

# THE ROBUSTNESS OF INCOMPLETE NEUTRALIZATION IN GERMAN

Timo B. Röttger<sup>a</sup>, Bodo Winter<sup>b,c</sup> & Sven Grawunder<sup>b</sup>

<sup>a</sup>IfL Phonetik, University of Cologne, Germany;

<sup>b</sup>Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Germany;

<sup>c</sup>Department of Cognitive and Information Sciences, University of California, Merced, USA

timo.roettger@uni-koeln.de; bodo@bodowinter.com; grawunder@eva.mpg.de

## ABSTRACT

“*Incomplete neutralization*” (IN) refers to the finding that subphonemic differences distinguish segments in a neutralizing context. Previous findings on IN have often been called into question because of methodological problems. We conducted a production and a perception study to address these previous concerns and to investigate the robustness of IN in German final stop devoicing. Underlying voiced stops were produced with significantly longer vowels preceding the stops, as well as with lower burst intensities than voiceless stops in word final position. The perception study confirmed that these differences are audible. The experiments extend previous work by showing that (1) IN effects occur even in completely non-orthographic experimental designs and (2) IN effects in perception occur even when listeners are subjected to many different voices.

**Keywords:** incomplete neutralization, final devoicing

## 1. INTRODUCTION

Previous research has shown that phonological final obstruent devoicing may be phonetically incomplete: In German, the word-final underlying voiced and voiceless obstruents exhibit small differences in preceding vowel duration, closure duration or burst duration (see [4, 9, 10]) and listeners are sensitive to this “*Incomplete Neutralization*” (IN) (see [8, 9, 10]). These results seemingly undermine a completely phonological account of final devoicing. However, the idea of IN has been challenged by studies which attempted to demonstrate that IN is merely an experimental artifact (e.g. [6]). A common criticism of IN experiments is that the difference between final voiced and voiceless obstruents is orthographically represented in the German writing system. If the experimental task asks participants to read out

written material, this might lead to spelling pronunciation.

Fourakis and Iverson [6] tried to get around this criticism by prompting participants only with non-neutralizing infinitive forms such as *meiden* ‘to avoid’. These words had to be conjugated (i.e. *mied* ‘avoided’) and the conjugated forms included the relevant stop in a neutralizing context. With four speakers, they did not find significant subphonemic differences between neutralized voiced and voiceless segments.

However, as has been rightly pointed out by Port & Crawford [9], the low participant numbers might have prevented the IN effect from reaching significance [9]. In the terms of Frick [7], the study did not provide a sufficiently ‘good effort’ to find an effect in order to allow the conclusion that the null hypothesis is being proved. One of our aims is thus to employ a similar methodology as did Fourakis and Iverson with more speakers and more experimental items.

Moreover, even though a re-analysis of the Fourakis and Iverson data shows significant differences between underlying voiced and voiceless stops [9], these differences could still be due to orthography because with real words. The literate participants of that study evidently had orthographic representations of the experimental stimuli – and these representations can be active during an experiment even if the task does not emphasize orthography (e.g. [11]). We are therefore using pseudowords to which our participants could not have had prior orthographic exposure.

With respect to the perception of incomplete neutralization, previous studies have usually used the experimental stimuli gathered in IN production studies and presented them to listeners in forced choice paradigms (e.g. [9, 10]). Our perception study addresses the methodological concern that listening to words from only a small set of speakers might make it too easy to distinguish

between underlying voiced and voiceless phonemes because listeners have the opportunity to familiarize themselves with idiosyncratic production patterns of particular speakers. We therefore use the spoken responses of all 16 speakers of the production study in the perception study, trying to probe whether the small subphonemic differences between final voiced and voiceless sounds can still be perceived under these conditions.

## 2. EXPERIMENT 1: PRODUCTION

### 2.1. Methodology

In this production task, pseudowords were presented auditorily in plural form and participants had to derive and produce the corresponding singular form. Participants heard sentences such as (1, cf. audio file 1) and had to respond by saying (2, cf. audio file 2).

