Paper in press at *Language, Cognition, and Neuroscience*
http://dx.doi.org/10.1080/23273798.2017.1286361

Surface Forms Trump Underlying Representations in Functional Generalizations in Speech Perception: The Case of German Devoiced Stops

Holger Mitterer — University of Malta and University of Tübingen

Eva Reinisch — Ludwig Maximilian University of Munich

Listeners can adapt their perceptual categories for speech sounds in response to speakers' unusual pronunciations. The present study tested whether this generalization is sensitive to surface or underlying properties of speech by exploiting the devoicing of voiced stops in German. This allows us to assess whether learning on phonetically voiceless stops that are underlyingly voiced generalizes to stops that share the same surface form (i.e., voiceless) or the same underlying representation (i.e., voiced). Our results showed only minimal generalization: learning for (surface) voiceless stops in offset position that are underlyingly voiced generalizes to surface and underlying voiceless stops in the same position but neither to voiced nor voiceless stops in intervocalic position. This suggests that listeners extract segments of sufficient acoustic similarity from the input and use them for generalization of learning in speech perception. The units of perception thereby appear context-sensitive rather than abstract phonemes or phonological/articulatory features.

Listeners flexibly adapt to speakers' idiosyncratic pronunciation variants by using lexical context to adjust category boundaries (Norris, McQueen, & Cutler, 2003). That is, if listeners repeatedly experience an acoustically ambiguous sound, for example, between /s/ and /f/ in words where it replaces /s/ (e.g., *police* where *poli[f]* is not an English word) listeners then tend to perceive such ambiguous sounds in line with the previously experienced context even in cases of lexical ambiguity (i.e., they perceive forms such as [nai$^{s}/_{f}$] as *nice* rather than *knife*). This process has come to be known as perceptual learning in speech or phonetic recalibration.

This finding has also fuelled the debate about the computational architecture of spoken-word recognition. Partly due to the failure of the literature to converge on what the sublexical units in spoken-word recognition are (Goldinger & Azuma, 2003; Remez, 1987), it has even been proposed that listeners may in fact use no sublexical units at all (Goldinger, 1998; Pierrehumbert, 2002). However, studies on perceptual learning indicated that listeners make use of sublexical units and that such sublexical units may be useful to allow for faster adaptation to an unfamiliar accent (Mitterer & McQueen, 2009; Reinisch & Holt, 2014). The perceptual-learning paradigm has contributed to the consensus that abstraction is an important part of spoken-word recognition (Goldinger, 2007), partly because it has been shown by computational modelling that strictly episodic models are not able to account for such findings (Cutler, Eisner, McQueen, & Norris, 2010).

This research was conducted while the second author was supported a grant from the Alexander von Humboldt Foundation and a Bavarian Equal Opportunities Sponsorship. The data for Experiment 2 were gathered while the first author was a visiting professor at the University of Tübingen. We would like to thank Anna-Maria Meck and Elaine Stix for help with collecting data.

That the abstract units involved in perceptual learning are necessarily the units involved in spoken-word recognition is reinforced by findings that learning on one set of words generalizes to other words (McQueen, Cutler, & Norris, 2006; Mitterer, Chen, & Zhou, 2011; note that words used for exposure and test differed in basically all perceptual learning studies). Additional evidence for the involvement of these units in spoken-word recognition comes from the paradigm itself. Firstly, for participants the test is a completely unrelated task with regard to exposure. Second, given that after the exposure phase, participants are asked to read new instructions for the test, at least several minutes pass before the learned categories could be applied during the test. Eisner and McQueen (2006) showed that learning transfers even across a gap of 12h. It is hence necessary to assume that the representations affected by perceptual learning are used for spoken-word recognition. Additionally, Mitterer and Reinisch (2013) demonstrated with a visual-world eye-tracking paradigm that learning that a given speaker produces /s/ or /f/ as an ambiguous [$^{s}/_{f}$], influences fixations on possible lexical referents as early as the content of the phonetic signal does. This indicates that perceptual learning influences *pre-lexical* representations of segmental categories.

As for the grain-size of these pre-lexical units, support has been found for various types of units. Some of the evidence could be construed as supporting the idea of the classical phoneme, which is context- and position-independent. Jesse and McQueen (2011) found that perceptual learning for /s/ and /f/ generalizes from offset position (learning on items such as [pəli$^{s}/_{f}$], based on *police*) to onset position (i.e., to minimal pairs such as *sin-fin*), which shows, for this case, position-independence. However, as noted by these authors, the segments /s/ and /f/ are not ideal to argue for position-independent abstract phonemes, because the acoustic implementation of /s/ and /f/ is largely position independent in Dutch (the language actually used in this experiment).

Building on these findings, Mitterer, Scharenborg, and McQueen (2013) tested whether learning also occurs when there are well-described allophonic differences between different versions of the same phoneme. They presented listeners with ambiguous segments between a dark /l/ and an approximant /r/, here transcribed as [$^{ɫ}/_{r}$], in lexically unambiguous contexts during an exposure phase (e.g., [wɪntə$^{ɫ}/_{l}$], which can only be the Dutch word for *winter* since *wintel* is a nonword in Dutch). They then tested whether this form of exposure influenced the perceptual boundary between different allophonic versions of the phonemes /l/ and /r/. At test, participants categorized three different speech-sound continua from different versions of /l/ to different versions of /r/. One continuum used the same allophones used during exposure ([ɫ] and [ɹ], in the nonwords *kwipter* and *kwiptel*), one used an alveolar trill as the implementation of the /r/ in the same nonwords, and one continuum used /l/ and /r/ in onset position, then implemented as light /l/ and an alveolar trill /r/ in the nonwords *repaas* and *lepaas*. The results showed a strong learning effect for the continuum using the same allophones as presented during exposure but no generalization to the other two continua. Mitterer et al. (2013) hence concluded that listeners make use of abstraction during speech perception, which allows generalization from the exposure to the test (non)words, but that this abstraction does not make use of position- and context-independent phonemes.

Further research along these lines confirmed that learning is quite context-specific, suggesting that the sub-lexical units in spoken-word recognition are rather different from classical phonemes. Reinisch, Wozny, Mitterer, and Holt (2014) tested whether generalization is possible for the same phoneme when the cues strongly differ. They first established via cuetrading experiments that the difference between /b/ and /d/ in American English is dominantly carried by the stop-release burst in a high-front vowel context ([idi]-[ibi]) but mainly by formant transitions in a low vowel context ([aba]-[ada]). They found that perceptual learning occurs in both contexts, but does not generalize to the other context, again indicating that phonemes may not be involved in functional reorganization of speech perception.

Reinisch et al. (2014) also tested featural accounts of pre-lexical abstraction. In linguistics, the prevailing phonological theories do not assume that the speech signal is decomposed into letter-sized segments such as phonemes or allophones but rather decomposed into phonological or articulatory features (Embick & Poeppel, 2014; Goldstein & Fowler, 2003; Lahiri & Reetz, 2010). Kraljic and Samuel (2006) presented a data set that would support such an account. They found that perceptual learning for stop voicing in American English in word-medial position can generalize from alveolar stops (i.e., /t/ versus /d/) to labial stops (i.e., /p/ versus /b/). However, as for the finding of the position-independence of the learning for /f/ versus /s/ (Jesse & McQueen, 2011), this contrast is not ideal to argue for abstract phonological features, since the acoustic implementation of the voicing feature is identical in the learning and generalization condition. Reinisch et al. (2014) therefore further tested generalization for place of articulation over a difference in manner of articulation, which leads to stronger acoustic differences between baseline and generalization conditions. They exposed listeners to labial versus alveolar stops and tested learning on stops and nasals (and vice versa). A feature account predicts that listeners in both cases do not learn about the segments /b/ or /m/ versus /d/ or /n/ but learn about the distinction between the place features [LABIAL] and [CORONAL]. This would predict that learning should generalize from [aba] versus [ada] to [ama] versus [ana], because both pairs of stimuli only differ in the place feature, with the levels labial and coronal. This prediction was not borne out, however; when exposed to stops, learning occurred only for stops but not for nasals and vice versa. Importantly this finding has been replicated with two different types of learning contexts. While Reinisch et al. (2014) had used an audiovisual adaptation phase (following Bertelson, Vroomen, & de Gelder, 2003) Reinisch and Mitterer (2016) confirmed the lack of generalization across manner of articulation with the typical lexically-guided perceptual learning paradigm discussed above (i.e., following Norris et al. 2003). Since the lexical paradigm provides the listeners variable contexts during exposure (i.e., here: 20 different words as compared to one exposure token for the audiovisual paradigm) if anything generalization should have been facilitated here. However, again, learning appeared to be specific to the exposure contrast and the recalibrated place of articulation contrast in stops did not generalize to the same place contrast in nasals.

