Journal of Phonetics 43 (2014) 11–25

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Assessing incomplete neutralization of final devoicing in German

T.B. Roettger [a,*], B. Winter [b], S. Grawunder [c], J. Kirby [d], M. Grice [a]

[a] IfL Phonetik, University of Cologne, Herbert-Levin-Str. 6, D-50931 Köln, Germany
[b] Department of Cognitive and Information Sciences, University of California, Merced, 5200 North Lake Rd., Merced, CA 95343, USA
[c] Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany
[d] School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh EH8 9AD, Scotland, UK

ARTICLE INFO

Article history:
Received 12 June 2013
Received in revised form
3 January 2014
Accepted 7 January 2014

ABSTRACT

It has been claimed that the long established neutralization of the voicing distinction in domain final position in German is phonetically incomplete. However, many studies that have advanced this claim have subsequently been criticized on methodological grounds, calling incomplete neutralization into question. In three production experiments and one perception experiment we address these methodological criticisms.

In the first production study, we address the role of orthography. In a large scale auditory task using pseudowords, we confirm that neutralization is indeed incomplete and suggest that previous null results may simply be due to lack of statistical power. In two follow-up production studies (Experiments 2 and 3), we rule out a potential confound of Experiment 1, namely that the effect might be due to accommodation to the presented auditory stimuli, by manipulating the duration of the preceding vowel. While the between-items design (Experiment 2) replicated the findings of Experiment 1, the between-subjects version (Experiment 3) failed to find a statistically significant incomplete neutralization effect, although we found numerical tendencies in the expected direction. Finally, in a perception study (Experiment 4), we demonstrate that the subphonemic differences between final voiceless and "devoiced" stops are audible, but only barely so. Even though the present findings provide evidence for the robustness of incomplete neutralization in German, the small effect sizes highlight the challenges of investigating this phenomenon. We argue that without necessarily postulating functional relevance, incomplete neutralization can be accounted for by recent models of lexical organization.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Many languages such as Catalan, Dutch, German, Polish, Russian and Turkish contrast voiced obstruents intervocalically but neutralize the contrast syllable or word finally in favor of voiceless obstruents. An example from German is given in (1) and (2): in syllable final position, the voicing of the alveolar stop is neutralized, leading to apparent homophony between e.g. *Rad* [ʁaːt] 'wheel' and *Rat* [ʁaːt] 'council'.

(i) Rad [ʁaːt] 'wheel'; Räder [ʁæːdɐ] 'wheels'

(ii) Rat [ʁaːt] 'council'; Räte [ʁæːtə] 'councils' (1)

(i) Radschlag [ʁaːtʃlaːk] 'cartwheel'

(ii) Ratschlag [ʁaːtʃlaːk] 'advice' (2)

This asymmetrical distribution is commonly described in terms of final devoicing, a process that is often described in purely phonological terms. In fact, final devoicing in German[1] has been called the "universally recognized archetype of phonological neutralization" (Fourakis & Iverson, 1984: 141) and described as a "classic example of a phonological rule" (Wiese, 1996: 204).

In traditional formal theories of phonology, *Rad* and *Rat* are thought to differ only in their "underlying" lexical representations, while the surface form of the voiced stop is thought to be phonetically indistinguishable from that of the corresponding voiceless stop. In other words, neutralization of the final voicing distinction is assumed to be phonetically complete, resulting in homophony between the two lexical items. However, numerous experimental studies have

* Corresponding author. Tel.: +49 221 4707047; fax: +49 221 470 5938.
E-mail address: timo.roettger@uni-koeln.de (T.B. Roettger).

[1] Kohler (1984) argues that German voiced and voiceless stops are better characterized as fortis and lenis. To remain consistent with the terminology adopted in the incomplete neutralization debate, we retain the terms "voiced", "voiceless" and "final devoicing".

0095-4470/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2014.01.002

argued that there are small acoustic and articulatory differences between words such as *Rad* and *Rat*, suggesting that in German this neutralization is in fact *incomplete* (Charles-Luce, 1985; Dinnsen, 1985; Dinnsen & Garcia-Zamor, 1971; Fuchs, 2005; Greisbach, 2001; Mitleb, 1981; O'Dell & Port, 1983; Port & Crawford, 1989; Port, Mitleb, & O'Dell, 1981; Port & O'Dell, 1985; Piroth & Janker, 2004). Further studies suggest that listeners can distinguish "devoiced"[2] stops from voiceless ones with above-chance accuracy (Kleber, John, & Harrington, 2010; Port & Crawford, 1989; Port & O'Dell, 1985).

The results obtained in the above mentioned experiments are difficult to reconcile with traditional linguistic descriptions of German (Jespersen, 1913; Trubetzkoy, 1939; Wiese, 1996; Zifonun et al., 1997) that assume abstract phonological categories devoid of gradient phonetic information. Accounts based on this view have problems incorporating intermediate categories as the purported "semi-voiced" final obstruents. Most early formal attempts to incorporate incomplete neutralization (e.g., Charles-Luce, 1985; Port & O'Dell, 1985) involved a proliferation of post-hoc repairs (such as the "phonetic implementation rules" of e.g., Dinnsen & Charles-Luce, 1984) which led Port & Crawford (1989: 257) to claim that incomplete neutralization poses "a threat to phonological theory" (see also Port & Leary, 2005).

More recent attempts to account for incomplete neutralization are rooted in psycholinguistic models of lexical organization. There is mounting evidence suggesting that, far from being impoverished, lexical representations are rich in information, and may contain both detailed phonetic information of individual word forms (e.g., Brown & McNeill, 1966; Bybee, 1994; Goldinger, 1996, 1997; Palmeri, Goldinger, & Pisoni, 1993; Pisoni, 1997) as well as completely inflected forms (e.g., Alegre & Gordon, 1999; Baayen, Dijkstra, & Schreuder, 1997; Butterworth, 1983; Bybee, 1995; Manelis & Tharp, 1977; Sereno & Jongman, 1997). Such models of lexical organization and access assume that German speakers have inflected forms such as *Räder* in their mental lexicon. Due to its phonological and semantic relations with the singular form *Rad*, these two forms will be closely connected to each other. Ernestus and Baayen (2006) consider the possibility of incomplete neutralization effects being due to the co-activation of these related forms, i.e., when speakers pronounce *Rad*, they also activate the non-neutralized *Räder*. If some or most of the co-activated forms contain a non-neutralized segment that is fully voiced, these voiced forms could influence the motor commands used in speech production in subtle ways, leading to the observed incomplete neutralization effects.

A similar account has been advanced to explain the finding that speakers are able to distinguish forms like *Rat* and *Rad* with above-chance accuracy. Kleber et al. (2010) found that there is a greater probability of identifying a stop as voiceless after lax than after tense vowels. They further found that, following tense vowels, the (putatively neutralized) stop voicing contrast in syllable final position was recoverable more often when the stop was alveolar than when it was velar. Since in German phonologically short/lax vowels tend to occur more often before bilabial and velar voiceless stops, this suggests that sensitivity to statistical patterns of the German lexicon may affect the perception of incomplete neutralization, and thus it seems plausible that knowledge of phonotactic probabilities might play a role in production as well.

It seems safe to say that the predominant response to incomplete neutralization studies has been one of skepticism. Given that several early studies found no evidence for incomplete neutralization (Fourakis & Iverson, 1984; Jassem & Richter, 1989), some researchers have considered the debate to be settled (e.g., Kohler, 2007, 2012). However, other researchers have continued to investigate the phenomenon (e.g., Kleber et al., 2010; Piroth & Janker, 2004), and studies have since been carried out on both incomplete neutralization of final devoicing in other languages (e.g., in Dutch (e.g., Warner, Jongman, Sereno, & Kemps, 2004), Catalan (e.g., Charles-Luce & Dinnsen, 1987), Polish (e.g., Slowiaczek & Dinnsen, 1985) and Russian (e.g., Dmitrieva, Jongman, & Sereno, 2010; Kharlamov, 2012)) as well as incomplete neutralization of other processes (Bishop, 2007; Braver & Kawahara, 2012; de Jong, 2011; Dinnsen, 1985; Gerfen, 2002; Gerfen & Hall, 2001; Simonet, Rohena-Madrazo & Paz, 2008).

Thus, the debate surrounding incomplete neutralization is still very much ongoing. However, the numerically small effect sizes common across incomplete neutralization studies have attracted serious criticism on methodological grounds (Kohler, 2007; Manaster-Ramer, 1996). Fuchs (2005: 25) points out that the debate surrounding incomplete neutralization has become increasingly a debate about methodology rather than the phenomenon per se. As such, our first and foremost aim in the present work is to address the methodological and conceptual concerns raised against previous studies, thereby placing the debate surrounding incomplete neutralization on firmer empirical footing. Our second aim is to interpret our findings in light of recent psycholinguistic models of lexical organization.

In Section 2, we summarize previous empirical findings as well as their critiques, with a particular focus on Fourakis and Iverson (1984) and Jassem and Richter (1989). In Sections 3–5 we discuss the results of three production experiments that were inspired by Fourakis and Iverson's study. Section 6 presents the results of a perception experiment. In Section 7, we discuss the implications of our work for an assessment of the status of incomplete neutralization in German in light of co-activation accounts.

2. Methodological debate and the problem of "proving the null"

Across different studies, numerous phonetic properties have been found to distinguish voiceless from devoiced stops in final position. These include the duration of the preceding vowel, the closure duration, the duration of the "voicing-into-the-closure", as well as the burst and aspiration durations. Across different studies and languages, the duration of the preceding vowel has been shown to be the most reliable correlate of obstruent "voicing" in final position. Thus in the present study we shall focus on this acoustic parameter. This has the added advantage that we avoid statistical issues surrounding multiple comparisons: with each additional measure taken into account we have an added probability of rejecting the global null hypothesis that there is no acoustic correlate of incomplete neutralization at all. Standard ways of correcting for multiple comparisons, such as Bonferroni correction, increase the probability of missing a true effect and according to Bender and Lange (2001: 347) the "easiest and best interpretable approach is to avoid multiplicity as far as possible". We do this by focusing on vowel duration.

The direction of the vowel duration difference mirrors the durational difference in the intervocalic context, i.e., vowels tend to be longer before final devoiced stops than before final voiceless stops. Numerically, incomplete neutralization effects of vowel duration are minute. For example, Port and Crawford (1989) report a difference of 1.2–6.2 ms between devoiced and voiceless stops in German, while Warner et al. (2004) report a difference of 3.5 ms in Dutch. The magnitude of the incomplete neutralization effect appears to be dialect- and speaker-dependent (Piroth & Janker, 2004), as well as highly sensitive to the phonetic, semantic and pragmatic context (Charles-Luce, 1985, 1993; Ernestus & Baayen, 2006; Port & Crawford, 1989; Slowiaczek & Dinnsen, 1985).

As German maintains an orthographic contrast between voiced/devoiced and voiceless stops in all positions, the biggest issue surrounding previous results was the influence of this orthographic representation.[3] Most of the above-mentioned experiments used stimuli that had to be read aloud by the participants,

[2] We refer to the segment in words such as *Rad* as "devoiced". This term is theoretically loaded because it assumes the presence of an underlying voiced segment. However, for this paper, we merely use the term as shorthand to refer to a segment corresponding to an intervocalic voiced segment within the same morphological paradigm, e.g., *Räder* [d] vs. *Rad* [t], without necessarily invoking a phonological process of devoicing.