- (1) Aus Dortmund kamen die Drude. [plural stimulus]  
'From Dortmund came the Drude.'
- (2) Ein Drud wollte nicht mehr. [singular response]  
'One Drud refused to continue.'

The critical phoneme of this example is /d/ which appears in auditory form in a voiced and non-neutralizing context (1), and has to be pronounced in a neutralizing context (2).

There were no time constraints. Before the actual experiment, participants worked through 8 demonstrations and 8 practice stimuli. There were no repetitions.

#### 2.1.1. Stimuli

The experimental items consisted of 24 pseudoword pairs such as (3-5):

- (3) Wiebe vs. Wiepe
- (4) Gaude vs. Gaute
- (5) Gage vs. Gake

There were 8 bilabial, 7 alveolar and 9 velar stimulus pairs, each of which followed one of the vowels /a, o, u, i, au/. In addition to the 48 critical items, there were 96 fillers with sonorants and fricatives instead of stops. The stimuli were randomized into four blocks and corresponding members of a pair were separated by at least one block. All stimuli were spoken by a male native speaker of German (second author).

#### 2.1.2. Stimuli norming & acoustic analyses

In order to assure that the intervocalic voicing distinction was indeed present in our auditory materials, we performed an acoustic analysis with Praat 5.2 [3]. Voice onset time was on average 41ms longer for voiceless stops (analysis by items:  $t_2(23)=17.46$ ,  $p<0.0001$ ); the closure duration was 21ms longer ( $t_2(23)=13.35$ ,  $p<0.0001$ ). The vowel preceding the critical stop was on average 28ms shorter ( $t_2(23)=7.63$ ,  $p<0.0001$ ).

A norming study with 10 German native speakers (5 male / 5 female) confirmed that the voicing contrast of the critical stimuli is very easy to perceive (98% accuracy; one sample t-test by subjects against chance:  $t_1(9)=22.81$ ,  $p<0.0001$ , by items:  $t_2(23)=82.41$ ,  $p<0.0001$ ).

#### 2.1.3. Acoustic analyses

Our experiment includes four dependent measures: (1) The duration of the vowel preceding the critical stop, (2) the closure duration between the end of the vowel and the release, (3) the duration of the burst and (4) the intensity of the burst.

#### 2.1.4. Statistics

All data were analyzed using R with the packages *lme4* [2] and *languageR* [1]. We constructed linear mixed effects models with Subjects and Items as random effects, and Voicing (underlying voiced vs. underlying voiceless) as fixed effect. We checked for normality and homogeneity by visual inspection of plots of residuals against fitted values. Throughout the paper, we present MCMC-estimated p-values that are Dunn-Šidák corrected for 4 tests (on each dependent measure).

#### 2.1.5. Participants

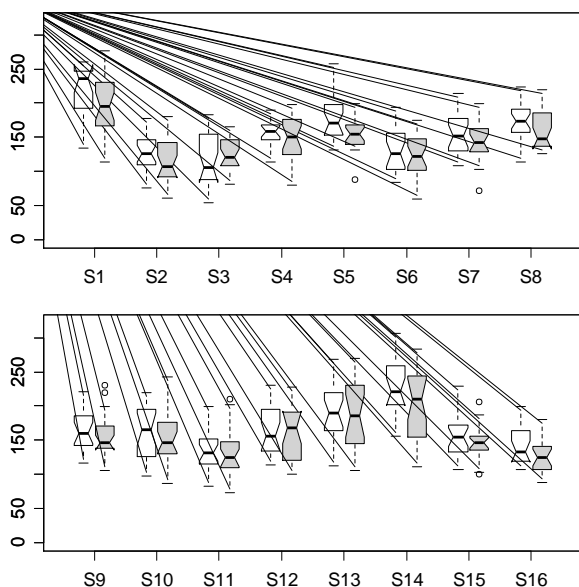
16 speakers participated in the experiment. All participants were native speakers of German without known hearing deficits or speaking impairments (mean age: 25 years; age SD = 3.7; 9 women, 7 men).

