factor nor the three-way interaction are significant, we cannot draw a strong conclusion about this link between production and perception. The interaction of Intended Voicing and Speaker has been discussed in the four factor analysis above.

3.3. Discussion

Both the four factor analysis using all 30 subjects, and the three factor analysis using only those subjects for whom production data are also available, show a significant interaction of Intended Voicing with Speaker and no other interactions or main effects with import for the perception of final voicing. Tests of the simple effect of Intended Voicing for each of the four speakers separately indicate that listeners are able to distinguish words with final /d/ and /t/ at significantly better than chance accuracy for speakers 3 and 5, the two speakers who produced longer vowels before final /d/ than /t/ (a vowel duration difference in the expected direction). The d' for even these speakers is quite small, even though the effect of Intended Voicing is significant for both. This confirms that this difference, although significant, is quite small, and is not on the order of a typical phonemic distinction which listeners can distinguish very accurately. Listeners cannot distinguish the productions of speaker 6, who produced no difference in vowel duration before /d/ and /t/ at all. Interestingly, listeners distinguish among productions of speaker 14 in the opposite direction from the speaker's intentions in producing the words. That is, listeners more often perceive words as ending in /t/ if speaker 14 was producing the /d/ member of the minimal pair. Productions of speaker 14 have a small effect in the opposite of the usual direction for both vowel duration and burst duration. Because the stimuli in this experiment vary in many ways, we cannot be sure from these results which cues listeners are using, but comparison across speakers suggests that at least vowel duration is a likely cue.

A particularly interesting result of this experiment lies in the lack of significance of many effects. First, the results show that listeners are equally sensitive to differences between final /t/ and /d/ whether they have had extensive previous experience with these pairs by virtue of participating in the production experiment or not. Second, although the vowel duration difference (of very similar absolute size for both long and short vowels) constitutes a greater proportion of the duration of the vowel for short than long vowels, the lack of an interaction of Intended Voicing with Vowel Length suggests that there are useful perceptual cues to underlying final voicing in both words with short and long vowels. Finally, we cannot conclusively claim that individual speaker's results in the production experiment are related to their ability to perceive the distinction as listeners. That is, listeners who themselves produce larger differences between underlying final /t/ and /d/ were not consistently better (although there is a trend in the expected direction) at distinguishing final /t/ and /d/ perceptually than listeners who, in their own speech, produce little difference or a difference in the opposite direction from the average.

4. Experiment 3: perception when vowel duration is manipulated

Although Experiment 2 shows that listeners can make use of small durational differences to distinguish underlying final voicing with better than chance accuracy, it does not show what cues are involved in doing so, since it allows all potential cues to vary at once. In this experiment we manipulated only the most likely perceptual cue, vowel duration, in order to determine whether it alone can serve as a useful perceptual cue to underlying final voicing.

4.1. Methods

We created vowel duration continua from productions taken from Experiment 1 by splicing a single period of the vowel into the signal repeatedly or by removing periods from the vowel, and asked listeners to identify the resulting stimuli as either the /t/- or /d/-final member of the minimal pair. We made such continua from four words from each of the four speakers used in Experiment 2 above, a /t, d/ pair of words with a phonemically long vowel and a pair with a short vowel (e.g., *smeet, smeed, wat, wad*), resulting in a total of 16 continua.[2] The two pairs of words for each speaker were selected to have several periods in the vowel during which formants, f_0, and amplitude remained as steady as possible, in order to avoid clicks in spliced stimuli. Furthermore, the four words from each speaker were chosen to have the f_0 values during that steady region within 10 Hz of the other words for that speaker, so that the size of continuum steps created by splicing single periods (approximately 5 ms) would not vary greatly.

For each of these 16 words, a single period in the middle of the steady region was selected, and was spliced into the signal an additional one to five times. To create the shorter stimuli, first that period, then additional periods up to a total of five, were removed from the signal. This generated an 11-step continuum with the original production in the middle of the continuum, and with the endpoints of the continuum differing by approximately 50 ms in vowel duration. The stimuli were all free of noticeable splicing clicks, and sounded like natural productions of the word from which they were made. The second author, a native Dutch speaker, confirmed that none of the stimuli crossed a phonemic boundary, for example that none of the stimuli made from words with phonemically short vowels sounded as if they had phonemically long vowels.

Although the 50 ms vowel duration range of the continua is much greater than the average vowel duration difference between /d/- and /t/-final words, it is well within the range of vowel duration variability for these words. For each pair of words, the difference between the vowel durations of the productions at the 5th and 95th percentiles was determined. The average vowel duration range for the pairs of words used in this experiment, across speakers and productions, was 108 ms. No word pair had a vowel duration range of less than 90 ms. This variability reflects primarily variability in vowel durations of both members of the pair, rather than a large effect of final /d/ vs. /t/ for a few productions, as demonstrated in Fig. 1.