[3] There are other concerns with incomplete neutralization studies. These include minimal pair awareness, second language proficiency of experimenter and participants and stimuli selection. These concerns have been dealt with at length in Fourakis and Iverson (1984), Manaster-Ramer (1996), Kohler (2007) and Winter and Roettger (2011).

inviting the criticism that participants used a form of hypercorrection or spelling pronunciation: as laboratory settings tend to elicit more formal and clear speech, participants might have produced words based on the written language in a way that they would not do in everyday speech.

In Fourakis and Iverson (1984) (henceforth FI), four native speakers were asked to conjugate neutralized verb forms such as *mied* ('avoid.PST.1+3SG') when presented auditorily with non-neutralized forms such as *meiden* ('to avoid'). Both the duration of the preceding vowel and the closure duration were measured. No statistically significant incomplete neutralization effect was obtained. Jassem and Richter (1989) (henceforth JR) conducted a very similar study in Polish in which participants answered questions constructed by the experimenter such that the answer could be expected to consist of a single word utterance. They measured the duration of the preceding vowel, voicing-into-the-closure/frication, closure/frication duration, and, where relevant, release duration. Again, four speakers were recorded and no incomplete neutralization effect was found.

In both cases, it was concluded that the lack of a statistically significant effect supports an orthography-based explanation of incomplete neutralization. Since then, many have cited FI and JR as evidence against incomplete neutralization (e.g., Kohler, 2007, 2012; Wiese, 1996). However, these studies have methodological shortcomings of their own. For example, FI did not use minimal pairs, but instead compared words such as *mied* and *riet* ('avoid.PST.1+3SG' and 'advice.PST.1+3SG'). As pointed out by Dinnsen and Charles-Luce (1984) and Port and Crawford (1989), this leaves the potential influence of the syllable onset uncontrolled for. In other words, the durational differences due to final voicing are confounded with durational differences due to properties of the initial consonant.

Both FI and JR interpret their null results as evidence for the absence of incomplete neutralization. There is a logical problem with "accepting the null", and most researchers would argue that it is not logically sound to accept null hypotheses (e.g., Cohen, 1990; Weitzman, 1984), in line with the saying that "absence of evidence is not evidence for absence". If anything, one can only demonstrate "sufficiently good effort" to disprove the null hypothesis (Frick, 1995). FI and JR only tested four speakers – less than most of the previous and following investigations of incomplete neutralization that *did* find an effect. Their null results may thus well be due to a lack of statistical power.

Another concern related to statistical power is that FI conducted statistical tests within speakers. Thus, for each individual test there were only a few data points. Indeed, an across-speaker re-analysis of the published FI data conducted by Port and Crawford (1989) did find significant differences consistent with incomplete neutralization. Given the low statistical power (due to lack of minimal pairs, a small number of speakers and the fact that subset analyses were conducted), it is possible that both studies committed a Type II error (i.e., failing to reject a false null hypothesis). This would not be the first time this has happened with respect to incomplete neutralization. For Dutch final devoicing, Baumann (1995) and Jongman, Sereno, Raaijmakers, and Lahiri (1992) failed to find significant incomplete neutralization effects, but Warner et al. (2004), with more speakers, did find effects.

At a bare minimum, any study that wants to demonstrate "sufficiently good effort" to disprove the null needs to have at least as many subjects and items as previous investigations *in support of* the purported phenomenon. Thus while the studies by FI and JR certainly suggest that effect sizes for incomplete neutralization are small, their results cannot be taken as counter-evidence against the phenomenon.

With regard to the perceptibility of incomplete neutralization, previous studies investigated accuracy in forced choice identification tasks. The identification accuracies were reported to be generally lower than in experiments with non-neutralized contrasts (see Brockhaus, 1995: 244, for an overview) and, in some studies, even barely above chance performance (Port & O'Dell, 1985). This leads to the question as to whether incomplete neutralization has any function in speech communication.

Previous studies used auditory stimuli for the perception experiments which come from a small set of speakers (e.g., Port & Crawford, 1989), or in some cases from just a single speaker (e.g., Kleber et al., 2010). This, together with many repetitions, gives participants ample opportunity to familiarize themselves with speaker characteristics. This in turn might make it easier for participants to detect subtle cues to voicing in a neutralizing context, enhancing the likelihood that they might be attending to cues that they would not use in listening situations outside of the laboratory.

Thus, although there is some evidence that listeners are able to exploit subtle cues to distinguish devoiced from voiceless stops in final position, the results must be interpreted with caution. While some see this as genuine evidence for incomplete neutralization as a perceptual phenomenon with potential real-world relevance, others are more inclined to view it as the result of task demands (e.g., Slowiaczek & Szymanska, 1989; Warner et al., 2004). Brockhaus (1995: 244), among many others, points out that it is not clear whether the perceptual difference between syllable-final devoiced and voiceless obstruents is actually "salient enough to be relied upon in normal communication". Although it is not known how accurate a contrast needs to be perceived in order to play a role outside the laboratory (Xu, 2010: 334), the low accuracy scores and high variability suggest that incomplete neutralization would have little if any functional relevance in everyday communicative situations.

In summary, a number of methodological shortcomings have been identified in previous studies arguing for the existence of incomplete neutralization. However, studies that failed to find incomplete neutralization effects are, themselves, subject to methodological criticism, especially since *failure* to find an effect cannot be taken as evidence for the *absence* of that effect. The present study aims to circumvent these concerns.

Our production studies are inspired by Fourakis and Iverson's (1984) study of German final devoicing, but employ a design that has increased statistical power (more speakers, more items). We also address the concern that incomplete neutralization is potentially a result of an orthographically induced contrast. It is known that speakers automatically activate orthographic representations even in completely auditory tasks (Dehaene et al., 2010; Perre, Midgley, & Ziegler, 2009; Seidenberg & Tanenhaus, 1979; Ziegler & Ferrand, 1998). Given that all previous studies on incomplete neutralization used real word stimuli, literate speakers inevitably know their written forms. Thus in our first experiment, we employed pseudowords, such as *Gobe* or *Gope,* in order to reduce the effect of orthography. Subjects were presented with a plural form auditorily in which the target consonant is intervocalic ([goːbə]), and were instructed to produce the singular form ([goːp]) in which the target consonant is word final. Pseudowords, which effectively have a frequency of 0, presumably lack existing orthographic representations. While it is still possible that participants think of our auditorily presented pseudowords in terms of orthography (for example, they might think of how they would spell a given pseudoword in order to produce its related singular form), the design minimizes the role of orthography *relative to other studies on incomplete neutralization*, in particular relative to FI. To the extent that orthography impacts the realization of incomplete neutralization, this should make the effect less likely to emerge.

This design, however, potentially introduces another confound: accommodation to the auditory stimuli. Phonetic accommodation, also known as phonetic convergence or phonetic imitation, involves the adaptation of a talker's speech to that of his or her interlocutor (e.g., Goldinger, 1998; Gregory & Hoyt, 1982; Natale, 1975a,b). This process happens even in situations with minimal social interaction: a number of laboratory studies have found that participants shift their pronunciation of single words towards productions of auditorily presented voices they have just heard (e.g., Babel, 2012; Goldinger 1996, 1997, 1998; Nielsen, 2011). Thus in a task in which participants are exposed to the intervocalic contrast auditorily (they hear e.g. [goːbə] or [goːpʰə]) and respond with the corresponding singular form right away, they may merely imitate the acoustics of the stimulus they have just heard. To address this issue we conducted two additional experiments, eliminating this potential confound by manipulating the acoustic cues of the intervocalic voicing distinction.

Finally, given the small effect sizes reported in the literature, we sought to evaluate the functional relevance of incomplete neutralization for speech communication. To assess the perceptibility of incomplete neutralization in a more ecologically valid design, we replicated earlier perception studies utilizing a number of different voices. If speakers consistently fail to fully neutralize the voicing contrast in final position, and/or if listeners are able to distinguish between voiceless and devoiced stops with greater than chance accuracy, this suggests that neutralization is indeed incomplete. Even if it arguably has no functional utility for communication, an explanation of this effect is nonetheless warranted, given its implications for foundational theories of phonological processing and lexical organization.

3. Production Experiment 1

3.1. Methodology

3.1.1. Participants and experimental procedure

Sixteen native speakers of German participated in the experiment (mean age: 25 years; nine women). All were undergraduates or PhD students in the humanities living in Cologne or in the area surrounding Cologne. Most of them grew up in this area and all participants claimed to speak non-dialectal Standard German. None of the participants were familiar with the concept of incomplete neutralization prior to the post-experiment debriefing.

The recording session was managed by a native speaker (the first author) and conducted entirely in German. Participants were seated in a well-illuminated sound-treated booth in front of a computer screen. They were given written instructions that stated that the experiment investigates German plural formation. None of the participants reported noticing the presence of minimal pairs in the post-experimental interview. This addresses previous concerns surrounding the idea that incomplete neutralization effects might be artificially enhanced because of hyperarticulation due to participants noticing the final voicing alternation (see discussion in Winter & Roettger, 2011). After the written instructions, the remaining procedure was conducted auditorily. In each trial, participants first heard a stimulus sentence such as (3) and then produced a corresponding sentence such as (4).

Plural stimulus :

Aus Dortmund kamen die ***Drude***. (3)

From Dortmund come.3PL.PST DEFDET.PL.NOM NONCE-PL
From Dortmund came the NONCE-PL.

Singular response :

Ein ***Drud*** *wollte nicht mehr*. (4)

INDFDET.SG.M.NOM. NONCE-SG want.3SG.PST NEG longer
One NONCE-SG did not want to continue.

The experiment was run using Superlab 2.04 (Abboud, 1991). At the beginning of each trial, a cross appeared in the center of the screen (+) and participants heard the plural sentence through headphones. After presenting a blank screen for 500 ms three question marks appeared on the screen. Participants were now asked to produce the corresponding singular sentence. The experiment was self-paced and there were no time constraints.

Prior to the actual experiment, participants listened to eight demonstration stimuli, each of which was a plural sentence followed by a singular response. None of these demonstration items were potential critical items, and none included a voiced/voiceless obstruent distinction. This was done so as not to bias our participants' responses with respect to incomplete neutralization. After the demonstration, participants performed eight practice trials where they had to produce the response sentences themselves. The actual experiment was divided into four blocks. After each block, there was an obligatory break of at least ten seconds. On average, the entire experiment (including instruction and debriefing) took about 30 min.

3.1.2. Speech material

The experimental items consisted of 24 pseudoword pairs such as (5)–(7) (see Appendix A):

Gobe [goːbə] vs. Gope [goːpʰə] (5)

Frade [fraːdə] vs. Frate [fraːtʰə] (6)

Schuge [ʃuːgə] vs. Schuke [ʃuːkʰə] (7)

All pseudowords were trochaic and complied with German phonotactic rules. There were eight bilabial, seven alveolar and nine velar stimulus pairs, each containing one of the vowels /aː, oː, uː, iː, a͡ʊ/. Each experimental item was introduced as a masculine noun inflected for the plural. Plural inflection was indicated through the regular plural marker for masculine nouns (/-ə/), the plural determiner /diː/ and number agreement on the verb. The German plural system exhibits many irregularities, and we chose the particular plural form used in this study because it is the most likely plural of monosyllabic masculine nouns (e.g., *Arm/Arme* 'arm/arms', *Stift/Stifte* 'pen/pens', etc.). We did not choose the commonly occurring plural ending *–en* because speakers are more insecure as to which singular form corresponds to pseudowords ending in *–en* (as a pilot study demonstrated), and because this marker often involves schwa deletion and a nasal release, which might in turn lead to an additional lengthening of the preceding vowel.