The results of Reinisch and colleagues (Reinisch et al. 2014, Reinisch & Mitterer 2016) indicate that generalization of learning is difficult to obtain and that learning is quite specific to the phonetic details of the exposure stimuli (for comparable findings highlighting the importance of acoustic similarity regarding generalization of learning across speakers, see Eisner & McQueen, 2005; Kraljic & Samuel, 2007; Reinisch & Holt, 2014). The repeated failure to find generalization of learning across similar phonological specifications raises the question whether maybe the “opposite” would hold, namely that *differences* in phonological specification are sufficiently powerful to prevent generalization of learning despite phonetic similarity. Although different phonological representations mostly lead to phonetic differences in the surface forms, phonological neutralizations provide the ideal testbed to address this question. Mitterer, Cho, and Kim (2016b) made use of the rule of phonological tensification in Korean to test the role of underlying representations in perceptual learning without confounding differences in surface form. Korean has a three-way voicing distinction in which lax stops (e.g., /k/) contrast with tense stops (e.g., /k*/) and aspirated stops (e.g., /kʰ/). Tensification is a phonological process which turns lax stops into tense stops if the preceding word ends in an obstruent and there is no strong prosodic boundary between the two words (e.g., the combination of /tʃuŋku**k**/ *Chinese* and /**p**atʃi/ *pants* is produced as [tʃuŋku**kp***atʃi], with the tense stop [p*] instead of the underlying lax stop /p/). In contrast to many other phonological processes (e.g., place assimilation in English), this process is phonetically complete across small prosodic boundaries, so that a derived tensified stop is phonetically equivalent to an underlying tense stop (Jun, 1998). Mitterer et al. (2016b) presented one group of listeners with ambiguous stops between a labial and alveolar tense stop in a context that allowed listeners to infer the identity of the ambiguous stop based on their lexical knowledge (e.g., [tʃuŋkuk{^{p*}/$_{t*}$}atʃi], where the ambiguous sound (transcribed as {^{p*}/$_{t*}$}) could only be interpreted as /p/, since /patʃi/ means *pants* while /tatʃi/ is a nonword). Another group heard the ambiguous stops in positions where they would likely be interpreted as *alveolar* (e.g. [tʃuŋkuk{^{p*}/$_{t*}$}oma], where /toma/ means *cutting board* while /poma/ is a nonword in Korean). At test, all participants categorized different segments as labial or alveolar. In the baseline condition, these were underlying lax stops in tensifying position, the same context as during exposure. As expected this gave rise to perceptual learning. To test a potential role of underlying representations, Mitterer et al. (2016b) tested whether learning also generalizes to underlyingly tense stops, which differ in the phonological representation but not in their surface form. Despite this difference in phonological representation generalization could be found.

While this finding indicates that a difference in underlying representations does not prevent learning, the other generalization conditions tested by Mitterer et al. (2016b) suggested that sharing an underlying representation may nevertheless foster generalization. Three generalization conditions were tested: lax stops that share the underlying representation and are acoustically very similar, aspirated stops that share neither underlying or surface form but are somewhat similar, and nasals that are dissimilar in all respects. Generalization was only found for lax stops. These results could be explained by two accounts. Either sharing the same underlying abstract representation fosters generalization of learning, or that learning can only generalize to phonetically highly similar tokens.

Interestingly, there is a parallel proposal that would be in line with a role for more abstract representations in speech perception. Bowers, Kazanina, and Andermane (2016) argued that the perceptual-learning might not be ideal to reveal all prelexical units in spoken-word recognition, because, though learning clearly seems to operate on position-specific allophones, this does not rule out that phoneme-type representations are still involved in spoken-word recognition. They make an interesting analogy with visual-word recognition: abstract letter codes are important in visual-word recognition, but learning to recognize an odd shape as a small letter *a* should not influence how a capital *A* is recognized. The analogy then is that though learning may be on more detailed representations, a more abstract phonemic representation may still be activated in spoken-word recognition at a later stage. Notably, the more abstract units proposed by Bowers et al. (2016) may support the notion that sharing an underlying representation may foster generalization of perceptual learning as shown in Mitterer et al. (2016b).

It is hence of theoretical importance to clarify the role of underlying representations in perceptual learning. This is the purpose of the current paper. We tested whether sharing an underlying representation can foster generalization even when the differences in acoustic realization are stronger than in the Korean tensification case. We identified a case where two segments share an underlying representation but differ more strongly in their acoustic realization than the quite similar tense and lax stops in Korean. If learning is mediated by underlying representations, we should find that learning generalizes despite the larger phonetic difference.

Such a larger phonetic distance between two versions of the same underlying segment is caused by final obstruent devoicing in German. Final obstruent devoicing is a process that occurs also in many other, relatively unrelated languages such as Russian and Maltese. Due to devoicing, underlyingly voiced stops are produced as voiceless in word-final position. It is important to note, however, that the knowledge about the underlying voicing of a final stop is not simply meta-linguistic knowledge, possibly based on orthography, but required in order to be a competent speaker of German, independent of the use of written language. When phonetically unvoiced final stops become word-medial through morphological alterations, their underlying voicing shows, so that *Bord* [bɔɐt] (Engl., *board* [nautical]) if turned into a verb becomes *borden* [bɔɐdn̩] (Engl. *to board*) while *Wert* [wɛɐt] (Engl., *value*) becomes *werten* [wɛɐtn̩] (Engl., *to evaluate*).

Even though this devoicing is phonetically incomplete in German (Roettger, Winter, Grawunder, Kirby, & Grice, 2014), a given word-final stop is quite unlikely to be identified correctly as underlying voiced or voiceless, with accuracies hovering just above chance. That is, even though there is incomplete neutralization for a range of tokens, it is not uncommon that a given token of a devoiced stop would be a good example of an underlyingly voiceless stop. Stated otherwise, the effect size of incomplete neutralization is rather small (cf. Roettger et al., 2014.). This raises the question whether such small differences are sufficient for listeners to make use of different units for voiceless versus devoiced stops. If this were the case, learning on devoiced stops should not generalize to true voiceless stops in word-final position.

Maybe even more important for the current purposes, the cues for place of articulation in underlying voiced stops can differ dramatically between different instances, for example, depending on their position in a word and hence phonological context. Figure 1 shows examples of German /d/ and /g/ in word-final position (upper row) and in word-medial position, with a following /ən/ syllable that is typically produced as syllabic [n̩] (lower row). While there is some overlap in the preceding cues (with a smaller F2-F3 distance for /g/ than /d/), the main cues differ strongly. When /d/ and /g/ surface as voiceless [t] and [k] the clearest cues are the spectral characteristics of the release burst. With following nasals, there is no release burst. Instead, the place of articulation is cued by progressive assimilation of the nasal, so that the /n/ in *borgen* (Engl, *to lend*) is produced as [ŋ] and the /n/ in *borden* (Engl, *to board*) is produced as [n].

The case of German final obstruent devoicing hence allows us to disentangle the two possible explanations why generalization was found from tensified to lax stops in Korean (Mitterer et al., 2016b). If underlying representations matter, we should find generalization from devoiced stops in final-position to voiced stops in medial position, despite the strong acoustic differences in the realization of the underlying /d/ and /g/. If, however, phonetic similarity matters, we should not find generalization from learning on devoiced stops to underlyingly voiced stops, but generalization should be found to the unvoiced stops in final position (underlyingly voiced and voiceless), and maybe to voiceless stops in medial position, which also contain a release burst.