Sixteen listeners participated in this experiment, all native Dutch speakers who had not participated in either Experiment 1 or 2. Stimuli were blocked by speaker and continuum, e.g., all stimuli created from the word *smeet* as produced by speaker 3 were presented in a block, and all four continuum blocks made from productions of speaker 3 were presented together. Each stimulus was presented five times, and the order of stimuli within each block was randomized. Subjects were given a list of the words which appeared in this experiment to familiarize themselves with. They were instructed that the stimuli would sound rather similar, and that the task would thus be difficult, but that they should respond with the word they thought the stimulus sounded

---

[2] The manipulation of primary interest in this experiment is the underlying final voicing, but it is important to control for any potential effect of phonemic vowel length as well. Listeners might be more sensitive to continuum differences in phonemically short vowels, where durational differences represent a greater proportion of the vowel, although Experiment 2 showed no such effect.
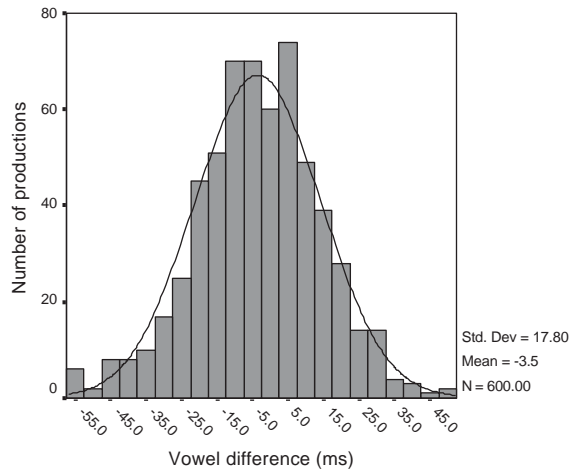
Fig. 1. Histogram showing the distribution of differences between vowel duration of a word with underlying final /t/ and vowel duration of a production of the matched word with final underlying /d/ by the same speaker. (Negative numbers indicate that the vowel is longer before /d/ than /t/.)

more like. Procedures for presentation of stimuli and collection of responses were otherwise as in Experiment 2.

4.2. Results

The percentage of each listener's 't' judgments to each stimulus (out of five presentations) was calculated, and an ANOVA was conducted on percent 't' judgments, with the factors Continuum Step (11 levels), Underlying Voicing of word from which the continuum was made (/t, d/), Phonemic Vowel Length (short, long), and Speaker from whose production the continuum was made (four levels). All are repeated measures factors. All four main effects were significant (Step: $F(10,150) = 16.40$, $p < 0.001$; Underlying Voicing: $F(1,15) = 7.76$, $p < 0.02$; Length: $F(1,15) = 5.29$, $p < 0.04$; Speaker: $F(3,45) = 3.99$, $p < 0.02$). The Continuum Step by Vowel Length interaction was significant ($F(10,150) = 5.50$, $p < 0.001$), as was the three-way interaction of Speaker by Underlying Voicing by Vowel Length ($F(3,45) = 5.84$, $p < 0.005$). The interaction of Speaker by Underlying Voicing ($F(3,45) = 3.26$, $p < 0.04$) was also significant. No other interactions reached significance.

Because the Continuum Step by Vowel Length interaction was significant, we conducted one-factor ANOVAs for the long and short vowel continua separately, collapsed across Speaker and Underlying Voicing (Fig. 2). The simple effect of Continuum Step was significant for both the long vowels ($F(10, 150) = 5.42$, $p < 0.001$) and the short ($F(10, 150) = 17.54$, $p < 0.001$). In both cases, stimuli with longer vowel duration received fewer 't' responses, i.e., a longer vowel was perceived as more likely to precede an underlying /d/. The significant interaction derives from the Continuum Step effect being stronger for the short than for the long vowels. Furthermore, the linear trend component of the Continuum Step factor was significant for both long vowels ($F(1, 15) = 12.82$, $p < 0.005$) and short ($F(1, 15) = 30.05$, $p < 0.001$), and none of the higher-order
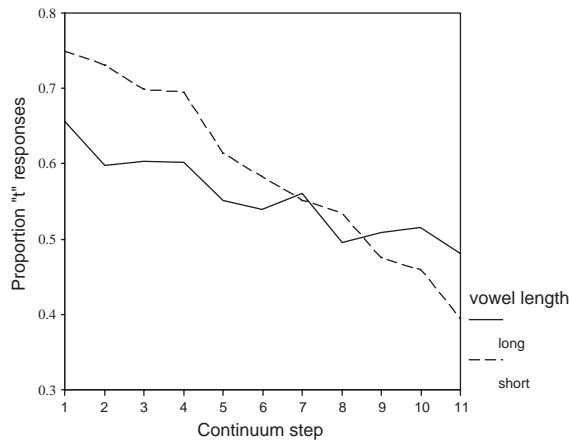
Fig. 2. Proportion of 't' responses, by continuum step in the vowel duration continuum and phonemic vowel length. Step 1 has the shortest vowel duration and step 11 the longest, with the original production at step 6.

trend components were significant. Trend analysis, used within an ANOVA, is a specific type of analytical comparison of the treatment means. It can be used to determine whether measurements increase or decrease in a linear fashion across conditions (Keppel, 1991). The lack of higher-order trend components shows that the data curve does not have any significant quadratic, cubic, or other such shape.