## 2.2. Results and discussion

Because we used pseudowords that were necessarily unknown by our participants, we had to exclude many responses (total: ~12%) that were either incorrectly remembered or produced with a lot of hesitation. There were 692 remaining items to be considered for the analysis.

We found a significant effect of Voicing on vowel duration ( $p < 0.001$ ) and burst intensity ( $p < 0.01$ ): vowels were on average 8 milliseconds shorter before underlying voiceless stops and the burst was 1.24 dB louder. We failed to find a difference between underlying voiced and voiceless stops for closure duration ( $p = 0.76$ ) and burst duration ( $p = 0.7$ ). Fig. 1 depicts the difference in vowel durations.

**Figure 1:** Vowel durations for underlying voiced (white) and voiceless (grey) stops in the neutralizing context for all 16 speakers; top row: speakers 1 to 8, bottom row, speakers 9 to 16.



To sum up, we were able to demonstrate an IN effect of vowel duration and burst intensity in an experimental paradigm that diminishes the influence of orthography to a minimum degree via purely auditory presentation and the use of pseudowords.

### 3. EXPERIMENT 2: PERCEPTION

The perception study assesses whether the difference in Experiment 1 is actually perceivable. The previous studies that have claimed that listeners can perceive the voicing distinction in a neutralizing context have used auditory stimuli only from a small set of speakers (e.g. [9]), or even from only a single speaker (e.g. [8]). This gives participants ample opportunity to familiarize themselves with speaker characteristics and this in turn might make it very easy to detect subtle cues for voicing in a neutralizing context. It is thus interesting to ask whether participants perform equally well with a multitude of voices. Therefore,

each listener in this experiment heard the 24 stimuli from all of the 16 different speakers that participated in the production experiment. This, to us, seems to be a design that more accurately mirrors the task of perception in the real world where listeners have to cope with inter-speaker variation.

### 3.1. Methodology

#### 3.1.1. Procedure

Participants heard sentences such as (2) taken from the first experiment and were asked to choose between two orthographic representations (e.g. *Drud* vs. *Drut*) presented on the left and the right side of the screen. Correct and incorrect responses, as well as voiced versus voiceless words were balanced for left and right positions; all items were randomized. Because we expected a response bias towards 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>.

#### 3.1.2. Stimuli

We randomly sampled subsets (192 items) of the items in the production study until we gained a subset in which the effects of the production study were significant (vowel duration,  $p < 0.05$ ; burst intensity,  $p < 0.01$ ).

We constructed 4 lists. In each of the lists, each stimulus pair (e.g. *Wieb* vs. *Wiep*) appeared once. Also, each speaker appeared at least once. Given that there were 24 items but only 16 speakers, 8 speakers appeared twice per block. All items were critical; there were no fillers.

#### 3.1.3. Statistics

We constructed a mixed logit model with the actual response as the dependent variable and the correct response as predictor (random effects: Subject and Items). As additional confirmation, we used one sample t-tests on subjects ( $t_1$ ), items ( $t_2$ ) and speaker voice ( $t_3$ ) with 0.5 as hypothetical mean value. We report two-tailed p-values.

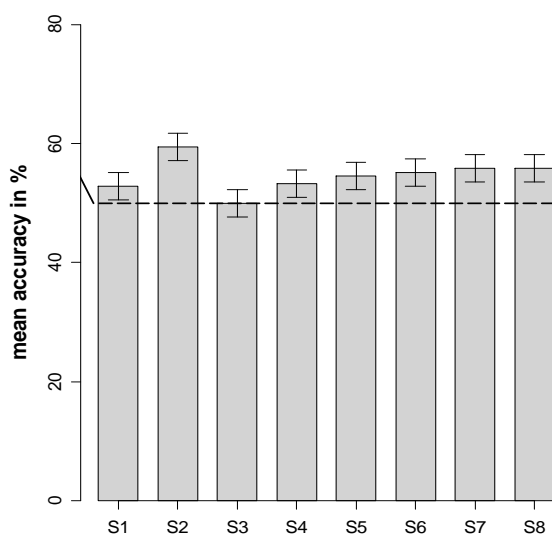
#### 3.1.4. Participants

We tested 8 native speakers of German, none of which participated in the preceding experiment.