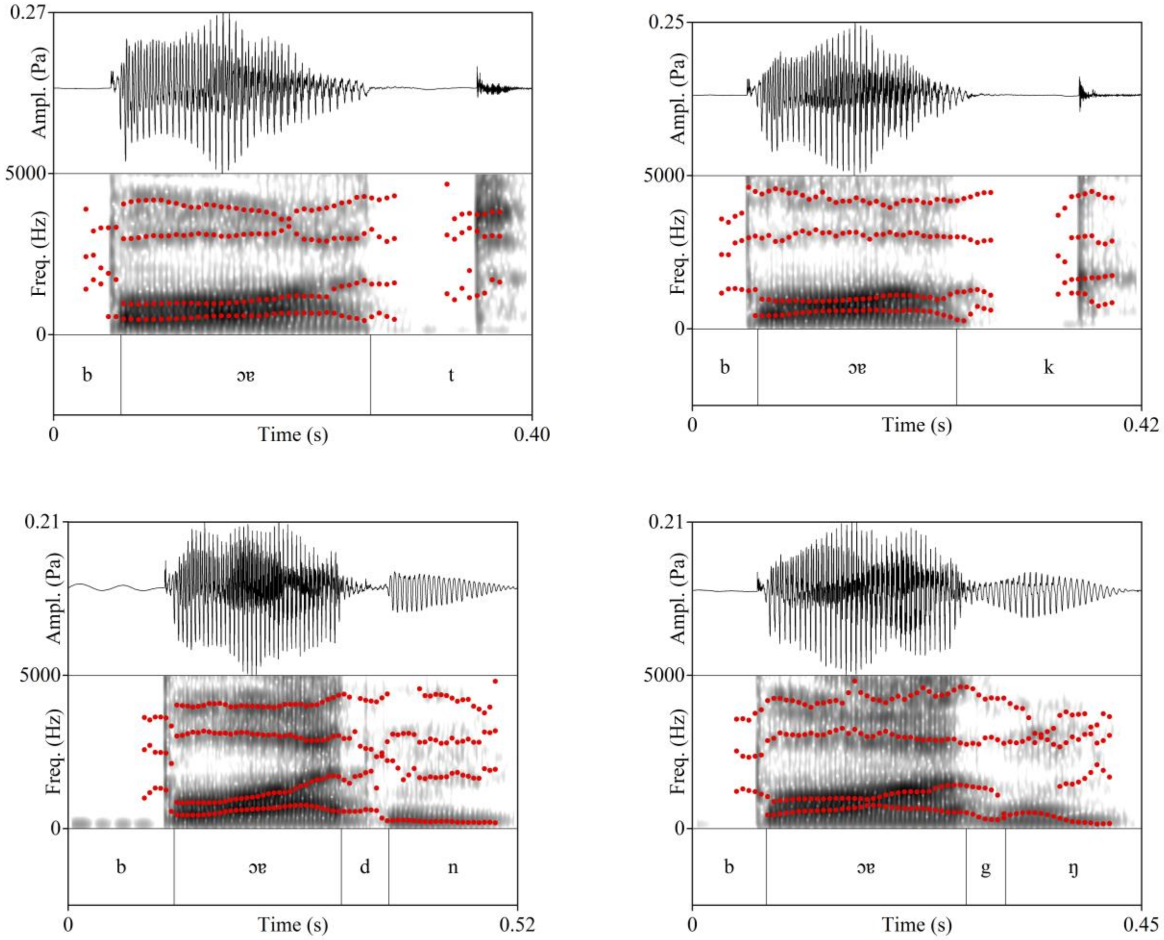


Figure 1: Oscillogram and spectrogram of the words *Bord, Borg, borden, and borgen* in German (Engl., *board* [nautical], *lend!* [imperative of lend], *to board*, *to lend*, respectively). Note that our segmentation did not attempt to draw a boundary between /o/ and /r/, which are produced as [ɔ] and [ɐ], respectively. Drawing a boundary between these two segments is quite difficult and the boundary partly arbitrary.

Experiment 1

This experiment set out to further test to what extent phonetic similarity and underlying representations constrain generalization of perceptual learning. The setup was similar to previous studies on lexically-guided perceptual learning following the paradigm first introduced by Norris et al. (2003). During exposure, participants heard ambiguous stops between [t] and [k]. Half of the participants heard these ambiguous stops in words ending on [k] and the other half in words ending on [t]. The respective other sounds were heard in their unambiguous form. Based on previous studies with this paradigm, we expect participants who heard words ending in [k] with an ambiguous stop to give more [k] responses during a phonetic categorization task at test than those participants who heard words ending in [t] with ambiguous stops.

The critical question is to what extend learning generalizes. During exposure, participants heard words ending in [t] and [k] that arise from final devoicing, so that the underlying stops are /d/ and /g/. Following this exposure to devoiced stops, listeners were tested on their categorization of four minimal pair continua involving the following conditions:

1) <u>devoiced final:</u> *Bord-borg* (board [nautical] or board [short for snowboard] – lend! [imperative of *borgen*]; also Science Fiction characters "Borg" from the Star Trek franchise), where phonological specification and phonetic realization matched the word-final devoicing context experienced during exposure.
2) <u>voiceless final:</u> *Wert-Werk* (value - work), where phonetic surface form but not phonological specification matched exposure.
3) <u>voiced medial:</u> *borden-borgen* (to board [a ship or airplane] – to lend), where phonological specification matched but the phonetic surface form strongly differed.
4) <u>voiceless medial:</u> *Werte-Werke* ([plurals of *Wert-Werk*]), with a similar phonetic surface form (released voiceless stop) but a different phonological specification in comparison to the exposure.

It is important to note that the word-medial voiceless tokens are less like the exposure tokens than voiceless final tokens. Importantly, the medial stops carry additional cues for place of articulation in the transitions out of the stop into the final schwa. For major place distinctions, these formant transitions into the following vowel are usually highly informative (Steriade, 2001).

If perceptual learning was sensitive to phonological properties of the sound contrast, then a learning effect should be evident in the two voiced conditions: the devoiced final because it fully matches the exposure condition, and the voiced medial because it matches the phonological specification of voicing. If, however, perceptual learning was specific to the phonetic realization of the sound contrast during exposure, then a learning effect should be found in both word-final pairs. Finally, if there was some generalization based on phonetic similarity, we might also find learning in the voiceless-medial condition.

Note that is experiment also provides another test of generalization across features. Previous experiments have mostly tested generalization of learning regarding place of articulation across manner of articulation (Reinisch & Mitterer, 2016; Reinisch et al., 2014). Manner of articulation is a stable feature—that is, it hardly ever varies over contexts—and is highly relevant for word recognition (Ernestus & Mak, 2004). A difference in manner as a highly relevant feature might hence block generalization. In the current experiment, we tested generalization across voicing, a feature that is not a phonological stable feature and is of lower relevance for word recognition (Ernestus & Mak, 2004). It may hence be more likely to allow for generalization of learning for place of articulation.

Method

Participants

Forty-four native speakers of German (15 males), students at the University of Munich took part for pay. They were aged between 18 and 30 and reported to have no speech or hearing problems.

Materials

100 German words and 100 phonotactically legal nonwords were selected for the lexical decision task that served as exposure (following the paradigm first used by Norris et al., 2003). 20 of the words ended in /d/, 20 ended in /g/. These were the critical words and no other instances of /d/, /t/, /g/, or /k/ occurred in the exposure set. Four minimal word pairs were selected for phonetic categorization at test (see examples above). Two pairs differed in the /d/-/g/ contrast, two in /t/-/k/. Each contrast once occurred word-finally (where /d/-/g/ would surface devoiced as during exposure) and once word-medially (see list of test conditions described above).

All stimuli were recorded spoken by a female native speaker of German. Critical words were recorded with the correct stop and the other critical stop that formed a nonword (e.g., *Fahrra[t]*, Engl., *bike*, was also recorded as *Fahrra[k]*). The speaker was asked to produce all words at a similar speaking rate and with a similar intonation contour (flat to slightly falling). Each word was recorded multiple times such that pairs could be closely matched.

Critical words and minimal pairs were then morphed in 21-step (from [0% word1, 100% word2] to [100% word1, 0% word2] in 5% steps) continua using STRAIGHT (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999). Time alignment ensured that only same types of segments were morphed (i.e., stops with stops, etc.). With the morphing technique, not only the critical stops' bursts but also their formant transitions and any other potential cues were morphed (for a more thorough description of the morphing procedure see e.g., Reinisch, Weber, & Mitterer, 2013). Three phoneticians selected the most ambiguous steps for exposure stimuli and the steps to be used as midpoints for the test continua.

Test continua consisted of five steps around the perceived midpoint of the continua as established in the norming procedure. The cross-over points for the perception of an alveolar vs. velar sound was at the 60% alveolar[1] morph for *Bord-Borg* (Engl., *board-lend*), 45% for *borden-borgen* (Engl., *to board- to lend*), 55% for *Wert-Werk* (Engl., *value-work*), and 60% for *Werte-Werke* (Engl., *values-works*). This most ambiguous point was used as the midpoint of the test continua (= Step 3 in Figures 2 and 3 below), steps two and four were the morphs differing by 10% in either direction, and steps one and five differed by 30% from the midpoint. That is, we used endpoints that should be reasonably good anchors for a phonetic-identification task. By using a continuum rather than only one stimulus during test, we motivate our participants to engage in phonetic identification (rather than adapt a certain strategy when answering to only

[1] We indicate morph ratios only with the percentage of the alveolar part, the amount of velar signal follows from that, so that a morph that is using the alveolar signal for 60% uses the velar signal to 40%.

one ambiguous sound), and the results reveal whether participants were then still influenced by the phonetic detail.