To investigate the Underlying Voicing by Speaker by Phonemic Vowel Length interaction, we performed two-factor ANOVAs (Underlying Voicing by Length) for each speaker separately. For speaker 3, only the main effect of Underlying Voicing was significant ($F(1, 15) = 7.51$, $p < 0.02$), with more 't' responses to words in which the speaker intended to have final /t/. For speaker 6, the interaction of Underlying Voicing and Length was significant ($F(1, 15) = 6.59$, $p < 0.03$), and tests of simple effects showed significantly more 't' responses to intended /t/ words for phonemically short vowels ($F(1,15) = 4.90$, $p < 0.05$) but a significant effect in the opposite direction for long vowel words ($F(1,15) = 4.62$, $p < 0.05$). For speakers 5 and 14, neither the main effects of Length or Underlying Voicing nor the interaction was significant.

4.3. Discussion

The significant effect of Continuum Step confirms that listeners can use vowel duration as a cue to underlying final voicing, when vowel duration is the only cue that is allowed to vary. As one would expect from the production results, longer vowel duration is associated with final /d/. By forcing listeners to choose between members of minimal pairs, and blocking stimuli by both speaker and word, we offer listeners the best possible conditions for using such a perceptual cue. Considering the small average size of the vowel duration difference found in Experiment 1 (3.5 ms), it seems unlikely that listeners make use of this cue in perceiving natural speech, but this experiment shows that under the best of conditions, they are able to do so.

The current experiment is an unusual use of identification continua, in that the endpoints of a continuum do not represent linguistically distinct categories in the usual sense. All members of a

given continuum sound like productions of the same two words, which would normally be considered homophonous. It is thus all the more impressive that Continuum Step has a significant effect on responses. The significance of the linear trend component of the Continuum Step effect, and lack of significance of any higher-order component, confirms that the response data across continuum step are linear. As is clear in Fig. 2, the results do not resemble a categorical perception curve. Although these results could potentially represent the middle, most linear appearing, portion of a categorical perception curve, we suspect that one would have to extend the continuum so far that the vowel would cross a phonemic boundary in order to find results resembling categorical perception. Thus, listeners can perceive a difference between words with final underlying /t/ and /d/, but not in the normal sense of perceiving a linguistic distinction. They perceive this difference linearly rather than categorically. This is a novel, and more precise, demonstration of what listeners' significantly better than chance but still rather low accuracy in Experiment 2 shows: listeners can hear a difference between final /t/ and /d/, but not as a categorical distinction.

As for the interaction of the factors other than Continuum Step, if there were a consistent effect of original voicing across speakers and vowel lengths, that is, if listeners in general responded more often with 't' to words intended to have final /t/, this would suggest that there were important perceptual cues other than vowel duration. Since the effect of Underlying Voicing and even the interaction of Underlying Voicing and Vowel Length are instead inconsistent across speakers, it is likely that the three-way interaction of Underlying Voicing, Vowel Length, and Speaker reflects idiosyncratic characteristics of the words on which the continua were based. Such idiosyncratic effects can occur because only one continuum (one word) was used for each combination of Underlying Voicing, Vowel Length, and Speaker.

5. Experiment 4: perception when closure duration is manipulated

Experiment 3 shows that listeners' judgments of underlying final voicing are affected by vowel duration. As shown in Experiment 1, average differences in vowel duration between underlying final /t/ and /d/ in real speech (even extremely careful real speech as in Experiment 1) are significant but very small, and are often variable. Fig. 1 shows that although some productions have vowel duration much longer before /d/ than /t/ (as much as 56 ms longer), nearly as many productions have differences nearly as large in the opposite direction (vowel duration as much as 50 ms longer before /t/ than /d/). When listeners are asked to do their best to perceive a rather small distinction (Experiment 3), the task itself might induce listeners to use whatever varies in the experimental stimuli as a perceptual cue. Since only vowel duration varied in Experiment 3, that would then appear to be a cue. Listeners adapt the vowel duration cue from the (intervocalic) distinction environment, where there is a larger durational difference in natural speech, to the neutralization environment where the durational differences show up as small differences in vowel duration. In the production data of Experiment 1, however, there is no significant effect of underlying voicing on closure duration in the neutralization environment (Table 2). It appears that there are no meaningful patterns in closure duration dependent on final voicing. The question remains then whether listeners could use closure duration as a cue to underlying voicing in an experimental setting in which only this cue is manipulated. To test this possibility, we performed a

continuum experiment similar to Experiment 3, but with closure duration instead of vowel duration varying.