### 3.2. Results and discussion

The overall average accuracy was 54% and participants ranged from 50% to 59%. Participants were able to hear the contrast in neutralizing position above chance-level ( $p < 0.001$ ), however, a mixed logit model with accuracy as dependent variable and voicing as predictor variable showed that participants were biased towards choosing the voiceless category ( $p < 0.001$ ). The analyses by subjects, items and speaker voice all reached significance ( $t_1(7) = 4.76$ ,  $p < 0.01$ ;  $t_2(23) = 3.93$ ,  $p < 0.001$ ;  $t_3(15) = 3.5$ ,  $p < 0.01$ ).

**Figure 2:** Mean accuracy values and 95% confidence intervals for the 8 participants; the dashed line indicates chance performance.



The accuracy averages are similar to some previous IN perception studies (e.g. [9]), however, one should ask the question whether a 54% accuracy – albeit significant – is actually indicative of the minor importance IN might play in everyday communication. To us, it seems that accuracy rates barely crossing the chance threshold reflect the minor functional relevance of IN (there are hardly any contexts in which a word-final phonetic difference between voiced and voiceless stops is needed to disambiguate words).

Either way, our results demonstrate that IN effects in perception are robust to the extent that they occur in a forced-choice paradigm even with a multitude of voices.

### 4. CONCLUSIONS

We have demonstrated that incomplete neutralization of German word-final stops occurs even in a completely auditory task that uses

pseudowords instead of real words. It therefore does not seem to be the case that our results are caused by orthography. Moreover, our perception study shows that the minute differences obtained in production can still be perceived if speakers do not have much opportunity to familiarize themselves with particular voices. These experiments therefore address important concerns with previous investigations of incomplete neutralization. Our results indicate that IN is a robust phenomenon whose phonological implications should be taken seriously.

### 5. REFERENCES

- [1] Baayen, R.H. 2009. *LanguageR: Data Sets and Functions with "Analyzing Linguistic Data: A Practical Introduction to Statistics"*. R package version 0.955.
- [2] Bates, D.M., Maechler, M. 2009. *Lme4: Linear Mixed-effects Models Using S4 Classes*. R package version 0.999375-32.
- [3] Boersma, P., Weenink, D. 2009. Praat: Doing phonetics by computer (Version 5.1.05) [Computer program]. Retrieved May 1, 2009, from <http://www.praat.org/>
- [4] Charles-Luce, J. 1985. Word-final devoicing in German: Effects of phonetic and sentential contexts. *Journal of Phonetics* 13, 309-324.
- [5] Ernestus, M., Baayen, H. 2006. The functionality of incomplete neutralization in Dutch: The case of past-tense formation. In Goldstein, L.M., Whalen, D.H., Best, C.T. (eds.), *Laboratory Phonology*. Berlin: deGruyter, 8, 27-49.
- [6] Fourakis, M., Iverson, G.K. 1984. On the 'incomplete neutralization' of German final obstruents. *Phonetica* 41, 140-149.
- [7] Frick, R.W. 1995. Accepting the null hypothesis. *Memory & Cognition* 23, 132-138.
- [8] 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.
- [9] Port, R., Crawford, P. 1989. Incomplete neutralization and pragmatics in German. *Journal of Phonetics* 17, 257-282.
- [10] Port, R., O'Dell, M.L. 1985. Neutralization of syllable-final voicing in German. *Journal of Phonetics* 13, 455-471.
- [11] Seidenberg, M.S., Tanenhaus, M.K. 1979. Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory* 5(6), 546-554.