As German plural formation is very complex, we needed to norm our stimuli with respect to their morphology. A list of the intended singular forms was given to a group of five participants who were asked to provide the respective plural forms. Indeed, the schwa-plural (/-ə/) was the most frequent response pattern (84% of all responses). However, as expected, some nonsense words were more consistently formed with this morpheme than others. The extent to which a stimulus was identified as schwa-plural was included in the statistical analyses presented below.

To further alleviate the concern of hyperarticulation due to minimal pair awareness, we included 96 fillers (2/3 of the total stimulus set), 70% of which contained an umlaut vowel. As plural forms with umlaut vowels sometimes do and sometimes do not require a vowel change (e.g., *Turm>Türme* 'tower/towers' but *Bär>Bären* 'bear/bears'), we hoped this would increase the salience of the fillers, simultaneously detracting attention from the critical stimuli. Forty different city names (randomized over stimulus pairs) were embedded in the carrier phrase to introduce an additional distracting element, but in all other respects the carrier phrase ('Aus (CITY NAME) kamen die (NONCE PLURAL)') remained constant. We avoided repetition of items to

further decrease the salience of the relevant minimal pairs. The 144 stimuli and the 16 demonstration and practice items were spoken by a native speaker of German (male, trained phonetician) and recorded in a sound-treated booth with an AKG C420 III microphone. All stimuli were randomized and divided into four blocks. Members of a stimulus pair were always within different blocks. At the beginning of the experiment, each participant was randomly assigned to one of eight block orders.

3.1.3. Acoustic analysis of stimuli

We performed acoustic analyses of the plural stimuli that were presented to participants to ensure that the stimuli have the typical acoustic characteristics of German voiced and voiceless stops (Keating, 1984; Kohler, 1984). Using Praat (Boersma & Weenink, 2011), we measured the duration of the vowel preceding the critical stop, the closure duration, the duration of the following vowel, the burst duration, the voice onset time and the median intensity of the burst. In addition, we analyzed the mean fundamental frequency (f_0), as well as the f_0 in the first quintile of the vowel following the stop release.

The vowel preceding the critical stop was on average 28 ms (SE = 3.7) longer before intervocalic voiced stops than before intervocalic voiceless stops ($\chi^2(1)=30.27$, $p<0.001$)[4]; there was no significant difference in the following vowel ($\chi^2(1)=0.04$, $p=0.99$). Voice onset times were on average 42 ms (SE = 2.3) longer for voiceless stops ($\chi^2(1)=65.57$, $p<0.0001$), and closures were on average 21 ms (SE = 1.56) longer for voiceless stops ($\chi^2(1)=52.8$, $p<0.0001$). There were no significant differences for burst duration ($\chi^2(1)=1.57$, $p=0.85$) or burst intensity ($\chi^2(1)=0.37$, $p=0.99$), nor were there differences of mean f_0 ($\chi^2(1)=2.17$, $p=0.7$) or of the first quintile f_0 in the second vowel ($\chi^2(1)=2.75$, $p=0.56$). Furthermore, all but one of the voiced stimuli had consistent voicing during the closure, meaning that vocal fold vibration was a reliable and consistent cue. Thus, the stimuli that were given to participants are relatively typical German voiced and voiceless stops. We found large differences in vowel durations before voiced and voiceless stops, the closure duration and the voice onset time in addition to voicing during the closure. This means that there are at least four robust cues for participants to distinguish between the voiced and the voiceless stimuli intervocalically. We presented these stimuli to five male and five female German participants who were able to retrieve the voicing status with 98% accuracy. Having described the phonetic details of the stimuli, we now turn to the measurements of the responses.

3.1.4. Acoustic analysis of responses

Participant responses were digitized at a sampling rate of 44.1 kHz (16 bit). The durations of the vowels preceding the final stops were measured by the first author. If the sound preceding the vowel/diphthong was a stop, the onset of the vowel was defined as the onset of voicing in cases of voiceless stop or as the end of the burst in cases of voiced stops. A sudden discontinuity in the spectrogram was taken as the onset of vowels following fricatives ([ʃ]), nasals ([m] and [n]), laterals ([l]) and palatal approximants ([j]) (e.g., [ʃuːk], [muːp], [bloːk] or [jiːt]). The end of the vowel was defined as the end of the second formant of the vowel, which usually coincided with a sudden drop in amplitude of voicing. To assess the interaction between incomplete neutralization and prosodic factors, we also coded certain aspects of the prosodic realization including the accent position and the presence of a potential prosodic boundary following the critical item.

3.1.5. Statistics

All data were analyzed with generalized linear mixed models, using R (R Core Team, 2012) and the package *lme4* (Bates, Maechler, & Bolker, 2012). For the production experiments (Experiments 1, 2 and 3), we used a Gaussian error distribution (assuming normality). We adhere to the random effect specification principles outlined in Barr, Levy, Scheepers, and Tily (2013). We included a term for random intercepts for participants and items, which quantifies by-participant and by-item variability in overall vowel duration (i.e., the fact that some speakers tend to produce longer or shorter vowels). The critical fixed effect in question was VOICING (i.e., voiced vs. voiceless in the plural form, where voiced = /b,d,g/ and voiceless = /p,t,k/), and for this fixed effect, we included correlated random slopes for participants and items (this quantifies by-participant and by-item variability in the effect of VOICING).

In our model selection process, we conceptually separated the fixed effects into control variables and the test variable (VOICING). ACCENT TYPE and PROSODIC BOUNDARY were two prosodic control variables. If either one of these had led to a significant interaction with VOICING, this would have indicated that the amount of incomplete neutralization depends on prosodic conditions, and could have been the result of hyperarticulation due to e.g. prosodic strengthening (e.g. Cho & Keating, 2009). VOWEL QUALITY and PLACE OF ARTICULATION (bilabial vs. alveolar vs. velar) were also included to explain residual variance. Since processing differences could lead participants to perceive some singular forms as better matches to their corresponding plural forms than others, we included the results of our stimulus norming as an additional control variable, PLURAL ASSOCIATION.

We first tested whether VOICING interacted with the control variables by performing a likelihood ratio test between a model containing interactions and a model containing main effects only. We then excluded the interaction between VOICING and all control variables. *P*-values were generated using likelihood ratio tests.

3.2. Results

VOICING had a significant effect on vowel duration in the singular form ($\chi^2(1)=13.76$, $p<0.0002$), with vowels estimated to be 8.6 ms longer before devoiced stops rather than to voiceless stops (SE = 2.03 ms). The effect of VOICING on vowel duration was fairly consistent across participants and items, as can be seen in Fig. 1. Overall, 14 out of 16 participants and 20 out of 22 items exhibited longer vowels preceding devoiced stops than preceding voiceless stops. Descriptive inspection of the data did not indicate that the effect was dependent on any item specific phonotactic characteristics (cf. Appendix A for a detailed listing of vowel differences for each stimulus pair separately). This was statistically validated: there were no interactions between VOICING and any of the control variables ($\chi^2(10)=9.45$, $p=0.49$). This means that the effect of VOICING on vowel duration did not depend on either of the prosodic variables (ACCENT TYPE or PROSODIC BOUNDARY),[5] nor on the variables VOWEL QUALITY, PLACE OF ARTICULATION, or PLURAL ASSOCIATION.

[4] Here and subsequently we report likelihood ratio tests between hierarchical linear regression models ("mixed models") with the fixed effect VOICING and random intercepts Item (no random effect for Speaker is needed as there is only one Speaker), as well as random slopes for Voicing dependent on Item. *P*-values were corrected for multiple testing by means of Dunn–Šidák correction.

[5] The target word was deaccented (EIN *Gop wollte nicht mehr*), or accented in prenuclear (*Ein Gop* WOLLTE *nicht mehr; Ein Gop wollte* NICHT *mehr*) or nuclear position (*Ein* GOP *wollte nicht mehr*).

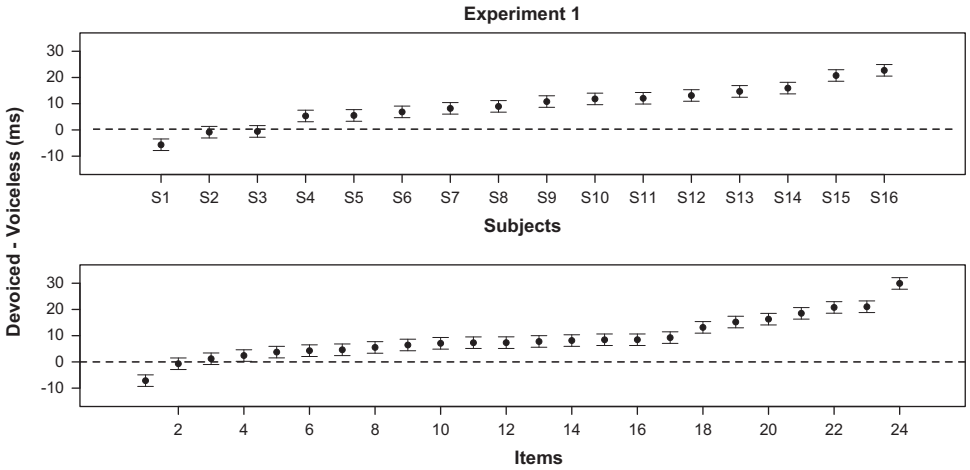

Fig. 1. Results of Experiment 1. Difference in vowel duration between stops in final position corresponding to voiced and voiceless stops in intervocalic position ('devoiced' and 'voiceless', respectively). Means are arranged according to size for all subjects (upper plot) and items (lower plot) separately. Error bars indicate standard errors taken from the model described in Section 3.1.5. Dashed lines indicate no difference between devoiced and voiceless stops.

Pseudowords are by definition unusual for our participants, and so we were concerned about problems with the task. To see whether our results might be disproportionately affected by a few extremely unusual responses, we performed subset analyses in which we excluded all responses where the pseudoword was either incorrectly remembered (e.g., substituting the vowel /iː/ for /eː/), or produced with considerable hesitation. This led to a removal of 10.02% of our data points, a considerable reduction of the size of the dataset. Nevertheless, even with these responses excluded, the main effect still obtains ($\chi^2(1)=13.34$, $p<0.0003$), with vowels being 8.6 ms longer (SE = 2.04 ms) before devoiced stops.

Our results indicate a successful extension of the FI study of incomplete neutralization in German, i.e., speakers produce longer vowels before stops corresponding to voiced stops in the plural form ('devoiced') than to the voiceless stops in the plural form ('voiceless'). However, there is a potential confound inherent to our design: namely, the pronounced differences in vowel duration between intervocalic voiceless and voiced stops in the acoustic stimuli (Section 3.1.3). This raises the possibility that the observed effect is simply an artifact of phonetic accommodation: since all stimuli were produced by a single speaker, participants may simply have been imitating the vowel duration differences when producing the singular forms with the stop in final position. This would be in line with previous research demonstrating that speakers shift their pronunciation towards productions of auditorily presented voices they have just heard (e.g., Babel, 2012; Goldinger, 1998; Nielsen, 2011). Experiments 2 and 3 were conducted to address this issue. By systematically manipulating the durational cue of the intervocalic stops in the plural forms, we can evaluate the potential impact of phonetic accommodation on the observed incomplete neutralization effect.