5.1. Methods

The methods for this experiment were very similar to those for Experiment 3. As far as possible, the same words were used as in Experiment 3, but some had to be replaced because they did not have long enough closure durations to allow for splicing. For each word, a 10 ms section near the middle of the stop closure was selected, and this 10 ms portion of the signal was spliced into the closure an additional 1–5 times to make the longer steps of the continuum. A portion of the signal with no clicks or noises louder than the general background noise was selected. To make the shorter continuum steps, first the same 10 ms portion of the signal, then 20, 30, 40, and 50 ms surrounding it, were spliced out. In no case did these deleted areas extend beyond the silence of the stop closure. This resulted in an 11-step continuum with the endpoints differing in closure duration by 100 ms, again with the original production in the center of the continuum.

As in Experiment 3, all stimuli sounded like natural productions of the word from which they were made. 100 ms is a relatively large range for closure duration to vary by, but a pilot study showed that differences in a 50 ms continuum are only marginally audible. 100 ms is within the natural range of variation for closure duration: approximately 98% of productions of final stops in Experiment 1 have closure durations between 25 and 125 ms. The range of closure durations for the particular words used in this experiment is somewhat smaller (average range of 63 ms from 5th percentile to 95th percentile of productions), so the range used in the continua is somewhat greater than would be expected for any single word pair. Procedures were otherwise identical to those in Experiment 3. The 16 subjects were native Dutch speakers who did not participate in any of the other studies in this article.

5.2. Results

As in Experiment 3, percentage of 't' responses across the five presentations of each stimulus was calculated, and was analyzed using a 4-factor repeated measures ANOVA, with the factors Continuum Step (11 levels), original Underlying Voicing (/t, d/), Phonemic Vowel Length (long, short), and Speaker (4 levels). The main effects of Continuum Step ($F(10,150) = 4.84$, $p < 0.001$), Vowel Length ($F(1,15) = 8.08$, $p < 0.02$), and Underlying Voicing ($F(1,15) = 9.40$, $p < 0.01$) were significant. The interaction of Vowel Length by Continuum Step was also significant ($F(10,150) = 6.51$, $p < 0.001$), as was the interaction of Speaker by Underlying Voicing by Vowel Length ($F(3,45) = 3.26$, $p < 0.04$). The interactions of Speaker by Vowel Length ($F(3,45) = 4.42$, $p < 0.01$) and Speaker by Underlying Voicing ($F(3,45) = 7.03$, $p < 0.005$) were also significant. No other main effects or interactions were significant.

Pursuant to the significant Continuum Step by Vowel Length interaction, we tested the effect of Continuum Step for short and long vowels separately, collapsed across Underlying Voicing and Speaker (Fig. 3). For the long vowels, there was a significant effect of Continuum Step, with longer closure durations receiving more 't' judgments ($F(10,150) = 8.99$, $p < 0.001$). The linear trend component of the Continuum Step effect was also significant ($F(1,15) = 11.26$, $p < 0.005$), and no higher order trend components were significant. For the short vowels, although the effect
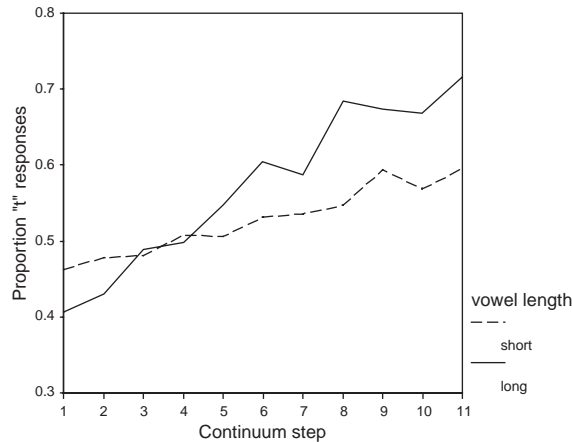
Fig. 3. Proportion of 't' responses, by continuum step in the closure duration continuum and phonemic vowel length. Step 1 has the shortest closure duration and step 11 the longest, with the original production at step 6.

was in the same direction and appears to be rather linear (Fig. 3), neither the effect of Continuum Step ($F(10,150) = 1.47$, $p > 0.1$) nor the linear trend component of the effect ($F(1,15) = 1.84$, $p > 0.1$), nor any higher order trend components, were significant. Thus, the interaction stems from the tendency for longer closure durations to receive more 't' judgments being stronger for continua containing long vowels.