Procedure

As in previous similar experiments, the experimental session consisted of an exposure and a test phase. During exposure, half of the participants were randomly assigned to a /d/-bias condition, and half to a /g/-bias condition. Both groups of participants were presented the same 100 words and nonwords except for the 40 critical items (see the Appendix) in which, depending on group, /d/- or /g/-words were replaced by the ambiguous morphs. The /d/-bias group heard the /d/-final words with stops that were ambiguous between [t] and [k], and the /g/-final words with a clear [k]. The /g/-bias group heard the /g/-final words with an ambiguous stop and the /d/-final words with a clear [t]. Exposure to the clear stops (clear [t] for the /g/-bias group and clear [k] for the /d/-bias groups) is necessary to indicate that the speaker does not simply neutralize the place contrast in word-final position. While it may be argued that these “contrast items” lead to selective adaptation or some form of perceptual contrast, this explanation has been ruled out by the finding that no changes in perception occur if the ambiguous and contrast stimuli are presented in meaningless contexts (Norris et al., 2003).

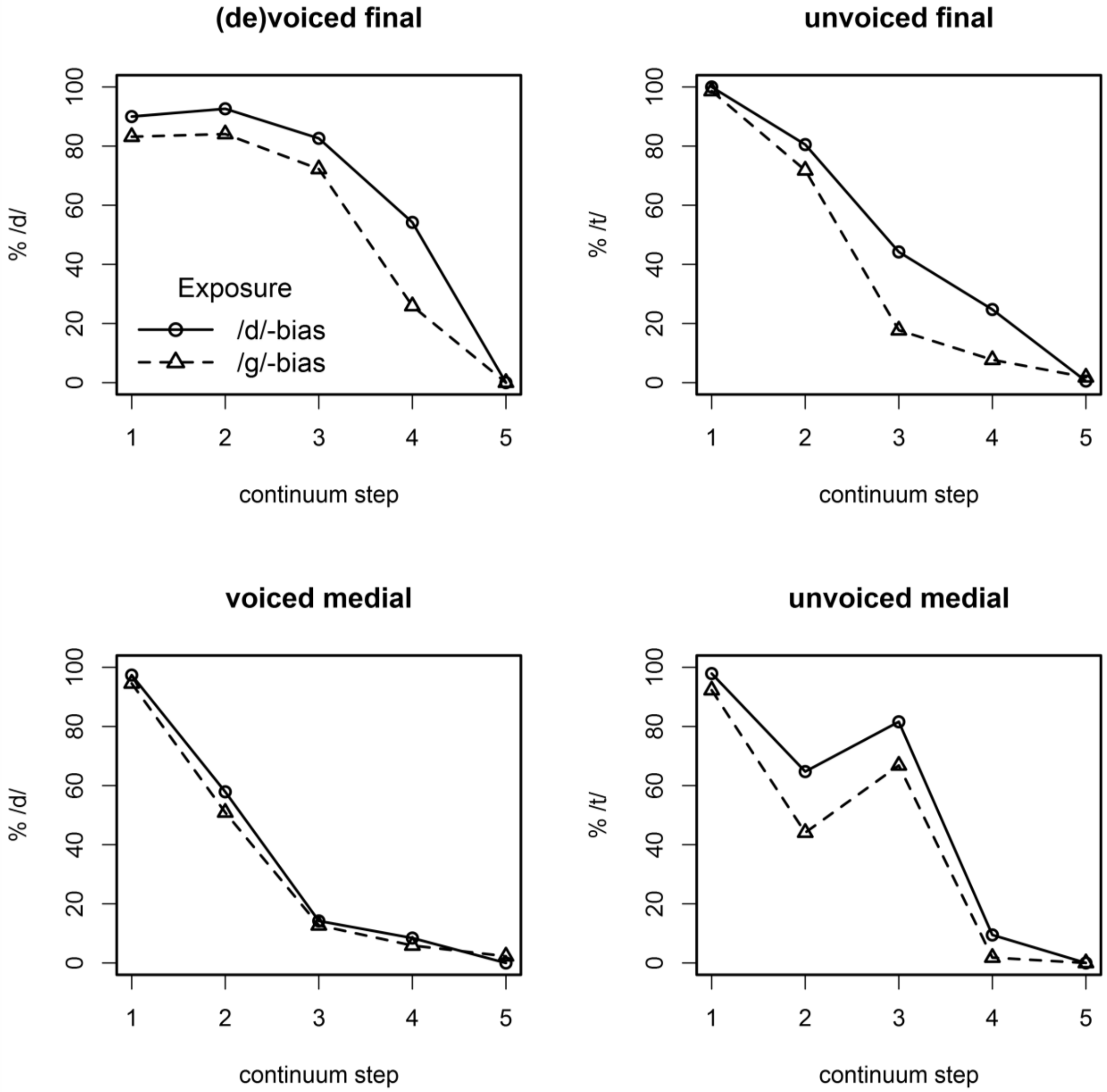
Participants were seated in a sound-proof booth and listened to the stimuli over headphones. Each stimulus was presented once, so that the lexical decision task consisted of 200 trials and lasted about 10-15 minutes. The task was to decide on every trial whether the presented stimulus formed a word or nonword (in German) by pressing the 1 or 0 key on the computer keyboard. Key labels and response options were shown on a computer screen. Stimuli were presented in random order. Every 50 trials participants were allowed a self-paced break.

The test phase followed immediately. After reading written instructions, all participants performed the same phonetic categorization task with the four minimal pair continua (control: *Bord-borg,* Engl., *board–lend*, generalization: *Wert-Werk, borden-borgen, Werte-Werke*, Engl.,*value-work, to board-to lend, values-works*, respectively). On each trial, participants were first presented the upcoming pair written on the screen with the word containing /d/ or /t/ on the left. As is typical for German, phonological voicing was coded orthographically in these words. Half a second later, the auditory stimulus was played over headphones. Participants had to indicate by button press which of the words they heard. Minimal pairs were presented intermixed in random order with the restriction that all words and all continuum steps were presented before they were repeated. Participants responded to 10 repetitions per word per step for a total of 200 trials. Every 50 trials they were allowed a break.

Results

Exposure

Three participants in the /d/-bias condition rejected more than 50% of the critical words during exposure and were therefore excluded from all further analyses. This is because previous studies showed that at least 10 critical items have to be experienced in order for perceptual learning to occur (Poellmann, McQueen, & Mitterer, 2011) and if the items are perceived as nonwords, they are unlikely to trigger recalibration (Norris et al., 2003; Sjerps & Reinisch, 2015). In the remaining set of participants, 95% of the critical words were accepted as the intended words.



Test

Figure 2: Percentage of word responses containing the alveolar consonant in the phonetic categorization task at test in Experiment 1, depending on continuum step (x-axis) and exposure condition (different lines). A learning effect would be reflected in the identification function of the /d/-bias group being “above” the identification function of the /g/-bias group.

Figure 2 shows the proportion of responses, in which participants selected the word with the alveolar stop (i.e., /d/ or /t/ rather than /g/ or /k/). Although the continuum endpoints were clearly identified as the intended sounds, the categorization function for the voiceless medial condition (*Werte-Werke*) was non-continuous. Similar discontinuities have been found previously with morphed stop continua. Importantly, however, this pattern emerged for both exposure groups suggesting that the shape of the categorization function is not affecting our critical results. As shown in Figure 2, for all but the voiced medial condition participants in the /d/-bias group gave more responses favouring the alveolar stop than the /g/-bias group.

Statistical analyses were carried out using generalized linear mixed-effects models as implemented in the lme4 (v.1.1.-10) package in R. The model was fitted with response as the dependent variable (the word containing the alveolar stop coded as 1, the velar as 0) for which a logistic linking function was used. Fixed factors were Exposure Group (/d/-bias coded as -0.5, /g/-bias as 0.5), Sound Position (word-medial coded as -0.5, final as 0.5), underlying Voicing (voiceless coded as -0.5, voiced as 0.5) and their interactions. Continuum Step (centred on 0, re-scaled to range from -0.5 to 0.5) was also entered but was not allowed to interact with the other fixed factors. This was because Step was used as a control factor here, and allowing interactions with other factor resulted in convergence problems with the generalized linear mixed-effect models. Participant was entered as a random factor with uncorrelated random slopes for all within-participant fixed factors (i.e., all but Group since this factor was manipulated between participants). Table 1 shows the results.