4. Production Experiment 2

4.1. Methodology

4.1.1. Participants and experimental procedure

Sixteen speakers participated in Experiment 2 (mean age: 27 years; 9 women). Background details of participants are as stated for E1. None of the participants in E2 had participated in E1. All details of the procedure were the same as in E1 if not stated otherwise.

In Experiment 2, we used a different carrier sentence. For each trial, participants first heard a sentence such as (8) and then produced a corresponding sentence such as (9).

Plural stimulus :

Peter weiβ nun, wie die Bauge aussehen. (8)

Peter know 3SG.PRS now how DET.PL.NOM NONCE-PL look
Peter knows now what the NONCE-PL look like.

Singular response :

Denn nur der Baug sieht so aus. (9)

Because only DET.SG.M.NOM NONCE-SG look.3SG.PRS like PART
As only the NONCE-SG looks like this.

4.1.2. Speech material, stimulus manipulation and norming

The experimental items consisted of 48 pseudoword pairs (see Appendix B). There were 24 stimulus pairs with labial and 24 with velar stops, each of which followed one of the vowels /iː, eː, aː, aʊ̯, oː, uː/.[6] To minimize measurement difficulties, alveolar stops (which may show coarticulatory effects of the following word) were excluded.

[6] Acoustical analyses of the stimuli show that the vowel preceding the critical stop was on average 23.78 ms (SE = 2.62) shorter before voiceless stops ($\chi^2(1)=48.547$, $p<0.0001$). The mean closure duration was 13.7 ms (SE = 1.74) longer for voiceless stops ($\chi^2(1)=40.32$, $p<0.0001$) and VOTs were on average 47.34 ms longer for voiceless stops (SE = 1.75, $\chi^2(1)=208.28$, $p<0.0001$). All of the voiced stimuli had voicing during the closure. Thus, as in E1, robust cues to the voicing status of the critical stop were present in intervocalic position.

Stimuli were balanced for vowel quality. As in E1, each experimental item was introduced as a masculine noun inflected for plural, and a norming study found that the schwa-plural (/-e/) was the most frequent response pattern (82% of all responses). Unlike E1, there were no fillers, making the contrast between the corresponding members of a minimal pair more obvious to participants and potentially leading to an enhancement of the effect under investigation (cf., Jassem & Richter, 1989), which in turn might make a potential confound effect of accommodation easier to detect. The 48 stimulus pairs were spoken by a native speaker of German (male, trained phonetician) along with the demonstration and practice items in a sound-treated booth recorded with an AKG C420 III microphone.

To evaluate the potential impact of phonetic accommodation, we manipulated the duration of the vowel preceding the intervocalic stops in the plural forms (e.g. /oː/ in /goːpə/). We took the mean difference in vowel duration preceding voiced and voiceless stops produced by the speaker as a baseline: Vowels preceding voiced stops were 16% longer than vowels preceding voiceless stops. We then manipulated vowel durations of both members of a minimal pair using TD-PSOLA (Time-Domain Pitch Synchronous OverLap-Add, Moulines & Charpentier, 1990) resynthesis as implemented in Praat (Boersma & Weenink, 2011) selecting a 10 ms Hanning window for analysis. Fundamental frequency was not manipulated. As this was a between-items design, each stimulus pair was manipulated only once and assigned to one of four sets. Stimuli in set A were edited to have a difference in vowel duration of 32% (henceforth *enhanced*); that is, vowels preceding voiced stops were 32% longer than vowels preceding voiceless stops (twice as long as the baseline condition). Stimuli in set B were edited to have a difference in vowel duration of 16% (henceforth *original*), similar to the baseline. Vowel durations of stimuli in set C did not differ at all (henceforth *neutralized*), meaning that vowel duration as a cue to intervocalic voicing was neutralized. Stimuli in set D were manipulated so that vowels preceding *voiceless* stops were 16% longer than those preceding voiced stops (henceforth *reversed*). In other words, set D contained stimuli where the effect of voicing on vowel duration was the mirror image of the baseline. The stimuli were judged to sound natural by a native speaker of German.

Additionally, we examined the perceptual robustness of the voicing distinction in the manipulated forms by conducting a norming study. Five native speakers of German (mean age: 25) were asked to decide whether the presented stimuli were voiced or voiceless in a forced-choice identification task. The norming study confirmed that the voicing contrast is very easy to perceive for all manipulation conditions: participants did not make any errors in identifying the voicing category. Even though we manipulated one perceptual cue to the voicing distinction, participants were able to rely on other cues like voicing during the closure, VOT and closure duration.

4.1.3. Stimulus presentation, acoustic analyses of responses and statistics

All stimulus presentations were randomized for each participant. The actual experiment was divided into four blocks. The first two blocks contained all 48 critical pairs, balanced for place of articulation of the stop, vowel quality and condition (voiced or voiceless). A subset of these items was repeated twice in blocks three and four. Corresponding members of a minimal pair in the first two blocks were separated by one block (so by at least 24 items). The acoustic analysis and statistical analysis was performed as specified for E1. In our model selection process, we separated the fixed effects into control variables (PLACE OF ARTICULATION, VOWEL QUALITY, PLURAL ASSOCIATION and REPETITION) and test variables (VOICING and MANIPULATION CONDITION). Statistical analyses were performed as specified for E1.

4.2. Results

We found an interaction between MANIPULATION CONDITION and VOICING ($\chi^2(1)=7.01$, $p=0.008$). For each manipulation step (*enhanced*⩾*original*⩾*neutralized*⩾*reversed*), the estimated difference between devoiced and voiceless stops became 1.49 ms smaller (SE = 0.57 ms). Interestingly, an incomplete neutralization effect was observed even in the *neutralized* and the *reversed* conditions, where the duration cue was at best uninformative. There was also a main effect of VOICING ($\chi^2(1)=12.76$, $p=0.00035$), with vowels being overall 4.3 ms shorter (SE = 1.02 ms) before voiceless stops than before devoiced stops (pooled across different manipulation conditions; see Fig. 2).

As in Experiment 1, we did not find any interactions between VOICING and any of the control variables ($\chi^2(8)=3.54$, $p=0.89$), suggesting that PLURAL ASSOCIATION, PLACE OF ARTICULATION, VOWEL QUALITY and REPETITION, did not have an effect. Subsequent inspection did not reveal any interaction of item specific phonotactic characteristics (see Appendix B for a detailed listing of vowel differences for each stimulus pair).

The results of Experiment 2 show that manipulation of the vowel duration in the plural stimulus affected the degree to which neutralization was incomplete. Nonetheless, there was still a significant overall effect of incomplete neutralization in the expected direction, even in the *reversed* condition. In other words, even though in a quarter of cases the input stimuli exhibited shorter vowel durations preceding *voiced* stops, participants produced shorter vowel durations preceding *voiceless* stops, suggesting that the effect of accommodation, if present, was at best small.

However, since this experiment employed a between-item design, all participants were prompted with items from all four manipulation conditions. Thus, we cannot rule out the possibility that stimuli of one condition might have influenced those of other conditions ("carry-over effects"). In addition, the manipulation conditions were not perfectly balanced; there was an overall duration advantage for vowels preceding voiced stops of +16% (adding up all conditions, 32%, 16%, 0%, –16%). This advantage is actually biased towards incomplete neutralization, as participants might have adapted to the overall 16% vowel duration difference, which might explain the persistence of the effect even in the *neutralized* and *reversed* conditions. To rule out this possibility, we conducted a third experiment with a between-*subjects* design and balanced manipulation conditions.

5. Production Experiment 3

5.1. Methodology

5.1.1. Participants and experimental procedure

Sixteen speakers participated in Experiment 3 (mean age: 24 years; 10 women). Background details of participants are as stated for E1 and E2. None had participated in the previous experiments. All details of the procedure were the same as in E2 if not stated otherwise.

(*footnote continued*)
As there was no interaction between manipulation condition and voicing for the parameters ($\chi^2(1)\leq 3.91$, $p\geq 0.27$), we may conclude that there were no differences of intervocalic voicing cues between conditions.

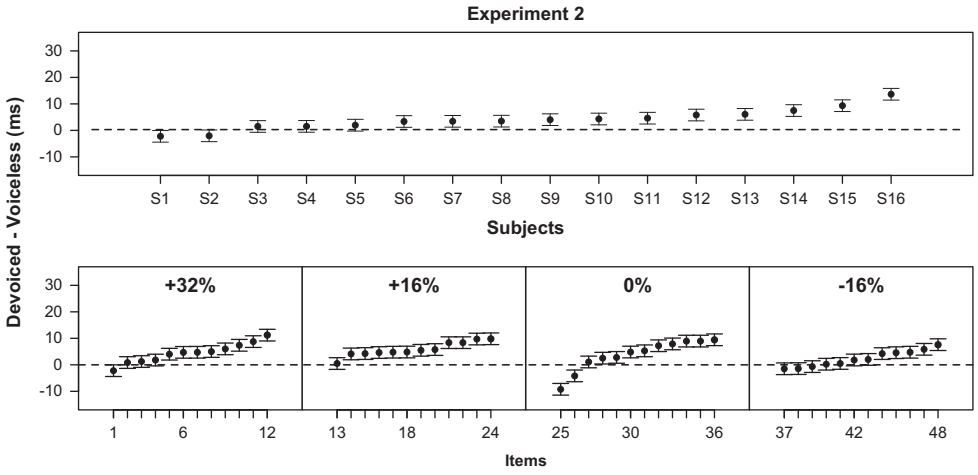

Fig. 2. Results of Experiment 2. Difference in vowel duration between devoiced and voiceless stops. Means are arranged according to size for all subjects (upper plot) and items (lower plot) separately. The lower plot shows vertical solid lines that separate the between-item manipulations of vowel duration. Dashed lines indicate no difference between voiced and voiceless stops.

5.1.2. Speech material, stimulus manipulation and norming

The 24 stimulus pairs consisted of a subset of the items used in E2 (cf. Appendix B).[7,8] Again, we manipulated the vowel durations as stated for E2. Each minimal pair was manipulated twice resulting in two stimulus sets: In set A, the difference in vowel duration was 32% (henceforth *enhanced*), that is vowels preceding underlying voiced stops were 32% longer than vowels preceding voiceless stops. In set B, the difference in vowel duration was 32% in the opposite direction (henceforth *reversed*).

As was done for E2, we examined the perceptual robustness of the voicing distinction in the manipulated forms by conducting a norming study. Five native speakers of German (mean age: 24) were asked to decide whether the presented stimuli were voiced or voiceless in a forced-choice identification task. All stimuli in both manipulation conditions were presented to all participants. The norming study confirmed that the voicing contrast is very easy to perceive: for the enhanced condition, participants were 100% correct in identifying the voicing of a stop, and for the reversed condition they were 99% (=3 incorrect tokens) correct.

5.1.3. Stimulus presentation, acoustic analyses of responses and statistics

Subjects were randomly assigned to one of two groups. Group A was presented with stimuli from set A only (*enhanced* stimuli), and group B was presented with stimuli from set B only (*reversed* stimuli). All stimuli were randomized for each participant. The actual experiment was divided into three blocks. In each block each stimulus was presented once resulting in three productions of each stimulus. The acoustic and statistical analyses were performed as specified for E2.