As in Experiment 3, we investigated the interaction of Underlying Voicing, Vowel Length, and Speaker by testing Underlying Voicing and Vowel Length for each speaker separately. Speakers 3 and 14 showed a significant effect only of Underlying Voicing (3: $F(1,15) = 14.76$, $p < 0.003$; 14: $F(1,15) = 6.76$, $p < 0.03$), both with words intended to have final /t/ receiving more 't' responses. Speakers 5 and 6 had significant effects only of Vowel Length (5: $F(1,15) = 10.88$, $p < 0.01$; 6: $F(1,15) = 5.41$, $p < 0.04$), both with more 't' responses to phonemically long vowels. Thus, as in the previous experiment, there is no consistent pattern to these effects across speakers, and the effects probably stem from the choice of items for continua.

5.3. Discussion

These results show that, at least within the context of a forced choice experiment, listeners can use closure duration as a cue to the underlying final voicing of a stop, even though closure duration does not vary systematically with final voicing in natural productions. This suggests that when only one acoustic characteristic varies within an experiment, listeners may use it as a cue to the distinction they are being asked to make, even if it does not serve as a cue in natural speech. Although the effect of closure duration is significant only for the long vowels, the size of the effect for both long and short vowels is actually rather similar to the size of the effects for the vowel duration continuum (Fig. 2).

If listeners chose randomly how to use the only variability in the stimuli as a cue, some would respond to short closure durations more often with 't,' and others with 'd.' However, since the direction of the effect is consistent across subjects, this cannot be the case. Although

closure duration does not vary systematically between final /t/ and /d/ in natural productions in the neutralization environment, it does vary systematically in the distinction environment, with closure duration significantly longer for /t/ than /d/ (Table 2). Thus, it appears that listeners locate the only possible perceptual cue in the experiment, and apply it based on its relevance in another environment, namely the intervocalic rather than the word-final environment. That is, when necessary, listeners can transfer perceptual cues learned in one environment to another.

6. Experiment 5: production of a primarily orthographic distinction

Past research on durational differences in cases of neutralization has sometimes suggested that speakers might produce the differences because of the orthographic difference between words, or that the differences might be limited to very careful speech (Fourakis & Iverson, 1984; Jassem & Richter, 1989). However, since highly literate adult speakers have considerable practice in connecting the orthographic and phonological forms of words, when repeating words in isolation to dictation, for example, speakers might be influenced by the spelling of a word even though they are not reading it. Seidenberg and Tanenhaus (1979) document orthographic influence during auditory recognition of words quite clearly, for example. Speakers may be less influenced by orthography in less careful speech, but it is very difficult to collect many test minimal pairs in less careful speech while still maintaining sufficient control over prosody. Any noise introduced into the data by prosodic variation might obscure durational effects of the small magnitude reported in Experiment 1.

In this study, we take the opposite approach. Rather than attempt to elicit pairs which differ in underlying form without drawing speaker's attention to the orthographic difference, we investigate a case in which the primary difference between members of the pairs is orthographic, rather than a difference in the underlying representation. Thus, we can use read, careful speech, and any sub-phonemic durational differences observed will be indicative of an effect of orthography rather than underlying form. To do this, we used pairs of words which consist of the same string of phonemes, but have a systematic difference in spelling: the infinitive or plural present tense and the plural simple past tense of Dutch verbs with /t/ or /d/ as the final consonant of the stem. These are pairs such as *kleden* /kledən/ 'to dress/dress (pl.)' and *kleedden* /kledən/ 'dressed (pl.).' Here, the past tense form is spelled with an additional 'd' to mark the past tense, and according to the rules of Dutch orthography, the vowel must then also be doubled to show that it is phonemically long. Since Dutch orthography is highly regular, this is a regular correspondence among verbs with a long vowel in the first syllable and a stem ending in /d/ or /t/.

The pairs in this experiment might be considered not to have completely identical underlying representations, as the past tense morpheme is /-tə/ or /-də/, so *kleedden* could be underlyingly /kled-də-n/ rather than /kled-ən/. However, it could be that the past tense morpheme has a null allomorph after stems ending in /t, d/. In any case, since Dutch has no geminates, /kleddən/ would have to be a very abstract form, more abstract than the final /d/ in the forms such as *smeed* /smed/ of Experiment 1, which are simply a case of positional neutralization. Any underlying difference between *kleden* and *kleedden* is at a more abstract morphophonological level.