Table 1: Results of the overall analysis of the data in Experiment 1.

Predictor	*b*	*z*	*p*
Intercept	-0.47	-2.13	<.05
Group	-1.09	-2.46	<.014
Position	0.11	0.39	.69
Voicing	1.66	5.26	<.001
Step	-5.22	-24.2	<.001
Group*Voicing	0.52	0.91	.36
Group*Position	-0.70	-1.11	.27
Voicing*Position	3.59	5.50	<.001
Group*Voicing*Position	-1.41	-1.09	.28

Table 2: Results of the analyses per condition.

	Test continuum							
	devoiced final		voiceless final		voice medial		voiceless medial	
Predictor	*b*	*p*	*b*	*p*	*b*	*p*	*b*	*p*
Intercept	4.45	<.001	-1.19	<.01	-2.71	<.001	-0.11	.62
Group	-1.54	<.01	-1.83	<.05	-0.63	.17	-0.42	.17
Step	-10.37	<.001	-6.91	<.001	-6.01	<.001	-3.87	<.001

Critically, in addition to effects of Step, Voicing, and an interaction between Voicing and Position, there was a significant effect of Exposure Group, confirming that participants in the /d/-bias group gave more alveolar responses than listeners in the /g/-bias group. In contrast to what Figure 2 suggests, however, there was no interaction between Group and either Voicing or Position or a three-way interaction between these factors. To follow up on this discrepancy, linear mixed-effects models were run separately for each of the four conditions shown in Figure 1. Table 2 summarizes the results. Only the two conditions with the critical contrast in word-final position showed significant effects of Group (in addition to an effect of continuum Step).

Discussion

The present study tested the role of phonological voicing and phonetic realization of a German stop contrast for the generalization of perceptual learning. We trained listeners to adjust their category boundaries for a place of articulation contrast in German stops that were phonologically voiced (as evident in related forms such as the plural) but realized as phonetically voiceless due to their word-final position. Robust learning was found at test for the minimal pair in which the stop contrast fully matched the exposure condition (i.e., devoiced final *Bord-borg,* Engl., *board-lend*) and this generalized to the pair that matched in phonetic realization and word position but not phonological voicing (i.e., voiceless final, *Wert-Werk,* Engl., *value-work*). This replicates the finding of Mitterer et al. (2016b) that a difference in underlying voicing does not preclude generalization of perceptual learning. It also extends this finding to a case where, on a sample level, the neutralization is incomplete. While the Korean tensification rule is usually considered to be complete when there is only a prosodic word boundary between the word carrying the tensifying context and the tensified word (Jun, 1998), final devoicing in German has been found to be phonetically incomplete (Roettger et al., 2014). Such phonetic differences—which only become apparent if a large sample of tokens is considered—apparently do not influence how a given, individual stop is perceived.

This experiment, however, does not provide a clear answer whether learning generalizes to other contexts. The situation is quite unclear for the two continua with the stops in word-medial position. On the one hand, a significant overall learning effect that did not interact over conditions might be taken as evidence that learning generalizes in these cases. On the other hand, the absence of a significant learning effect for these conditions if tested by themselves indicates that it would be premature to assume that learning generalizes to these conditions. Therefore, Experiment 2 aimed at clarifying whether perceptual learning generalizes to these cases, by slightly improving the exposure conditions and adding more power to the test. To achieve that, we used only the two test continua for which no learning was found in Experiment 1 and increased the number of trials for these continua. Moreover, learning effects tend to dissipate during the test phase (Mitterer & de Ruiter, 2008; Reinisch & Mitterer, 2016), and by using only two continua, we sample more trials for these continua briefly after exposure. Additionally, we changed the exposure stimulus for the subset of exposure items that had low acceptance rates so that these words were more likely to be accepted as real words. That is, for a /d/-final word that was often rejected as a word when in its ambiguous form, we used a morph that was slightly more /d/-like to boost the acceptance rate for this item. Sjerps and Reinisch (2015) showed that higher acceptance rates lead to stronger perceptual learning. These changes from Experiment 1 to Experiment 2 should hence increase the chances to find a learning (or generalization) effect. The descriptive data in Experiment 1 suggest that that we might find a learning effect for the phonetically more similar condition (unvoiced medial) but are unlikely to find a learning effect for the phonetically dissimilar condition (voiced medial).

Experiment 2

Method

Participants

Forty-eight native speakers of German (6 male), students at the University of Tübingen took part for pay. They were aged between 19 and 30 and did not report any speech or hearing problems.

Materials

The same materials were used in Experiment 1 with the following changes. For seven exposure words with acceptance rates below 70% in the ambiguous form, we chose a member of the continuum 5% closer to the lexically correct underlying form to be used as the ambiguous sound during exposure. For instance, in Experiment 1 the item *Sarg* (Engl., coffin) had been presented as the morph with 45% /d/ and was accepted as an existing word in only 68.1% of the cases. Therefore, in Experiment, we now presented a mix of 40% /d/ to make it sound more word-like. Such changes were made for additional six /d/-final words.

In Experiment 2, only two of the four test continua used in Experiment 1 were used. These were the word-medial voiced (*borden-borgen*, *Engl., to board-to lend*) and voiceless (*Werte-Werke,* Engl., *values-works*) continua, which provided no clear learning effect in Experiment 1. By focusing on two continua, the number of test trials for the critical conditions (see Procedure for details) could be increased. These test continua were also slightly adjusted based on the results of Experiment 1. This was to allow for a more balanced perception of the continua. For *borden-borgen*, we used a continuum ranging from a mix using 30% *borden* (and hence 70% *borgen*) to a mix containing 70% *borden* in equal steps of 10%. For *Werte-Werke*, we used a continuum from a mix using 35% *Werte* to 75% *Werte*, again in steps of 10%. By using continua that are better centered around the point of maximal ambiguity, we increased the chances of finding a learning effect.

Procedure

The procedure was the same as in Experiment 1 except that only the two test continua with medial stops were used: *Werte-Werke* (Engl., *values-works*) *and borden–borgen* (*Engl., to board-to lend*). As in Experiment 1, half of the participants were randomly assigned to a /d/-bias exposure condition, half to a /g/-bias exposure condition. Given the smaller number of tokens to be categorized in the test condition, we increased the number of repetitions per token to 15, so that the test phase consisted of 150 trials. The larger number of trials per continuum step should allow a better estimation of the individual categorization function and hence facilitate finding possible generalization effects. Participants again had a chance to take a break after every 50 trials.

Results

Exposure

Four participants in the /d/-bias condition rejected more than 50% of the critical words and were therefore excluded from all further analyses. In the remaining set of participants, 90% of the critical words were accepted as the intended words.

Test

Figure 3 shows the proportion of responses, in which participants selected the word with the alveolar stop (i.e., /d/ or /t/ rather than /g/ or /k/). The results indicate little effect of the exposure condition on the results from the test phase.

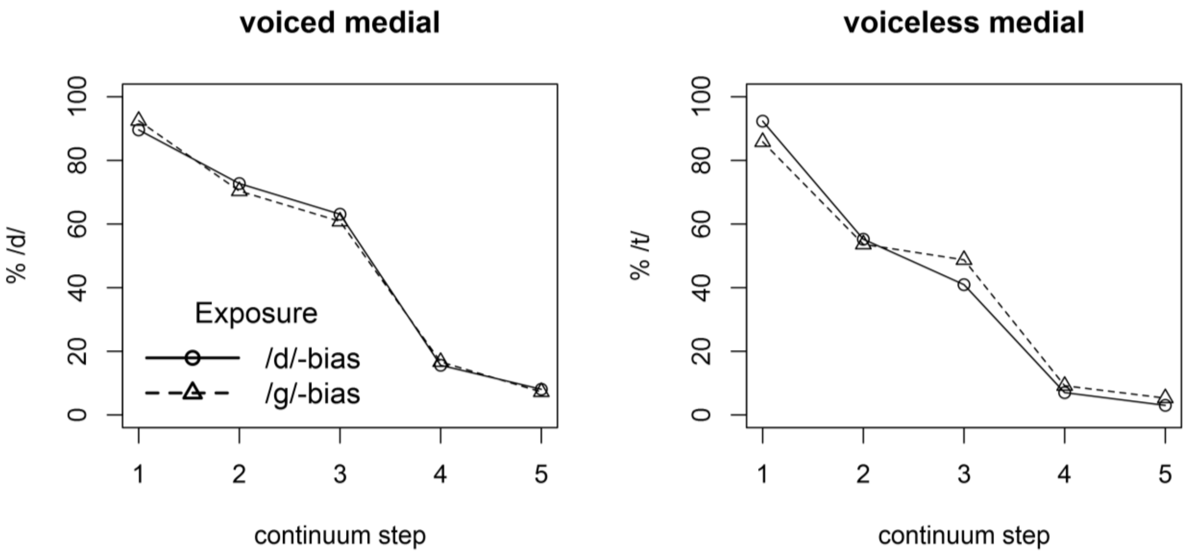


Figure 3: Percentage of word choices containing the alveolar consonant in the phonetic categorization task at test in Experiment 2, depending on continuum step (x-axis) and exposure condition (different lines). A learning effect would be reflected in the identification function of the /d/-bias group being “above” the identification function of the /g/-bias group.