5.2. Results

As opposed to Experiment 2, there was no interaction between MANIPULATION CONDITION and VOICING ($\chi^2(1)=1.25$, $p=0.26$). However, numerically there was a small impact of manipulation condition, with the incomplete neutralization effect being 2.69 ms (SE=2.42) smaller in the *reversed* condition. For the *enhanced* condition, the predicted difference between devoiced and voiceless stops was 4.1 ms (SE=3.12). The difference between the two manipulation conditions, while not significant, resembles the effect seen in E2; however, the main effect of VOICING did not reach significance ($\chi^2(1)=1.62$, $p=0.2$), with vowels being only 1.75 ms shorter (SE=1.32 ms) before voiceless stops (cf. Fig. 3). Experiment 3 thus marks a failure to replicate the incomplete neutralization effect. Subsequent inspection of the results did not indicate any interaction of underlying voicing with item-specific phonotactic characteristics (cf. Appendix B for a detailed listing of vowel differences for each stimulus pair).

Before we proceed to the perception experiment, we summarize the results of the three production experiments and discuss their implications for the incomplete neutralization debate.

5.3. Discussion of production results

In Experiment 1 we found a difference in vowel duration depending on the voicing status of the intervocalic stop in the (plural) stimulus form. Thus, we were able to demonstrate that neutralization of the voicing contrast in final position is incomplete in terms of the duration of the preceding vowel, even when the influence of orthography was minimized by using auditory presentation of pseudowords (which are presumed to lack pre-existing orthographic representations). The pattern was found to be consistent across different individuals and stimuli even when controlling for variation between different participants and items. Furthermore, we found no interactions between the incomplete neutralization effect and any of the other variables that we controlled for. This is noteworthy, as it suggests that the incomplete neutralization effects were not altered by prosodic characteristics or place of articulation, suggesting relative independence from these factors.

[7] Acoustical analyses of the stimuli show that the vowel preceding the critical stop was 16.56 ms (SE=3.86) shorter before voiceless stops ($\chi^2(1)=14.12$, $p=0.00017$). The closure duration was 13.71 ms (SE=1.95) longer for voiceless stops ($\chi^2(1)=27.62$, $p<0.0001$). VOTs were on average 47.54 ms long for voiceless stops (SE=2.84, $\chi^2(1)=94.04$, $p<0.0001$). All of the voiced stimuli had voicing during the closure. So as stated for E1 and E2, there were robust cues for the voicing status of the critical stop in intervocalic position.

[8] Due to a coding error, one stimulus pair had to be excluded from the analysis.

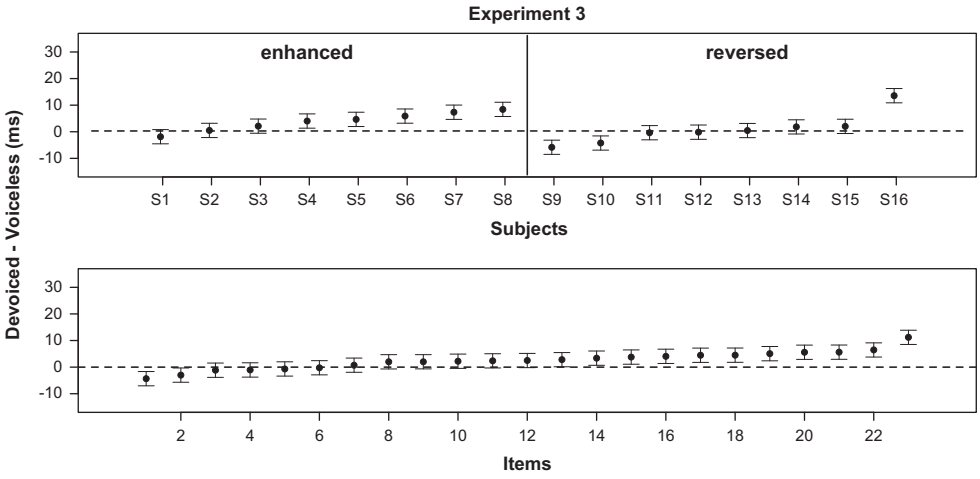

Fig. 3. Results of Experiment 3. Difference in vowel duration between devoiced and voiceless stops. Means are arranged according to size for all subjects (upper plot) and items (lower plot) separately. The upper plot shows a vertical solid line that separates the between-participants manipulation. Dashed lines indicate no difference between devoiced and voiceless stops. The lower plot shows 23/24 items because one item had to be excluded due to a coding error.

Finally, our debriefing indicated that participants perceived the task to be a morphological one – none were aware of the fact that we were looking specifically at minimal pairs such as [goːbə] and [goːpʰə]. This suggests that our distraction devices (instructions, difficult fillers, different city names) were successful, and that task demands and strategic responses were unlikely to play a significant role. We can then safely conclude that we have demonstrated the existence of an incomplete neutralization effect while avoiding the methodological shortcomings that may have impacted previous findings.

Experiment 2 replicated the findings of Experiment 1 and ruled out a potential confound, namely accommodation to the input stimuli. We demonstrated an incomplete neutralization effect of vowel duration in four manipulation conditions. Participants produced incomplete neutralization effects in the expected direction even when they were prompted with intervocalic cues providing evidence in the opposite direction. Although there was a statistically significant difference between the manipulation conditions, and thus potentially a residual effect of accommodation, this effect was numerically very small.

The incomplete neutralization effect was even smaller in Experiment 3, where we manipulated vowel durations in a between-subjects design. Whereas in the *enhanced* condition, there still was a numerical difference between vowels before devoiced and voiceless stops that was of similar magnitude as in the other experiments, this difference was even further diminished in the *reversed* condition. The latter condition is strongly biased against an incomplete neutralization effect, as all of the stimuli are manipulated so as to make accommodation counteract the vowel duration differences predicted by incomplete neutralization. It should also be pointed out that a between-subjects design inherently reduces statistical power. It is therefore unsurprising that we failed to replicate an incomplete neutralization effect in Experiment 3. As has been observed repeatedly in the literature on final devoicing in Dutch and German, incomplete neutralization effects are brittle and can be difficult to detect with inferential statistics (Baumann, 1995; Fourakis & Iverson, 1984; Jongman et al., 1992; Warner et al., 2004).

We now turn to the role of orthography. Given that literate adult speakers constantly and habitually associate phonological with orthographic forms (Perre et al., 2009; Seidenberg & Tanenhaus, 1979; Ziegler & Ferrand, 1998), participants might have mentally constructed orthographic representations “on the fly”. Thus, a given participant that has just heard a pseudoword such as [goːbə] might have activated an orthographic mental representation of that word, despite our solely auditory task design. It should be emphasized, however, that the magnitude of the effect that we obtained for vowel duration is comparable to previous studies that *did* have orthographic representations as the input.

Although we minimized the role of orthography at least to the same extent as Fourakis and Iverson (1984), our use of pseudowords comes with its own set of problems. For one thing, as pseudowords are necessarily unknown and unfamiliar (and thus have a frequency of 0), they may be more likely to be hyperarticulated (see Whalen, 1991, 1992). This, however, does not seem to be the case in our data. The overall vowel durations in Experiment 1, for example, are lower than previously reported ones: our mean was 156 ms (SD = 44 ms), whereas Port and O'Dell (1985: 459) reported 202–305 ms and Charles-Luce (1985: 315) reported 184–211 ms, suggesting that relative to these other experiments, our participants were if anything hyperarticulating *less*.

Furthermore, pseudowords always introduce the possibility of analogy to real words. The difference in frequency of different V-C sequences in the lexicon is the main source of potential analogical asymmetries. For example, tense/long vowels tend to precede voiced bilabial/velar stops (e.g., /liːbə/ ‘love’, /fliːgə/ ‘fly’), while lax/short vowels before voiced bilabial/velar stops are very rare (e.g., /ɛbə/ ‘tide’). We addressed this issue by manual inspection of the data and adding place of articulation as an effect to our statistical models. We found no noteworthy pattern, suggesting that any item-specific effects of individual pseudowords are marginal.

Finally, we checked for the possibility that singular-plural formation was more difficult for some stimuli than for others, which could introduce a potential confound. Data collected in our norming studies were used to predict production results, but no effect of plural formation preference was observed. We conclude that any processing difficulties due to idiosyncratic properties of the stimuli are of minor importance for our results.

We now turn to the perception experiment.

6. Perception experiment

The production experiments confirmed that neutralization of the voicing contrast in final position is indeed incomplete. We have ruled out a number of potential methodological reasons for this incompleteness. However, as mentioned in Section 1, it is not clear what, if any, functional role incomplete neutralization plays in speech communication. To further our understanding of the perceptibility of incomplete neutralization, our fourth experiment sets out to replicate and extend earlier studies of incomplete neutralization in perception. Previous studies used auditory stimuli from a small set of speakers, or even just a single speaker (e.g., Kleber et al., 2010; Port & Crawford, 1989). But are listeners able to discriminate between final

stops corresponding to intervocalic voiced and voiceless counterparts when they are confronted with a multitude of speakers? For a more ecologically valid assessment of incomplete neutralization in perception, our experiment confronted listeners with productions from all of the participants of Experiment 1.

6.1. Methodology

6.1.1. Participants and experimental procedure

Sixteen listeners participated in the experiment, none of whom had participated in any of the preceding experiments. All participants were native speakers of German with no reported hearing deficits (mean age: 30 years; five women). Two of the participants were authors of this study (the first and the second author, both from the Cologne/Rhine region), neither of whom performed remarkably better or worse than naïve participants, thus showing that even extensive familiarity with the training stimuli does not affect the results of this experiment. The remaining participants were either living in Cologne or in Leipzig. Regardless of their origins, all participants claimed to speak non-dialectal Standard German.

Participants heard the response sentences spoken by the speakers of Experiment 1. They were asked to decide whether the presented stimulus corresponded to an intervocalic voiced or voiceless stop by choosing the appropriate written presentation of a word (e.g., *Gob* vs. *Gop*). These were presented on the left and the right side of the screen (counterbalanced), and participants had to press a left or right button on the computer keyboard. Because we expected ceiling effects in the direction of the voiceless response, the instructions emphasized that exactly half of the stimuli were from the set <b,d,g> and half were from the set <p,t,k>. In order to control for the possibility of a speed-accuracy trade-off, we also measured reaction times. The procedure was run using E-Prime 2.0 (Schneider, Eschman, & Zuccolotto, 2002).

6.1.2. Speech material

The experiment was designed to capture immediate success in perceiving the distinction between stops corresponding to voiced and voiceless stops in intervocalic position as well as long-term success over many trials and repetitions. Recall that Experiment 1 contained 748 critical stimuli (16 speakers∗24 items∗2 voicing conditions). In order to mitigate the potential effects of participant fatigue, we sampled a subset of these data (192 items) to use as stimuli in the perceptual study.

In order not to handpick particular items, semi-random subsets of 192 items (12 items per speaker) were sampled from the set of critical stimuli (the sampling was semi-random in order to insure that each speaker and each item was represented). In order to make sure that the stimuli included acoustic evidence of incomplete neutralization, we chose the first subset with a significant incomplete neutralization effect. Out of this subset we constructed four lists that constituted the four blocks of the experimental procedure. In each block, each stimulus pair (e.g., *Gob* vs. *Gop*) and each speaker appeared at least once. Each devoiced/voiceless combination came from the same speaker (e.g., *Gob* and *Gop* in list 1 were both from speaker 4). Given that there were 24 item pairs but only 16 speakers, 8 speakers had to be re-used and their voices appeared twice per block.