Table 4
Orthographic minimal pairs for Experiment 5

/t/ pairs			
baten	baatten	/batən/	'to avail/they availed'
blaten	blaatten	/blatən/	'to bleat/they bleated'
heten	heetten	/hetən/	'to be called/they were called'
keten	keetten	/ketən/	'to fool/they fooled'
kloten	klootten	/klotən/	'to screw around/they screwed around' (vulgar)
loten	lootten	/lotən/	'to draw lots/they drew lots'
poten	pootten	/potən/	'to plant/they planted'
stoten	stootten	/stotən/	'to jolt/they jolted'
zweten	zweetten	/zwetən/	'to sweat/they sweated'
/d/pairs			
baden	baadden	/badən/	'to bathe/they bathed'
besteden	besteedden	/bəstedən/	'to spend/they spent'
doden	doodden	/dodən/	'to kill/they killed'
kleden	kleedden	/kledən/	'to dress/they dressed'
kneden	kneedden	/knedən/	'to knead/they kneaded'
laden	laadden	/ladən/	'to load/they loaded'
loden	loodden	/lodən/	'to plumb/they plumbed'
noden	noodden	/nodən/	'to bid/they bid'
schaden	schaadden	/sxadən/	'to damage/they damaged'
smeden	smeedden	/smedən/	'to forge/they forged'
waden	waadden	/wadən/	'to wade/they waded'

6.1. Methods

We selected 20 orthographic minimal pairs consisting of a regular infinitive verb and its plural past tense form, with a long vowel as the stressed stem vowel and with /t/ or /d/ as the final consonant of the stem. 9 pairs had stem-final /t/, and 11 had /d/. A full list of items appears in Table 4. These pairs are all spelled with a single VC sequence in the infinitive (e.g., *kleden*, *poten*) and a doubled VVCC sequence in the past tense form (e.g., *kleedden*, *pootten*). Fifteen native Dutch speakers who did not participate in any other experiment in this study were recorded. Recording procedures were identical to those for Experiment 1. As in Experiment 1, speakers read from randomized word lists in which the presence of the pairs was fully apparent, although pairs did not appear together, and each speaker produced the full set of materials twice. (One production of one item by one speaker was excluded because of a pronunciation error.)

For each production, vowel duration and consonant duration were measured using the same criteria as in Experiment 1. Closure voicing was included as part of the consonant (both for /t/ and /d/). Only vowel and consonant duration were measured, because in this experiment the relevant difference is between orthographically single vs. double consonants and vowels, rather than between segments which differ in voicing.

Table 5
Average durations for all measurements of Experiment 5 (ms). Durations are shown by orthographic condition (medial orthographic VC or VVCC) and by voicing of the medial consonant (/t/ or /d/)

Measurement	Voiceless		Voiced	
	VC	VVCC	VC	VVCC
Consonant duration	109	113	47	50
Vowel duration	181	178	206	206

6.2. Results

Results were averaged across productions, and across items for the by subjects analysis and across subjects for the by items analysis. They were analyzed for effects of Orthographic Difference (VC vs. VVCC) and Voicing of the medial consonant on either consonant or vowel duration. (The Orthography factor is repeated measures in both analyses, and the Voicing factor is repeated measures in only the by subjects analysis.) Results appear in Table 5. VC items (e.g., *kleden*) had significantly shorter consonant duration than VVCC items (e.g., *kleedden*) ($F1(1,14) = 11.51$, $p < 0.005$, $F2(1,18) = 16.73$, $p < 0.005$). The main effect of Voicing on consonant duration was also significant ($F1(1,14) = 624.23$, $p < 0.001$, $F2(1,18) = 1312.69$, $p < 0.001$). The interaction was not significant either by subjects or items. The size of the orthographic effect, 3.4 ms, is remarkably similar to the size of the voicing effect on vowel duration in Experiment 1.

For vowel duration, there was a significant interaction of the Orthography and Voicing factors ($F1(1,14) = 4.73$, $p < 0.05$, $F2(1,18) = 6.78$, $p < 0.02$). Orthography significantly affected vowel duration for pairs with medial /t/ ($F1(1,14) = 7.95$, $p < 0.02$, $F2(1,18) = 11.69$, $p < 0.01$) but not pairs with medial /d/ ($F1$ and $F2 < 1$). Vowel duration was shorter for 'VVtt' words (178 ms) than for 'Vt' words (181 ms) by 3 ms, again about the same magnitude as the vowel duration effect in Experiment 1.

6.3. Discussion

Syllable structure is very important in the Dutch orthographic system. In Dutch bisyllabic words, a single vowel letter represents a phonemically long vowel if in an open syllable, but a short vowel if in a closed syllable. For example, 'a' represents the long vowel /a/ in *paden* /padən/ 'paths,' but the short vowel /ɑ/ in *padden* /pɑdən/ 'toads.' In order to represent a long vowel in a closed syllable, one must write an orthographically double vowel, leading to orthographic alternations between morphologically related forms, as in *noot* /not/ 'nut' vs. *noten* /notən/ 'nuts'. (See Nunn (1998) for discussion of this and other orthographic alternations in Dutch.) In Dutch this system is entirely regular aside from some proper names and recent borrowings. Given these orthographic conventions, Dutch allows phonologically identical items to exhibit systematic differences in spelling with the infinitive (VC) forms contrasting orthographically to the simple past tense (VVCC) forms. Our results show that this orthographic distinction influences duration of both consonant and vowel (only in one environment for the vowel) at a sub-phonemic level to about the same extent that underlying voicing influences duration in Experiment 1.