Table 3: Results of the overall analysis of Experiment 2.

	b	z	p
(Intercept)	-0.488	-3.170	.002
Group	-0.010	-0.033	.974
Voicing	-1.010	-2.860	<.005
Step	-1.691	-24.426	<.001
Group*Voicing	0.109	0.154	.878

Statistical analyses were carried out as in Experiment 1 using generalized linear mixed-effects models. The model was fitted with response as the dependent variable (the word containing the alveolar stop coded as 1, the velar as 0) for which a logistic linking function was used. Fixed factors were Exposure Group (/d/-bias coded as -0.5, /g/-bias as 0.5), underlying Voicing (voiceless coded as -0.5, voiced as 0.5) and their interaction. Continuum Step (centred on 0, re-scaled to range from -0.5 to 0.5) was also entered but was not allowed to interact with the other fixed factors. Participant was entered as a random factor with uncorrelated random slopes for all within-participant fixed factors. In line with what the visual inspection of Figure 3 suggests, Table 3 shows that there is no effect of exposure Group on the identification of the test continua. The effect of voicing reflects that more alveolar responses were provided for the voiced continuum than for the voiceless continuum. However, this did not differ across exposure groups.

Since Experiment 2 did not show any generalization of perceptual learning to word-medial stops, we performed a Bayesian analysis (Rouder, Speckman, Sun, Morey, & Iverson, 2009) to test whether the data in fact support the null hypotheses. For this analysis we used the BayesFactor package (v0.9.12, Morey, Rouder, & Jamil, 2015). Since the BayesFactor package does not allow an analogue to a linear mixed-effects model, we calculated a Bayesian version of an ANOVA. To this end, we calculated the logOdds of the proportion of alveolar responses for each participant in each cell of the design. These were then used for a Bayesian ANOVA for repeated measures using the default priors implemented in the BayesFactor package.[2] The outcome of this analysis is presented in Table 4, with the critical model comparison for each effect (Navarro, 2015, p. 584). Each comparison tests the contribution of a factor by comparing a model with that factor to a simpler model without that factor (comparable to a Type II error in an ANOVA). A Bayes Factor (BF) larger than three then indicates evidence for an effect of this factor, while a BF < 1/3 provides evidence for the null-hypothesis. The results indicate that there were more alveolar responses in the voiced continuum (i.e., effect of Voicing with a BF > 3).

[2] The r-command was: allBFs = anovaBF(logitAlv ~ continuum*Group + participant, data = aggregatedLogOdds, whichRandom = "participant")

Table 4: Results of Bayesian analysis of Experiment 2

Effect	Numerator Model	Denominator model	Bayes Factor
Group	Group + Voicing + pp	Voicing + pp	0.25
Voicing	Group + Voicing + pp	Group + pp	8.67
Group * Voicing	Group + Voicing + Group:Voicing + pp	Group + Voicing + pp	0.29

Note: *pp* indicates the random factor participant.

Critically, the data provides evidence that the factor Exposure Group did not influence the amount of alveolar responses for either of the continua. This is because both BFs involving the factor Exposure Group, the main effect and the interaction with Voicing, are smaller than one third.

Discussion

The results of Experiment 2 clarify the uncertainties that remained Experiment 1. The descriptive data from Experiment 1 (see Figure 2) suggested that there might be learning for the phonetically similar voiceless medial condition (*Werte-Werke*, Engl., *values-works*) but not for the voiced medial condition (*borden-borgen*, Engl. *to lend-to board*). In contrast to that, the results suggest that there is no generalization in either case. The results show no generalization of perceptual learning from word-final underlyingly voiced but phonetically devoiced stops to phonologically identical but phonetically different voiced stops in word-medial position. Perceptual learning for place of articulation in word-final phonetically devoiced but phonologically voiced stops hence seems unconstrained by the underlying representation of these stops. However, learning also did not generalize to the phonetically somewhat similar voiceless stops in medial position, even though these stimuli shared the release burst as an important cue for place of articulation with the exposure stimuli. Apparently, the additional cues in the formant transitions into the schwa were potent enough to prevent generalization.

General Discussion

The purpose of the current study was to clarify under what conditions perceptual learning generalizes to other perceptual categories than the ones heard during exposure. Earlier studies (Mitterer, et al., 2013; Reinisch & Mitterer, 2016; Reinisch et al., 2014) had shown little evidence of generalization, suggesting that pre-lexical units in speech perception might be context-dependent allophones rather than more abstract phonemes. A recent result (Mitterer et al., 2016b), however, suggested that more abstract representations may still play a role by fostering generalization between different surface forms when they are versions of the same underlying form. Next to this potential role of underlying representations, an alternative interpretation for this generalization was the strong acoustic similarity between exposure and generalization items, with the main cues to place of articulation in the release burst.

Therefore, the current paper investigated learning based on devoiced stops in German as they occur in offset position. First of all, devoicing of underlying voiced stops is incomplete and hence provides a stronger motivation for listeners to activate the underlying form. Secondly, the cues to place of articulation differ strongly between word-final and word-medial implementations. Word-final voiced stops surface as unvoiced stops, in which the strongest place cues are in the release burst. In word-medial position—especially when followed by /ən/—there is no such release burst, and place of articulation is cued by formant transitions and the following nasal, which undergoes place assimilation. Under these circumstances, no generalization was found despite a shared underlying representation. This hence suggests only a very limited role for abstract phonological units in speech perception: Differing underlying representations do not obstruct generalization and a shared underlying representation does not necessarily foster generalization. Instead, perceptual units seem to be strongly bound to phonetic surface forms, and differences in underlying representations do not seem to make much of a difference in generalization of perceptual learning. As such, it is likely that the generalization obtained by Mitterer et al. (2016b) was caused by the strong phonetic similarity of exposure and test items, rather than a shared underlying representation.

It is important to note that for lexically-guided perceptual learning, it is crucial that the critical exposure items are recognized as the intended words (Sjerps & Reinisch, 2015). This was the motivation to make slight changes to the exposure items for Experiment 2. However, listeners also sometimes need to adapt to stronger changes, for instance, when confronted with complete mispronunciations from a second language learner. Under such circumstances, other learning mechanisms may be invoked, for which other forms of generalization may be found. In that vein, Eisner, Melinger, and Weber (2013) found that learning by native speakers of British English from exposure to a German speaker who devoices stops in coda position generalized to the onset position, so that *town* seemed to be recognized as *down* (based on a cross-modal priming measure). However, the interpretation of this data set is complicated by the fact that there was no priming in a control condition (auditory *town* did not prime visual *town*), and replication would be required to show that these findings were not an artefact of these particular stimuli. Nevertheless, other perceptual learning studies have indicated that perceptual learning in speech is different, especially in terms of efficiency, when the listener can immediately understand the word uttered (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Mitterer & McQueen, 2009). It remains an open question whether other types of learning are more likely to generalize over phonetic differences between exposure and test items than lexically-guided learning. One possibility is that the apparent finding of generalization in the case of Eisner et al. (2013) is in fact not learning about voicing neutralization but just a general adaptation in accepting more variance, an adaptation that has been observed for speech with intermittently overlaid noise (Huettig & McQueen, 2009) and casual speech with many phonetic reductions (Brouwer, Mitterer, & Huettig, 2012).