6.1.3. Statistics

There are several ways of analyzing this type of perception data. The traditional way works with d', a sensitivity index derived from Signal Detection Theory (Green & Swets, 1966). d' takes into account a subject's response bias (here, their inclination to respond with “voiced” or “voiceless”) in order to compute a measure of perceptual sensitivity. We calculated d' per subject, per item and per speaker voice and performed one-sample *t*-tests against $d'=0$ for each of these measures. While this traditional analysis already takes bias into account, we cannot use it to analyze the effects of other measures on accuracy, such as response times and trial order. For this, we used a mixed logistic regression model with “accuracy” (0 or 1) as the dependent measure. As fixed effects we included mean-centered RESPONSE TIMES, TRIAL ORDER and REPETITION. We included correlated random slopes for subject, item and speaker voice. If the *intercept* of this model is significantly above zero, we can conclude that participants are able to perceive the voicing contrast in the neutralization context with above chance accuracy.

6.2. Results

Fig. 4 displays d' per subject, item and speaker voice. Overall, d' was fairly low, indicating that sensitivity to the voicing of final stops, when response bias was controlled for, was poor. *T*-tests indicate that d' is significantly above zero by subjects ($t(15)=7.1$, $p<0.0001$), items ($t(23)=4.419$, $p=0.00019$) and speaker voices ($t(15)=3.79$, $p=0.0017$), with estimates of 0.25 (SE=0.036), 0.29 (SE=0.066) and 0.27 (SE=0.072), respectively.

In the mixed logistic regression analysis, there were no effects for TRIAL ORDER ($\chi^2(3)=5.53$, $p=0.14$) or REPETITION ($\chi^2(3)=5.01$, $p=0.17$). The absence of an effect of REPETITION indicates that participants were no more likely to respond correctly the second time they heard the same item spoken by the same voice. This suggests that there was no familiarization effect. The absence of an effect of TRIAL ORDER on accuracy indicates that there was no overall learning effect either. There was, however, a significant effect of RESPONSE TIME ($\chi^2(3)=8.88$, $p=0.03$). Although faster responses were less accurate, the decrease was very small, with only a 5% decrease in accuracy per SD of response times (log odds: −0.053, SE=0.027).

Crucially, the intercept of this analysis was positive and significant ($p<0.0001$), with an estimated overall accuracy of 55% (log odds: 0.35, SE=0.09). This indicates that listeners were, on average, more likely to respond correctly than incorrectly. If we add CORRECT VOICING as a predictor (whether the spoken word was the intervocalic counterpart of a voiced or voiceless stop), we can divide up the results according to whether tokens have voiced and voiceless counterparts and look at differences in accuracies for these two conditions. With this model, participants were not significantly above chance for devoiced stops (51.4%, log odds: 0.057, SE=0.07), but they were for voiceless stops ($p<0.0001$), with 58.66% (log odds: 0.29, SE=0.05), indicating that they were 1.3 times more likely to respond correctly when listening to a voiceless stop.

6.3. Discussion

The accuracy average of just 55% is barely above chance performance, and contrasts starkly with the near 100% accuracy averages obtained in the norming studies of Experiments 1–3. In addition, participants performed worse when responding to devoiced stops (average accuracy of just 51%), which was not significantly different from chance performance. In turn, the overall significant accuracy scores might be due to a ceiling effect, i.e., participants correctly identified voiceless stops more often than devoiced stops. Even though similar results were obtained in previous perception studies on incomplete neutralization (e.g., Port and O'Dell (1985) report mean accuracy values of 59%), the present results are the lowest accuracy scores reported in the literature.[9]

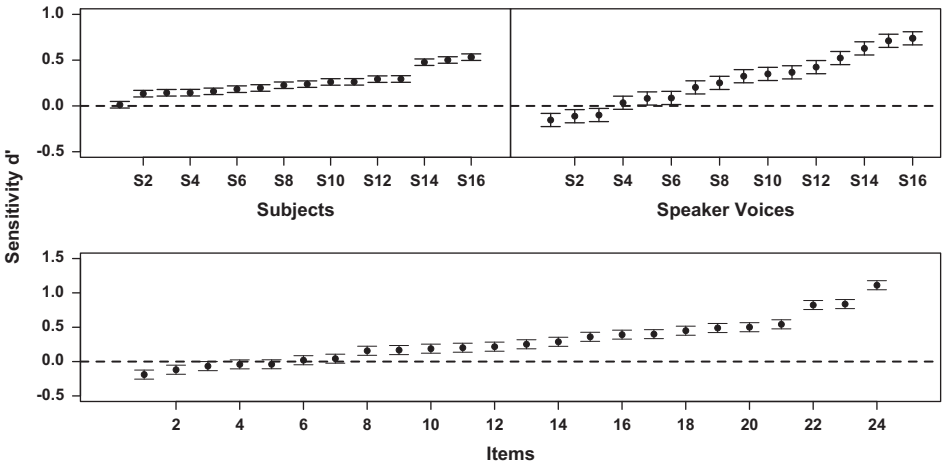

Fig. 4. Results of Experiment 4. *d'* sensitivity values arranged according to size for subjects (upper left plot), speaker-voices (upper right plot) and items (lower plot) with standard errors. Dashed lines indicate chance performance. There was no subject that scored below chance ($n=16$). There were only 5 items that scored below chance ($n=24$). And there were only 3 voices that scored below chance ($n=16$).

These low accuracy values naturally lead one to the question of whether incomplete neutralization plays any role in speech perception outside of the laboratory whatsoever. As there are only a handful of minimal pairs which are distinguished by the voicing specification of the final stop, and as these minimal pairs most often have different syntactic contexts which help to disambiguate them (e.g., the adjective *tot* 'dead' and the noun *Tod* 'death' would never appear in the same syntactic position), one could argue that the role of incomplete neutralization as a perceptual cue outside of a controlled laboratory context is negligible and unlikely to have a great deal of functional relevance in everyday speech communication. However, there is a great body of evidence demonstrating that fine phonetic detail (which may not be immediately perceptually detectable) is used in lexical access and spoken word recognition (e.g., Davis, Marslen-Wilson, & Gaskell, 2002; Hawkins & Nguyen, 2003). Thus, the present results should not necessarily be interpreted as evidence against the functional relevance of incomplete neutralization.

7. General discussion

A substantial number of experiments over the last three decades have reported minor acoustic differences between obstruents in a phonologically neutralizing context corresponding to voiced and voiceless counterparts intervocalically (which we have referred to as 'devoiced' and 'voiceless', respectively). The bulk of these studies were focused on German, although the finding has been advanced for other languages as well. As noted earlier, the findings of many of these studies have been called into question on methodological grounds, but studies purporting to provide counter-evidence such as Fourakis and Iverson (1984) or Jassem and Richter (1989) were at least as problematic (see Section 2). The aim of the present study was to put the debate surrounding incomplete neutralization on a firmer empirical footing by using an auditory design similar to that of Fourakis and Iverson, but with a larger sample of subjects and items.

We found vowel duration to be a robust acoustic correlate of devoiced and voiceless stops in syllable-final position: vowels were longer before devoiced stops than before voiceless stops. By using a different methodology from previous studies, this study contributes to the converging evidence that neutralization of German final stops is incomplete. Our finding that incomplete neutralization emerges even in a completely auditory task, and when using pseudowords instead of real words, was replicated in two additional experiments that mitigated the potential for accommodation to an intervocalic input stimulus.

We also conducted a perception experiment, in which we found that although participants were able to distinguish devoiced from voiceless final stops, their accuracy was barely above chance performance (55%, as opposed to 98–100% intervocalically in the norming studies). Moreover, this overall above-chance accuracy was largely driven by greater accuracy in correctly identifying voiceless stops; participants were at chance when identifying devoiced stops. This speaks to limitations with respect to the perception of incomplete neutralization, indicating that, at least within a forced choice paradigm, pre-stop vowel length is not a robust cue to voicing category in final position. Thus, while the present experiments provide robust evidence for incomplete neutralization in production, it remains unclear whether listeners actually use these small differences in perception.

The acceptance of any phenomenon should never be based on a single study, and several studies, such as Fourakis and Iverson (1984), have been overemphasized relative to the totality of incomplete neutralization studies (see Winter and Roettger, 2011, for discussion). Only by accumulating converging evidence from different methodologies can we be more certain about whether neutralization is complete or not. To date, studies finding evidence of incomplete neutralization (both for German and for other languages) outnumber those finding counter-evidence, suggesting that statements like those of Wiese (1996: 205) that "[t]hese results are rather tentative […] given that the recognition of non-neutralized devoicing was found in a minority of cases only" can safely be said to have been superseded. Positive results for incomplete neutralization characterize the majority of studies on this topic and several of the methodological concerns raised in earlier work have now been successfully addressed.

Since the body of evidence is in favor of incomplete neutralization, we now turn to how it can be accounted for. Accepting that neutralization was incomplete was previously thought to entail changes in phonological theory; early work assumed that the obtained differences had to be explained in terms of differences in abstract representations or different ordering of implementation rules (e.g., Brockhaus, 1995; Charles-Luce, 1985; Piroth & Janker, 2004; Port & O'Dell, 1985).

[9] It might be argued that the very low accuracy scores may also be due to the dialectal background of the subjects. Even though subjects reported speaking standard German, they came from the Central Franconian and Saxon dialect area. Dialects or regional varieties are still spoken in both areas, and most subjects are likely to have been exposed to them. For example in Saxon dialects, the Central German lenition rule operates, i.e., the voiced/voiceless contrast is neutralized (through fortis stop lenition) in all positions including intervocalically. As a result, the performance of Saxon listeners in perceiving the voicing contrast in intervocalic stops is generally lower (John, 2004). This interpretation, however, stands in contrast to the very high accuracy scores in intervocalic position we found in the norming studies of Experiments 1–3.

Given the subtle and perhaps barely detectable nature of incomplete neutralization, it is reasonable to question whether it is really necessary to capture such a small effect in terms of reflexes of abstract linguistic entities. However, there are alternative approaches of incomplete neutralization, to which we now turn.

Lexical representations are now commonly assumed to contain considerably fine-grained and redundant types of information, including phonetic detail of completely inflected forms (Alegre & Gordon, 1999; Baayen et al., 1997; Brown & McNeill, 1966; Bybee, 1994, 1995; Butterworth, 1983; Goldinger, 1996, 1997; Manelis & Tharp, 1977; Palmeri et al., 1993; Pisoni, 1997; Sereno & Jongman, 1997). Ernestus and Baayen (2006) propose the possibility of incomplete neutralization effects being due to the co-activation of paradigmatically related forms, i.e., when speakers pronounce *Rad*, they also activate the non-neutralized *Räder*. This co-activation of the related voiced forms could influence the speech production mechanism in subtle ways, leading to incomplete neutralization.

This hypothesis is based on the concept of spreading activation (e.g., see references in Collins & Loftus, 1975). With respect to morphological relations, there is evidence for the automatic activation of morphological "neighbors" in perception, whereby words with more or more frequent neighbors may be recognized more quickly and/or more accurately than those with fewer or less frequent neighbors (e.g., Andrews, 1989; Sears, Hino & Lupker, 1995). One might object that most previous evidence for activation of lexical neighbors comes from perception studies. However, recent studies have also demonstrated the effects of lexical neighbors on speech production. Baese-Berk and Goldrick (2009), for instance, showed how production of VOT is modulated on-line depending on whether or not a word is presented in context with its minimal pair neighbor, while Munson (2007) found greater vowel-space expansion for words with larger numbers of neighbors (see also Wright, 2004).