We suggest that in the current experiment, consonant duration is longer for VVCC than VC items because the consonant in VVCC items is *orthographically* both the coda of the first syllable and the onset of the second, but is only an onset in VC items. In both VVCC and VC items, the consonant is *phonologically* only an onset (or is ambisyllabic). The vowel in the VVCC items is shorter than in the VC items because it is in an orthographically closed syllable. While it has been shown that vowels in phonologically closed syllables are generally shorter than the same vowel in an open syllable (Maddieson, 1985; Jongman, 1998), these effects have not been shown for orthographic distinctions. The fact that the orthographic distinctions can cause such effects, even when the phonological syllables are identical, may stem from the importance of the open/closed syllable distinction in Dutch and its systematic reflection in the orthography. Perhaps in a language without such consistent orthographic patterns, such durational differences would not appear. It is possible that the morphophonological difference between the infinitive and past tense pairs (the addition of the past tense morpheme) may cause the durational effects. However, it is clear that abstract differences between forms, either orthographic or morphophonological, can cause the same sorts of durational differences which have been found for underlying phonemic differences.

7. General discussion

Our results in Experiment 1 suggest that durational differences of the type which have been called incomplete neutralization (e.g., effects of underlying final voicing in a language with final devoicing) are pervasive, if often extremely small. We find such differences even in Dutch, a language for which some past research has found complete neutralization (Jongman et al., 1992; Baumann, 1995). Although we did find these differences (specifically an effect on vowel duration, and an effect on burst duration in long vowels only), the effects we found are much smaller than those found in previous studies of other languages, 3.5 ms for vowel duration in the present study compared to the 10–15 ms found for German or Polish. This suggests both that a large number of subjects and items are necessary in order to be sure of such effects, and that these effects may be even more widespread than previously recognized, if sufficiently large and careful studies are used to search for them.

In the past literature on German, those studies which do find significant incomplete neutralization effects are often inconsistent in respect of which duration is found to be affected. The same is true for Dutch when comparing our study to others: The only incomplete neutralization effect Baumann (1995) finds is on closure duration (and that for only one syntactic condition), while Ernestus and Baayen (in press) find an effect only on burst duration. The only effect we find for both long and short vowels is on vowel duration. This inconsistency in what is affected also suggests that incomplete neutralization effects are small, variable, and task dependent.

Experiment 2 shows that listeners can also use durational differences to distinguish words with differing underlying forms. They can use whatever acoustic differences speakers produce to achieve significantly better than chance, although still quite poor, recognition of supposedly homophonous final voicing pairs. This indicates that listeners' ability to distinguish these pairs is highly dependent on the pattern of differences individual speakers produce. By systematically

manipulating only one cue, Experiment 3 confirms that vowel duration, the most likely cue based on the production results, is perceptually useful. However, the results of Experiment 4 suggest that additional cues can be perceptually useful. When closure duration, rather than vowel duration, is the only acoustic feature to vary in the experiment, listeners appear to use that as a perceptual cue to underlying final voicing, even though closure duration does not vary with underlying final voicing in the production results. Closure duration does vary with voicing of medial (non-neutralized) stops, so it seems that when listeners are asked to make a difficult distinction, and only one acoustic feature of the stimuli varies, they can 'borrow' a perceptual cue from another environment in order to make a distinction in a neutralized environment.

While past research suggests that the largest durational effects are shown for languages that maintain an orthographic contrast between the two underlying forms, our results in Experiment 5 indicate that durational differences of this magnitude are not exclusively caused by underlying differences in the phoneme string. Rather, such differences can also stem from orthographic (or perhaps morphophonological) differences that do not represent underlying phonological differences. Ernestus and Baayen's (in press) finding of sub-phonemic durational differences in non-words spelled with final 'p, t' vs. 'b, d' also suggests an influence of orthography rather than underlying representation, unless speakers construct underlying representations for the non-words based on orthography. In addition, Whalen (1991, 1992) reports that higher frequency words are produced faster and thus with shorter segment durations than lower frequency words, indicating that frequency of occurrence too can be a source of durational differences. These findings, in combination with the present results, suggest that such effects should be referred to as sub-phonemic durational differences, rather than incomplete neutralization effects. Wright (2002) does not examine duration, but finds small spectral differences between words with 'easy' and 'hard' lexical neighborhoods. This suggests that sub-phonemic differences may not be limited to duration, and adds another possible source of such effects, lexical neighborhood difficulty. Small durational or spectral differences, far smaller than those between segments that are linguistically distinct, can be caused by differences in orthography, frequency, or even neighborhood density. Thus, speakers may produce acoustic differences much smaller than those between phonemes for a variety of reasons, not just because of underlying form differences.