As already hinted at in the introduction, Bowers et al. (2016) recently provided some evidence that appears to contradict the series of perceptual learning studies quoted here (including the present one). They found that selective adaptation for stops can generalize over syllable position, which they argued supported the assumption of context-independent phonemes. However, since Bowers et al. used a different paradigm (selective adaptation, see Remez, 1987, for discussion) further research will be necessary to settle this issue. Importantly, it remains unclear what the function of the more abstract unites proposed by Bowers et al. (2016) would be. The perceptual-learning paradigm reflects units that are critically involved in solving the central problem in speech perception, the invariance problem. Moreover, the units involved in perceptual-learning are used early in speech perception (Mitterer & Reinisch, 2013). Both these arguments indicate that the perceptual-learning paradigm certainly can reveal units that are used in speech perception at a pre-lexical level. In contrast, finding selective adaptation from one set of stimuli to another does not necessarily show that two sets share an abstract phoneme. Selective adaptation can also be found if the two sets of stimuli overlap in simple or complex acoustic properties (Samuel & Kat, 1996) or even at the response level (Remez,1987). More research with diverse allophones—that show different amounts of acoustic-phonetic overlap—hence seems necessary to make more definite statements.

Related to this issue, it could be argued that the adaptation in the perceptual-learning paradigm could be the consequence of episodic representations which are not yet integrated with the speech-perception system proper and the (linguistic) units it uses. The learning of new words provides an analogy here, in which early work suggested that initial word learning is based on episodic representations that might reside in the hippocampus, which only become part of the cortical linguistic system after memory consolidation, which is facilitated by sleep (Gaskell & Dumay, 2003). However, we think this account is unlikely. First, the example of newly learned words does not supply an example case where such a strong distinction between episodic representations and linguistic presentations can be made, since more recent work shows immediate integration of new words in the linguistic system (Kapnoula & McMurray, 2016). Second, even if a distinction between episodic and linguistic representations could be made for perceptual learning of speech sounds, one should be able to find that sleep has a strong influence on the effect (cf. Gaskell & Dumay, 2003). This was explicitly tested by Eisner and McQueen (2006), who tested the stability of perceptual learning over a 12h period during which participants did or did not sleep. Two groups went through exposure around 9am or 9pm and were tested twelve hours later around 9pm or 9am the following day, respectively. Contrary to the assumption that sleep should strongly influence perceptual adaptation of speech sounds, the two groups showed clear perceptual adaptation effects that were equivalent in size. That is, sleep did not matter. Finally, it is also difficult to see how, computationally, an account of perceptual adaptation based on episodic representations could explain the data that show generalization to other words (McQueen et al., 2006; Mitterer et al., 2011). To do so, these newly generated episodic representations would have to be connected to all other words containing these sounds to influence perception. While we cannot fully reject an account in which perceptual learning influences transitory episodic representations rather than units of speech perception, we think that such an account is difficult to maintain considering the data gathered with the paradigm.

Another important result of the present study is that it confirms previous studies suggesting that perceptual learning is unlikely to operate on articulatory or phonological features. Previous studies mostly considered generalization from stops to nasals (Reinisch & Mitterer, 2016; Reinisch et al., 2014). Given the finding that manner of articulation is a feature of high importance for word recognition (Ernestus & Mak, 2004), it might be argued that this blocks generalization. The current data provides another data point of failure of generalization across features: here generalization of the adaptation of a place of articulation contrast across voicing that would be considered a “minor” feature for word recognition (Ernestus & Mak, 2004). This shows that the failure to generalize can be observed with different combinations of exposure, test and generalization contrasts.

An additional contribution of the current experiments is that perceptual learning and its generalization was tested for the contrast between velar and alveolar consonants, while the earlier experiments tested the contrast between labial and alveolar consonants. It may be argued that earlier experiments are difficult to interpret, because, from the point of view of production, ambiguous segments in between a labial and an alveolar place of articulation do in fact not exist. The difference between these two places of articulation is categorical rather than gradient, because they involve different articulators (tongue versus jaw/lips). This argument falters with the current experiments, because the velar-alveolar distinction is a gradient one of constriction location between tongue and palate, and intermediate cases do occur (cf. Mann, 1980). The current results hence show that these articulatory differences do not seem to matter and reinforce the conclusions from earlier experiments.

The current paper also indicates that featural decomposition is not an important part of pre-lexical processing in spoken-word recognition. Note that the assumption of featural decomposition makes different predictions about generalization than the assumption of context-independent phonemes. A phonemic account would predict that learning on /d/-final words generalizes to all other words containing /d/, but not to words containing /t/. A featural-decomposition account would predict generalization to other phonemes with the same place feature, including words containing /t/. Including the current data, there is now a total of seven attempts to find evidence for a featural decomposition of the speech signal that turned out negative (Mitterer et al., 2016b, 2013; Reinisch & Mitterer, 2016; Reinisch et al., 2014). At this juncture, it becomes difficult to argue that these are just null results. Moreover, the only two studies that did find generalization and could be interpreted in terms of generalization across “features” (Kraljic & Samuel, 2006; Mitterer et al., 2016b) both used strongly overlapping acoustic/*phonetic* properties between their exposure sets and test words and, hence acoustic similarity could just as well explain these results. Taken together there is now sufficient evidence to argue that, with a paradigm that shows functional generalization that is useful for speech perception, one does not find evidence for a featural decomposition of the speech signal, despite the widespread support for this idea in linguistics. Recently, Embick and Poeppel (2014) argued that there is a regrettable disconnect between linguistic and psychological research into speech processing, so that psychologists tend to ignore linguistic research and vice versa. We wholeheartedly agree with this assessment (see, e.g., Mitterer, et al. 2016a). However, Embick and Poeppel also argue that psychologists should take note of the prevalence of featural theories in linguistics and reconsider their preference for segments. The current line of investigations provides a cautionary note here. While features are highly useful to parsimoniously explain phonological processes, it seems that speech processing does not make much use of featural decomposition of the speech signal. Other findings from the field of speech perception also support this view (Ettlinger & Johnson, 2009; Kang, Johnson, & Finley, 2016; but see Mesgarani, Cheung, Johnson, & Chang, 2014). Instead, listeners seem to make use of segment-sized units for functional generalizations in speech perception, and the relevant tests indicate that these may be diphones or context-dependent allophones.

The current data also carry implications for another debate regarding the phonetics-phonology interface: the role of incomplete neutralization. As reviewed in the introduction, there are, on a sample level, phonetic differences between devoiced and underlyingly voiceless stops (Roettger et al., 2014). In the current study, all critical exposure items were devoiced, that is, phonologically voiced tokens produced as devoiced stops. Listeners have access to the underlying voicing—independent of orthography—because voicing differences surface in morphologically related forms. Despite listeners’ access to the underlying voicing specification, learning generalized to underlyingly voiceless tokens. This suggests that the small phonetic differences between true voiceless stops and devoiced stops is not sufficient for listeners to use as different pre-lexical representations for these two categories. This suggests that the phonetic differences caused by incomplete neutralization may not be relevant in perception, but simply arise through production dynamics. Importantly, this conclusion is limited to devoicing in German, since, in other cases of incomplete neutralization, such as incomplete phonological assimilations, there is evidence that listeners exploit small phonetic differences (Gow, 2002; Mitterer, Csépe, Honbolygo, & Blomert, 2006).

The current results also bear relevance to the controversy regarding the involvement of orthography in spoken-word recognition (Mitterer & Reinisch, 2015; Pattamadilok, Morais, Colin, & Kolinsky, 2014). If orthographic representations were automatically activated during spoken-word recognition, it may well be expected they should also be able to influence perceptual learning. However, the current results show that different orthographic representations do not hinder generalization of learning (learning on “d”-final items like *Fahrrad* (Engl. *bike*), *Rennpferd* (Engl. *race horse*), etc., generalized to “t”-final *Wert,* Engl. *value*). Similarly, sharing the same orthographic representation (as *borden*, Engl., *to board*, does with *Fahrrad, Rennpferd,* etc.) does not lead to generalization. This indicates that orthographic representations, even if activated in speech perception, seem to have little functional relevance.

In summary, our results reinforce the conclusions from previous studies (Mitterer et al., 2013; Reinisch & Mitterer, 2016; Reinisch et al., 2014) that listeners neither analyze the speech stream as a stream of context-independent phonemes nor decompose the speech signal into context-free phonological/articulatory features at a prelexical level. Instead, listeners seem to make use of context-dependent segments, and those segments are strongly constrained by their phonetic surface form. Listeners seem to extract segments of sufficient acoustic similarity from the input and use them for generalization of learning in speech perception. Consequently, the “alphabet” of the listener in speech perception may be much larger than the number of phonemes assumed for a given language.