The co-activation account has two advantages over traditional accounts of incomplete neutralization. First, from a functionalist perspective, an account that treats incomplete neutralization as an artifact of lexical representations is more attractive than an account based on phonetic or phonological rules and/or representations that are extracted from auditory information. Under the co-activation account, speakers would not need to extract any subtle contrast from the signal (e.g. incomplete neutralization effects in *Rad* vs. *Rat*) so long as they can perceive the contrast between the corresponding paradigmatic neighbors (e.g. *Räder* vs. *Räte*). The presence of the contrast somewhere in the paradigm leads automatically to the prediction of incomplete neutralization. Thus the acoustic cues found in the neutralized position have no functional utility and are not reliably used in regular communication to differentiate between minimal pairs. This interpretation is in line with the low accuracy scores in perception experiments.

Second, the co-activation account makes testable predictions for future experiments. For one thing, it predicts recency effects: if the response is delivered immediately following the stimulus, the effect should be stronger than after a longer time interval. This is because spreading activation generally recedes over a relatively short time span. Furthermore, there should be frequency effects: words that have very frequent neighbors with voiced stops in intervocalic position should exhibit stronger incomplete neutralization effects than words with very infrequent neighbors (see e.g., Bybee, 2001, for the role of frequency in analogy). Moreover, it predicts incomplete neutralization effects to be dependent on lexical density. This means that a word with many voiceless lexical neighbors should surface with no (or at least weaker) incomplete neutralization effects compared to a word with many voiced lexical neighbors. Recall that in German, voiced and voiceless stops following tense/long vowels are unequally distributed depending on the place of articulation: tense/long vowels tend to precede voiced bilabial and velar stops (e.g., /liːbə/ 'love', /fliːgə/ 'fly'), while lax/short vowels are very rare in this environment (e.g., /ɛbə/ 'tide'). Given the co-activation hypothesis, forms with many lexical neighbors containing a voiced stop should show stronger incomplete neutralization effects. In line with that, we would expect that pseudoword pairs ending in alveolar stops (e.g., *Frade/Frad*) co-activate more lexical neighbors with voiced stops than pseudoword pairs ending in bilabial or velar stops (e.g., *Gobe/Gob, Schuge/Schug*). This predicts that in German the degree of incomplete neutralization may be modulated by place of articulation. This hypothesis, however, could not be confirmed; in none of our models did we find a statistically significant effect of place of articulation or even a numerical trend in the predicted direction.

However, we would like to point out once again that the experimental task employed in this and similar studies is subject to pragmatic limitations due to small effect sizes and considerable degrees of variation. Showing a statistically significant effect of incomplete neutralization is already difficult, and finding significant differences in effect sizes due to factors such as frequency asymmetries in the mental lexicon is more challenging still. This points to the practical limits of investigating incomplete neutralization, and of using incomplete neutralization as a test bed for investigating the cognitive architecture of the lexicon: in the absence of viable strategies of strengthening the effect, research on incomplete neutralization will always have to cope with high Type II error rates.

Table A1
Critical stimuli of E1.

[+voice]		[−voice]		Place	Δ devoiced − voiceless in ms
Blode	[bloːdə]	Blote	[bloːtʰə]	Alveolar	7.1
Drude	[dʁuːbə]	Drute	[dʁuːtʰə]	Alveolar	15.2
Flabe	[flaːbə]	Flape	[flaːpʰə]	Bilabial	−7.2
Frade	[fʁaːdə]	Frate	[fʁaːtʰə]	Alveolar	7.8
Froge	[fʁoːgə]	Froke	[fʁoːkʰə]	Velar	20.8
Frube	[fʁuːbə]	Frupe	[fʁuːpʰə]	Bilabial	4.3
Gage	[gaːgə]	Gake	[gaːkʰə]	Velar	3.7
Gaude	[gaʊ̯də]	Gaute	[gaʊ̯tʰə]	Alveolar	9.2
Gobe	[goːbə]	Gope	[goːpʰə]	Bilabial	4.6
Griede	[gʁiːdə]	Griete	[gʁiːtʰə]	Alveolar	7.3
Jiede	[jiːdə]	Jiete	[jiːtʰə]	Alveolar	29.9
Klabe	[klaːbə]	Klape	[klaːpʰə]	Bilabial	6.4
Mube	[muːbə]	Mupe	[muːpʰə]	Bilabial	13.1
Nauge	[naʊ̯gə]	Nauke	[naʊ̯kʰə]	Velar	8.1
Priege	[pʁiːgə]	Prieke	[pʁiːkʰə]	Velar	8.4
Pruge	[pʁuːgə]	Pruke	[pʁuːkʰə]	Velar	21.0
Quade	[kvaːdə]	Quate	[kvaːtʰə]	Alveolar	−0.7
Quobe	[kwoːbə]	Quope	[kwoːpʰə]	Bilabial	8.4
Roge	[ʁoːgə]	Roke	[ʁoːkʰə]	Velar	18.5
Schmaube	[ʃmaʊ̯bə]	Schmaupe	[ʃmaʊ̯pʰə]	Bilabial	2.4
Schriege	[ʃʁiːgə]	Schrieke	[ʃʁiːkʰə]	Velar	7.3
Schuge	[ʃuːgə]	Schuke	[ʃuːkʰə]	Velar	16.3
Stauge	[ʃtaʊ̯gə]	Stauke	[ʃtaʊ̯kʰə]	Velar	5.5
Wiebe	[viːbə]	Wiepe	[viːpʰə]	Bilabial	1.2

8. Conclusion

The primary goal of this paper was to assess whether or not neutralization in German final stop voicing is indeed incomplete. We demonstrated the robustness of an effect on production in three production experiments, ruling out a number of claims that the incompleteness is a purely methodological artifact, and arguing that even if non-functional, the robustness of incomplete neutralization warrants explanation. We would like to emphasize that our results are crucially independent of whatever mechanism actually explains incomplete neutralization. Phonologists have been justifiably skeptical of the previous evidence arguing for incomplete neutralization, but as we have reviewed above, incomplete neutralization does not necessarily have to be explained in terms of representational differences; more parsimonious accounts are suggested by existing experimental work on lexical co-activation. Such accounts seem to us to be fruitful avenues for further investigations (cf., discussion in Winter & Roettger, 2011). Manaster-Ramer (1996: 487) used the incomplete neutralization debate as a call for an increased collaboration between phonologists and phoneticians. In Manaster-Ramer's words (Manaster-Ramer, 1996: 487), "Phonologists cannot afford to be neutral" with respect to incomplete neutralization. We have shown that the phenomenon can be seen in a different light if psycholinguistic and cognitive evidence is taken into account. We would like to extend Manaster-Ramer's call in the hopes that we may gain new perspectives on old problems by engaging with work from related disciplines.

Acknowledgments

This work has benefited from feedback we received for presentation at the ICPhS 2011 and LabPhon 2012. We would like to thank all anonymous reviewers and the editors of Journal of Phonetics for useful comments and suggestions. Furthermore, we would like to thank Klaus Kohler for his remarks on our work. Finally, we thank all of our participants for their participation in our tiresome experiments. Remaining errors are our own.

Table B1
Critical stimuli of E2 and E3 (in bold).

[+voice]		[-voice]		Place	E2 Δ devoiced−voiceless in ms	E3 Δ devoiced−voiceless in ms
Bauge	**[baʊ̯gə]**	**Bauke**	**[baʊ̯kʰə]**	**Velar**	4.0	NA
Bege	[beːgə]	Beke	[beːkʰə]	Velar	11.2	
Blebe	**[bleːbə]**	**Blepe**	**[bleːpʰə]**	**Bilabial**	−0.4	4.5
Bloge	**[bloːdə]**	**Bloke**	**[bloːtʰə]**	**Velar**	4.8	5.6
Dage	[daːgə]	Dake	[daːkʰə]	Velar	−0.5	
Dabe	[daːbə]	Dape	[daːpʰə]	Bilabial	17.3	
Diege	[diːgə]	Dieke	[diːkʰə]	Velar	4.2	
Dobe	**[doːbə]**	**Dope**	**[doːpʰə]**	**Bilabial**	4.2	−0.7
Drube	**[dʁuːbə]**	**Drupe**	**[dʁuːpʰə]**	**Bilabial**	8.9	5.0
Duge	[duːgə]	Duke	[duːkʰə]	Velar	5.5	
Fage	**[faːgə]**	**Fake**	**[faːkʰə]**	**Velar**	9.8	4.0
Faube	[faʊ̯bə]	Faupe	[faʊ̯pʰə]	Bilabial	0.5	
Flabe	[flaːbə]	Flape	[flaːpʰə]	Bilabial	−0.2	
Flebe	**[fleːbə]**	**Flepe**	**[fleːpʰə]**	**Bilabial**	−0.2	2.0
Frebe	[fʁeːbə]	Frepe	[fʁeːpʰə]	Bilabial	4.7	
Froge	**[fʁoːgə]**	**Froke**	**[fʁoːkʰə]**	**Velar**	8.9	3.4
Frobe	[fʁoːbə]	Frobe	[fʁoːpʰə]	Bilabial	−0.5	
Frube	[fʁuːbə]	Frupe	[fʁuːpʰə]	Bilabial	6.0	
Gage	[gaːgə]	Gake	[gaːkʰə]	Velar	0.5	
Gaube	**[gaʊ̯bə]**	**Gaupe**	**[gaʊ̯pʰə]**	**Bilabial**	9.4	6.4
Gauge	[gaʊ̯gə]	Gauke	[gaʊ̯kʰə]	Velar	1.8	
Glege	**[gleːgə]**	**Gleke**	**[gleːkʰə]**	**Velar**	7.2	2.2
Gliebe	**[gliːbə]**	**Gliepe**	**[gliːpʰə]**	**Bilabial**	4.8	−0.3
Gobe	[goːbə]	Gope	[goːpʰə]	Bilabial	8.3	
Griebe	[gʁiːbə]	Griepe	[gʁiːpʰə]	Bilabial	8.4	
Hege	[heːgə]	Heke	[heːkʰə]	Velar	4.6	
Klabe	**[klaːbə]**	**Klape**	**[klaːpʰə]**	**Bilabial**	4.8	−0.4
Krobe	**[kʁoːbə]**	**Krope**	**[kʁoːpʰə]**	**Bilabial**	4.7	3.8
Miebe	**[miːbə]**	**Miepe**	**[miːpʰə]**	**Bilabial**	4.6	5.6
Naube	[naʊ̯bə]	Naupe	[naʊ̯pʰə]	Bilabial	5.8	
Nauge	**[naʊ̯gə]**	**Nauke**	**[naʊ̯kʰə]**	**Velar**	2.8	−0.1
Nuge	[nʊːgə]	Nuke	[nʊːkʰə]	Velar	1.8	
Priege	[pʁiːgə]	Prieke	[pʁiːkʰə]	Velar	1.2	
Pruge	**[pʁuːgə]**	**Pruke**	**[pʁuːkʰə]**	**Velar**	2.0	11.2
Roge	[ʁoːgə]	Roke	[ʁoːkʰə]	Velar	1.1	
Schlabe	**[ʃlaːbə]**	**Schlape**	**[ʃlaːpʰə]**	**Bilabial**	−0.7	2.5
Schmaube	**[ʃmaʊbə]**	**Schmaupe**	**[ʃmaʊ̯pʰə]**	**Bilabial**	8.8	−0.2
Schriege	**[ʃʁiːgə]**	**Schrieke**	**[ʃʁiːkʰə]**	**Velar**	7.6	2.8
Schuge	**[ʃuːgə]**	**Schuke**	**[ʃuːkʰə]**	**Velar**	7.9	2.0
Spage	**[ʃpaːgə]**	**Spake**	**[ʃpaːkʰə]**	**Velar**	5.0	−0.0
Stauge	[ʃtaʊ̯gə]	Stauke	[ʃtaʊ̯kʰə]	Velar	9.7	
Strege	**[ʃtʁeːgə]**	**Streke**	**[ʃtʁeːkʰə]**	**Velar**	4.7	4.4
Sube	**[zʊːbə]**	**Supe**	**[zʊːpʰə]**	**Bilabial**	0.2	2.3
Triege	**[tʁiːgə]**	**Trieke**	**[tʁiːkʰə]**	**Velar**	5.3	0.7
Wiebe	[viːbə]	Wiepe	[viːpʰə]	Bilabial	7.4	
Wube	[vuːbə]	Wupe	[vuːpʰə]	Bilabial	4.1	
Wuge	[vuːgə]	Wuke	[vuːkʰə]	Velar	0.8	
Zebe	[tseːbə]	Zepe	[tseːpʰə]	Bilabial	5.8	

Appendix A

See Table A1.