We now return to the meaning of sub-phonemic durational differences for linguistic theory. Incomplete neutralization effects are often taken to mean that underlying phonemic differences are not completely neutralized, but rather leave behind small differences in the same direction as one finds where the distinction is fully maintained. If there are reliable differences between supposedly neutralized segments which are smaller than the differences which might ever cue a linguistic contrast, and the differences are too small to be useful in perceiving natural speech, then the entire concept of contrast is called into question (see Port (1996) for discussion). However, it appears from our results and those of other researchers that one can find such small reliable differences below the level of contrast for a variety of reasons, only one of which is differences in underlying representations. We have documented sub-phonemic durational differences for pairs which differ in orthographic form, while Whalen finds them for pairs which differ only in word frequency (Whalen, 1991, 1992), and Wright (2002) finds that lexical neighborhood difficulty yields similar spectral differences.

It has been well known in phonetics for decades that the environment of a segment (surrounding segments, prosodic environment, etc.) conditions small non-contrastive differences

in the duration and spectral characteristics of that segment. Rather than viewing incomplete neutralization as a phenomenon that calls the concept of linguistic contrast into question, we should allow for the fact that speakers produce small acoustic differences, below the level of phonemic contrast, for a variety of reasons other than segmental and prosodic environment. Differences in underlying form, orthography, morphology, word frequency, and lexical neighborhood difficulty might all cause such effects. Any of these factors can bring about sub-phonemic differences. A full model of how speakers produce speech, and a full linguistic model of phonetics, must include the possibility of several influences, other than the underlying phonemic string, on the exact specification of the acoustic signal.

Acknowledgements

References

Baayen, H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Baumann, M. (1995). *The production of syllables in connected speech*. Unpublished Ph.D. dissertation, University of Nijmegen.

Charles-Luce, J. (1985). Word-final devoicing in German: Effects of phonetic and sentential contexts. *Journal of Phonetics*, *13*, 309–324.

Charles-Luce, J. (1993). The effects of semantic context on voicing neutralization. *Phonetica*, *50*, 28–43.

Dinnsen, D. (1985). A re-examination of phonological neutralization. *Journal of Linguistics*, *21*, 265–279.

Dinnsen, D., & Charles-Luce, J. (1984). Phonological neutralization, phonetic implementation and individual differences. *Journal of Phonetics*, *12*, 49–60.

Ernestus, M., & Baayen, H. (in press). The functionality of incomplete neutralization in Dutch: The case of past tense formation. *Papers in Laboratory Phonology VIII*.

Fourakis, M., & Iverson, G. (1984). On the incomplete neutralization of German final obstruents. *Phonetica*, *41*, 140–149.

Jassem, W., & Richter, L. (1989). Neutralization of voicing in Polish obstruents. *Journal of Phonetics*, *17*, 317–325.

Jongman, A. (1998). Effects of vowel length and syllable structure on segment duration in Dutch. *Journal of Phonetics*, *26*, 207–222.

Jongman, A., Sereno, J., Raaijmakers, M., & Lahiri, A. (1992). The phonological representation of [voice] in speech perception. *Language and Speech*, *35*, 137–152.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd Ed). Englewood Cliffs, NJ: Prentice-Hall.

Kopkalli, H. (1993). *A phonetic and phonological analysis of final devoicing in Turkish*. Unpublished Ph.D. dissertation, University of Michigan.

Kim, H., & Jongman, A. (1996). Acoustic and perceptual evidence for complete neutralization of manner of articulation in Korean. *Journal of Phonetics*, *24*, 295–312.

Lahiri, A., Schriefers, H., & Kuijpers, C. (1987). Contextual neutralization of vowel length: Evidence from Dutch. *Phonetica*, *44*, 91–102.

Maddieson, I. (1985). Phonetic cues to syllabification. In V. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged* (pp. 203–221). Orlando: Academic Press.

Nunn, A. M. (1998). *Dutch orthography: A systematic investigation of the spelling of Dutch words*. Ph.D. dissertation, Holland Academic Graphics, The Hague.

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

Port, R., & Crawford, P. (1989). Incomplete neutralization and pragmatics in German. *Journal of Phonetics*, *17*, 257–282.

Port, R., & O'Dell, M. (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics*, *13*, 455–471.

Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 546–554.

Slowiaczek, L., & Dinnsen, D. (1985). On the neutralizing status of Polish word-final devoicing. *Journal of Phonetics*, *13*, 325–341.

Whalen, D. H. (1991). Infrequent words are longer in duration than frequent words. *Journal of the Acoustical Society of America*, *90*(4), 2311.

Whalen, D. H. (1992). Further results on the duration of infrequent and frequent words. *Journal of the Acoustical Society of America*, *91*(4), 2339–2340.

Wittenburg, P., Nagengast, J., & Baumann, H. (1998) NESU—the Nijmegen experiment setup. In: A. Trapp, N. Hammond, C. Manning (Eds.), *CIP98 conference proceedings* (pp. 92–93). York: CTI.

Wright, R. (2002). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology VI*. Cambridge University Press.