References

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*, 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Bowers, J. S., Kazanina, N., & Andermane, N. (2016). Spoken word identification involves accessing position invariant phoneme representations. *Journal of Memory and Language*, *87*, 71–83. https://doi.org/10.1016/j.jml.2015.11.002

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, *27*(4), 539–571. https://doi.org/10.1080/01690965.2011.555268

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10* (pp. 91–111). Berlin: de Gruyter.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*, 222–241.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. https://doi.org/10.3758/BF03206487

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America*, *119*(4), 1950–1953. https://doi.org/10.1121/1.2178721

Eisner, F., Melinger, A., & Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology*, *4*, 148. https://doi.org/10.3389/fpsyg.2013.00148

Embick, D., & Poeppel, D. (2014). Towards a computational(ist) neurobiology of language: correlational, integrated and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, *0*(0), 1–10. https://doi.org/10.1080/23273798.2014.980750

Ernestus, M., & Mak, W. M. (2004). Distinctive phonological features differ in relevance for both spoken and written word recognition. *Brain and Language*, *90*(1–3), 378–392. https://doi.org/10.1016/S0093-934X(03)00449-8

Ettlinger, M., & Johnson, K. (2009). Vowel discrimination by English, French and Turkish speakers: evidence for an exemplar-based approach to speech perception. *Phonetica*, *66*(4), 222–242. https://doi.org/10.1159/000298584

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*, 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279. https://doi.org/10.1037//0033-295X.105.2.251

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 49–54). Dudweiler, Germany: Pirrot.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics*, *31*, 305–320. https://doi.org/10.1016/S0095-4470(03)00030-5

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. In N. O. Schiller & A. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 159–207). Berlin: Mouton de Gruyter.

Gow, D. W. (2002). Does English coronal place assimilation create lexical ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 163–179. https://doi.org/10.1037//0096-1523.28.1.163

Huettig, F., & McQueen, J. M. (2009). AM radio noise changes the dynamics of spoken word recognition.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, *18*, 943–950. https://doi.org/10.3758/s13423-011-0129-2

Jun, S.-A. (1998). The accentual phrase in the Korean prosodic hierarchy. *Phonology*, *15*(02), 189–226.

Kang, S., Johnson, K., & Finley, G. (2016). Effects of native language on compensation for coarticulation. *Speech Communication*, *77*, 84–100. https://doi.org/10.1016/j.specom.2015.12.005

Kapnoula, E. C., & McMurray, B. (2016). Newly learned word forms are abstract and integrated immediately after acquisition. *Psychonomic Bulletin & Review*, *23*, 491–499. https://doi.org/10.3758/s13423-015-0897-1

Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction. *Speech Communication*, *27*, 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, *13*, 262–268. https://doi.org/10.3758/BF03193841

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*, 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, *38*, 44–59.

Mann, V. A. (1980). Influence of preceding liquid on stop consonant perception. *Perception & Psychophysics*, *28*, 407–412.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113–1126. https://doi.org/10.1207/s15516709cog0000_79

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science*, *35*, 184–197. https://doi.org/10.1111/j.1551-6709.2010.01140.x

Mitterer, H., Cho, T., & Kim, S. (2016a). How does prosody influence speech categorization? *Journal of Phonetics*, *54*, 68–79. https://doi.org/10.1016/j.wocn.2015.09.002

Mitterer, H., Cho, T., & Kim, S. (2016b). What are the letters of speech? Testing the role of phonological specification and phonetic similarity in perceptual learning. *Journal of Phonetics*, *56*, 110–123. https://doi.org/10.1016/j.wocn.2016.03.001

Mitterer, H., Csépe, V., Honbolygo, F., & Blomert, L. (2006). The recognition of phonologically assimilated words does not depend on specific language experience. *Cognitive Science*, *30*(3), 451–479. https://doi.org/10.1207/s15516709cog0000_57

Mitterer, H., & de Ruiter, J. P. (2008). Recalibrating color categories using world knowledge. *Psychological Science*, *19*(7), 629–634. https://doi.org/10.1111/j.1467-9280.2008.02133.x

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *Plos One*, *4*(e7785). https://doi.org/10.1371/journal.pone.0007785

Mitterer, H., & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language*, *69*, 527–545. https://doi.org/10.1016/j.jml.2013.07.002

Mitterer, H., & Reinisch, E. (2015). Letters don’t matter: No effect of orthography on the perception of conversational speech. *Journal of Memory and Language*, *85*, 116–134. https://doi.org/10.1016/j.jml.2015.08.005

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*, 356–361. https://doi.org/10.1016/j.cognition.2013.07.011

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes Factors for Common Designs (Version 0.9.12-2). Retrieved from https://cran.r-project.org/web/packages/BayesFactor/index.html

Navarro, D. (2015). *Learning Statistics with R: A tutorial for psychology students and other beginners*. Adelaide: University of Adelaide.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Pattamadilok, C., Morais, J., Colin, C., & Kolinsky, R. (2014). Unattentive speech processing is influenced by orthographic knowledge: Evidence from mismatch negativity. *Brain and Language*, *137*, 103–111. https://doi.org/10.1016/j.bandl.2014.08.005

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 101–139). Berlin: Mouton de Gruyter.

Poellmann, K., McQueen, J. M., & Mitterer, H. (2011). The time course of perceptual learning. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences 2011 [ICPhS XVII]* (pp. 1618–1621). Hong Kong: Hong Kong: Department of Chinese, Translation and Linguistics, City University of Hong Kong.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 539–555. https://doi.org/10.1037/a0034409

Reinisch, E., & Mitterer, H. (2016). Exposure modality, input variability and the categories of perceptual recalibration. *Journal of Phonetics*, *55*, 96–108. https://doi.org/10.1016/j.wocn.2015.12.004

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 75–86. https://doi.org/10.1037/a0027979

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics*, *45*, 91–105. https://doi.org/10.1016/j.wocn.2014.04.002

Remez, R. E. (1987). Neural models of speech perception: a case history. In S. Harnad (Ed.), *Categorical Perception: The groundwork of cognition* (pp. 199–225). Cambridge, Mass.: Cambridge University Press.

Roettger, T. B., Winter, B., Grawunder, S., Kirby, J., & Grice, M. (2014). Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics*, *43*, 11–25. https://doi.org/10.1016/j.wocn.2014.01.002

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Samuel, A. G., & Kat, D. (1996). Early levels of analysis of speech. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 676–694. https://doi.org/10.1037/0096-1523.22.3.676

Sjerps, M. J., & Reinisch, E. (2015). Divide and conquer: How perceptual contrast sensitivity and perceptual learning cooperate in reducing input variation in speech perception. *Journal of Experimental Psychology. Human Perception and Performance*, *41*(3), 710–722. https://doi.org/10.1037/a0039028

Steriade, D. (2001). Directional asymmetries in place assimilation: a perceptual account. In E. Hume & K. Johnson (Eds.), *The role of speech perception in phonology* (pp. 219–250). New York, NJ: Academic Press.

Appendix

Table A1: Critical /d/ and /g/-final exposure words and their SUBTLEX-DE (Brysbaert et al., 2011) frequency per million.

/d/-final words (English translation)	SUBTLEX-DE Frequency	/g/-final words (English translation)	SUBTLEX-DE Frequency
Abschied (goodbye, *n*)	11.69	Abflug (departure)	5.16
Absurd (absurd)	8.86	Analog (analogue)	0.24
Billard (Billiard)	2.05	Anzug (suit)	35.00
Blöd (stupid)	49.65	Aufschlag (serve)	3.23
E-Herd (electrical oven)	0.00	Belag (covering)	0.63
Freibad (open pool)	0.24	Beleg (receipt)	0.83
Kamerad (buddy)	6.14	Berg (mountain)	32.05
mild (mild)	1.30	Briefumschlag (envelope)	0.71
paranoid (paranoid)	7.48	Erfolg (success)	46.97
Raubmord (holdup murder)	0.04	Hamburg (Hamburg)	3.39
Rennpferd (race horse)	0.79	Herzog (duke)	7.87
Rollfeld (airfield)	1.10	Hochburg (stronghold)	0.20
Schild (shield)	21.97	Monolog (monologue)	1.65
Schwarzwald (black forrest)	0.16	Privileg (priviledge)	3.35
Sinnbild (allegory)	0.47	Sarg (coffin)	12.24
sobald (as soon as)	116.15	schräg (skew)	5.00
Spielschuld (gaming debt)	0.08	Umweg (detour)	3.07
Unschuld (innocence)	8.66	Verlag (publisher)	2.48
Urwald (jungle)	0.71	Wahlsieg (election victory)	0.24
Vorbild (model)	8.86	Zwerg (dwarf)	9.33

Note: E-Herd is short for *Elektrischer Herd*, Engl., *electrical oven*, with Herd having a frequency of 5.64 per million.