Appendix B

See Table B1.

References

Abboud, H. (1991). *SuperLab*. Wheaton, MD: Cedrus.
Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, *40*, 41–61.
Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 802–814.
Baayen, R. H., Dijkstra, T., & Schreuder, S. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, *37*, 94–117.
Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*, 177–189.
Baese–Berk, M., & Goldrick, M. (2009). Mechanism of interaction in speech production. *Language and Cognitive Processes*, *24*, 527–554.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.
Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999–0.
Baumann, M. (1995). *The production of syllables in connected speech (Unpublished Ph.D. dissertation)*. University of Nijmegen.
Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how?. *Journal of Clinical Epidemiology*, *54*, 343–349.
Bishop, J. B. (2007). Incomplete neutralization in Eastern Andalusian Spanish: Perceptual consequences of durational differences involved in s-aspiration. In *Proceedings of the 16th ICPhS*, Saarbrücken (pp. 1765–1768).
Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer (Version 5.2) [Computer program].
Braver, A., & Kawahara, S. (2012). Complete and incomplete neutralization in Japanese monomoraic lengthening. Ms. Rutgers University.
Brockhaus, W. (1995). *Final Devoicing in the Phonology of German*. Max Niemeyer Verlag: Tübingen.
Brown, R., & McNeill, D. (1966). The 'tip of the tongue' phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*, 325–337.
Butterworth, B. (1983). Lexical representation. In: B. Butterworth (Ed.), *Language production*, vol. 2 (pp. 257–294). London: Academic Press.
Bybee, J. (1994). A view of phonology from a cognitive and functional perspective. *Cognitive Linguistics*, *5*, 285–305.
Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*, 425–455.
Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.
Charles-Luce, J. (1985). Word-final devoicing in German: Effects of phonetic and sentential contexts. *Journal of Phonetics*, *13*, 309–324.
Charles-Luce, J., & Dinnsen, D. (1987). A reanalysis of Catalan devoicing. *Journal of Phonetics*, *15*, 187–190.
Charles-Luce, J. (1993). The effects of semantic context on voicing neutralization. *Phonetica*, *50*, 28–43.
Cho, T., & Keating, P. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*, 466–485.
Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.
Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.
Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 218–244.
de Jong, K. J. (2011). Flapping in American English. In: M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *Blackwell Companion to Phonology* (pp. 2711–2729). Oxford: Wiley-Blackwell.
Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunez Filho, G., Jobert, A., et al. (2010). How learning to read changes the cortical networks for vision and language. *Science*, *330*, 1359–1364.
Dinnsen, D. A. (1985). A re-examination of phonological neutralization. *Journal of Linguistics*, *21*, 265–279.
Dinnsen, D. A., & Charles-Luce, J. (1984). Phonological neutralization, phonetic implementation and individual differences. *Journal of Phonetics*, *12*, 49–60.
Dinnsen, D. A., & Garcia–Zamor, M. (1971). The three degrees of vowel duration in German. *Papers in Linguistics*, *4*, 111–126.
Dmitrieva, O., Jongman, A., & Sereno, J. (2010). Phonological neutralization by native and non–native speakers: The case of Russian final devoicing. *Journal of Phonetics*, *38*, 483–492.
Ernestus, M., & Baayen, R. H. (2006). The functionality of incomplete neutralization in Dutch: The case of past-tense formation. In: L. M. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Laboratory Phonology*, 8 (pp. 27–49). Berlin: de Gruyter.
Fourakis, M., & Iverson, G. K. (1984). On the 'Incomplete Neutralization' of German final obstruents. *Phonetica*, *41*, 140–149.
Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*, 132–138.
Fuchs, S. (2005). Articulatory correlates of the voicing contrast in alveolar obstruent production in German. *ZAS Papers in Linguistics*, 41.
Gerfen, C. (2002). Andalusian Codas. *Probus*, *14*, 247–277.
Gerfen, C., & Hall, K. (2001). Coda aspiration and incomplete neutralization in Eastern Andalusian Spanish. Manuscript, University of North Carolina at Chapel Hill. Retrieved from: ⟨http://www.unc.edu/~gerfen/papers/GerfenandHall.pdf⟩.
Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.
Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. In: K. Johnson, & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 33–65). San Diego: Academic Press.
Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.
Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.
Gregory, S. W., & Hoyt, B. R. (1982). Conversation partner mutual adaptation as demonstrated by Fourier series analysis. *Journal of Psychological Research*, *11*, 35–46.
Greisbach, R. (2001). Experimentelle Testmethodik on Phonetik und Phonologie. Untersuchungen zu segmentalen Grenzphänomenen im Deutschen. Frankfurt a. M.: Lang.
Hawkins, S., & Nguyen, N. (2003). Effects on word recognition of syllable-onset cues to syllable-coda voicing. In: J. Local, R. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology VI* (pp. 38–57). Cambridge: Cambridge University Press.
Jassem, W., & Richter, L. (1989). Neutralization of voicing in Polish obstruents. *Journal of Phonetics*, *17*, 317–325.
Jespersen, O. (1913). *Lehrbuch der Phonetik* (2nd ed.). Leipzig: G.B. Teubner.
John, T. (2004). *Eine akustische Analyse der Lenis/Fortis—Opposition in Varietäten des Sächsischen (Unpublished MA thesis)*. University of Kiel.
Jongman, A., Sereno, J. A., Raaijmakers, M., & Lahiri, A. (1992). The phonological representation of [voice] in speech perception. *Language and Speech*, *35*, 137–152.
Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language*, *60*, 286–319.
Kharlamov, V. (2012). *Incomplete neutralization and task effects in experimentally-elicited speech: Evidence from the production and perception of word-final devoicing in Russian (Ph.D. dissertation)*. University of Ottawa.
Kleber, F., John, T., & Harrington, J. (2010). The implications for speech perception of incomplete neutralization of final devoicing in German. *Journal of Phonetics*, *38*, 185–196.
Kohler, K. J. (1984). Phonetic explanations in phonology: The feature fortis/lenis. *Phonetica*, *31*, 150–174.
Kohler, K. J. (2007). Beyond Laboratory Phonology. The phonetics of speech communication. In: M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology* (pp. 41–53). Oxford: Oxford University Press.
Kohler, K. J. (2012). Neutralization?! The phonetics–phonology issue in the analysis of word–final obstruent voicing. ⟨http://www.ipds.uni-kiel.de/kjk/pub_exx/kk2012_3/⟩. Retrieved 14.02.13.
Manaster–Ramer, A. (1996). A letter from an incompletely neutral phonologist. *Journal of Phonetics*, *24*, 477–489.
Manelis, L., & Tharp, D. A. (1977). The processing of affixed words. *Memory and Cognition*, *5*, 690–695.
Mitleb, F. (1981). Temporal correlates of "voicing" and its neutralization in German. *Research in Phonetics*, *2*, 173–191 (Bloomington, Indiana: Indiana University).
Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.
Munson, B. (2007). Lexical access, lexical representation, and vowel production. In: J. S. Cole, & J. I. Hualde (Eds.), *Laboratory phonology*, 9 (pp. 201–228). Berlin: de Gruyter.
Natale, M. (1975a). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, *32*, 790–804.

Natale, M. (1975b). Social desirability as related to convergence of temporal speech patterns. *Perceptual Motor Skills*, *40*, 827–830.
Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*, 132–142.
O'Dell, M., & Port, R. (1983). Discrimination of word-final voicing in German. *Journal of the Acoustical Society of America*, *73*(S1), S31 (A).
Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 309–328.
Perre, L., Midgley, K., & Ziegler, J. C. (2009). When beef primes reef more than leaf: Orthographic information affects phonological priming in spoken word recognition. *Psychophysiology*, *46*, 739–746.
Piroth, H. G., & Janker, P. M. (2004). Speaker-dependent differences in voicing and devoicing of German obstruents. *Journal of Phonetics*, *32*, 81–109.
Pisoni, D. (1997). Some thoughts on 'normalization' in speech perception. In: K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego: Academic Press.
Port, R., & Crawford, P. (1989). Incomplete neutralization and pragmatics in German. *Journal of Phonetics*, *17*, 257–282.
Port, R., & Leary, A. (2005). Against formal phonology. *Language*, *81*, 927–964.
Port, R., & O'Dell, M. (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics*, *13*, 455–471.
Port, R., Mitleb, F. M., & O'Dell, M. (1981). Neutralization of obstruent voicing in German is incomplete. *Journal of the Acoustical Society of America*, *70*(S13), F10.
R Core Team (2012). R: A Language and Environment for Statistical Computing. Vienna. ⟨http://www.R-project.org⟩.
Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh, PA: Psychology Software Tools, Inc.
Sears, C. R., Hino, Y., & Lupker, S. J. (1995). Neighborhood size and neighborhood frequency-effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 876–900.
Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 546–554.
Sereno, J., & Jongman, A. (1997). Processing of English inflectional morphology. *Memory and Cognition*, *25*, 425–437.
Simonet, M., Rohena-Madrazo, M., & Paz, M. (2008). Preliminary evidence for incomplete neutralization of coda-liquids in Puerto Rican Spanish. In: L. Colantoni, & J. Steele (Eds.), *Selected proceedings of the 3rd conference on laboratory approached to spanish phonology* (pp. 72–86). Somerville, MA: Cascadilla Press.
Slowiaczek, L., & Dinnsen, D. (1985). On the neutralizing status of polish word final devoicing. *Journal of Phonetics*, *13*, 325–341.
Slowiaczek, L., & Szymanska, H. (1989). Perception of word-final devoicing in polish. *Journal of Phonetics*, *17*, 205–212.
Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck and Ruprecht.
Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub–phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics*, *32*, 251–276.
Weitzman, R. A. (1984). Seven treacherous pitfalls of statistics, illustrated. *Psychological Reports*, *54*, 355–363.
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.
Wiese, R. (1996). *The Phonology of German*. Oxford: Clarendon Press.
Winter, B., & Roettger, T. B. (2011). The nature of incomplete neutralization in German. *Grazer Linguistische Studien*, *76*, 55–74.
Wright, R. A. (2004). Factors of lexical competition in vowel articulation. In: J. J. Local, R. Ogden, & R. Temple (Eds.), *Laboratory phonology*, 6 (pp. 26–50). Cambridge: Cambridge University Press.
Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, *38*, 329–336.
Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin and Review*, *5*, 683–689.
Zifonun, G., Hoffmann, L., Strecker, B., Ballweg, J., Braußе, U., Breindl, E., et al. (1997). *Grammatik der deutschen Sprache (Band 1)*. Berlin: de Gruyter.