PERCEPTUAL LEARNING IN SPEECH IS PHONETIC, NOT PHONOLOGICAL: EVIDENCE FROM FINAL CONSONANT DEVOICING

Eva Reinisch[1] and Holger Mitterer[2]

[1]Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich, Germany;
[2]Department of Cognitive Science, University of Malta
evarei@phonetik.uni-muenchen.de and holger.mitterer@um.edu.mt

ABSTRACT

Listeners flexibly recalibrate the perceptual categorization of sounds in response to speakers' unusual pronunciation variants. Recent studies have shown that generalization of this recalibration can inform us about the nature of prelexical units used for speech perception. The present study tested whether this generalization is sensitive to phonetic or phonological properties of speech. Using the example of German word-finally devoiced stops, we found that generalization does not extend to sounds that match in phonological specification (here: voiced) when they are dissimilar in their acoustics (phonetically voiceless in word-final position to phonetically voiced word-medially). However, phonologically voiceless stops in word-final and word-medial position did show the effect. This supports suggestions 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.

Keywords: speech perception, perceptual learning, phonetic recalibration, final devoicing

1. INTRODUCTION

Listeners flexibly adapt to speakers' idiosyncratic pronunciation variants by using lexical context to adjust category boundaries [2-10]. 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) they 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*). A question that has received quite some attention but still remains unresolved is what kinds of units listeners are recalibrating [6,10]. The present study addresses this issue by testing whether recalibration is sensitive to phonetic or phonological properties of speech using the example of German word-final consonant devoicing. Understanding the workings of perceptual learning may inform us about speech processing more generally.

While previous studies of perceptual learning implicitly or explicitly assumed that the categories for recalibration are phonemes [5], studies testing generalization of recalibrated categories across speakers suggest that phonetic similarity is more important than a mere match in phoneme category. For example, generalization of a recalibrated fricative contrast across speakers was found only if the two speaker's fricative spaces (/f/-/s/ or /s/-/ʃ/) were acoustically or perceptually similar [4,9].

Results for generalization across sound contrasts and positions in a word, however, are mixed. While generalization of a recalibrated voicing contrast (/d/-/t/) across place of articulation (to /b/-/p/) could be explained by recalibration of a phonological or articulatory voicing feature [4], generalization of a recalibrated place of articulation contrast across manner of articulation – that would also be predicted under a phonological feature account – did not occur [10]. Alternative suggestions for the units of recalibration include allophones [6] and variably-sized chunks of speech depending on the types of contrasts involved [6,10].

Given this mixed evidence about the targets of perceptual recalibration, the present study set out to further narrow down when generalization is possible and when not. Specifically, we tested generalization of a recalibrated place of articulation contrast for phonologically voiced stops in German (i.e., /d/-/g/). However, during exposure the critical stops were presented in word-final position where, through final devoicing, they were realized as voiceless. Following this exposure, listeners were tested on their categorization of four minimal pair continua involving the following conditions:

- "voiced" final: *Bord-borg* (board as in [on board of a ship] or [short for snowboard] – lend! [imperative of *borgen*]; also SiFi characters "Borg"), where phonological specification, realization, and position matched the word-final devoicing context experienced during exposure.

- voiceless final: *Werk-Wert* (work - value), where realization and position but not phonological specification matched exposure.
- voiced medial: *borden-borgen* (to board [a ship] – to lend), where phonological specification but neither phonetic realization nor position matched exposure.
- voiceless final: *Werke-Werte* ([plurals of *Werk-Wert*]), where phonetic realization but neither phonological specification nor position matched exposure.

If perceptual learning was sensitive to phonological properties of the sound contrast, then a recalibration effect should be evident in the two voiced conditions: the "voiced" 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 all pairs but the voiced medial one, unless generalization was also position-specific. This should additionally preclude an effect for the voiceless medial condition (but see [2] for evidence of generalization across word position).

2. METHODS

2.1. Participants

Forty-four native speakers of German, students at the University of Munich took part for pay.

2.2. Materials

100 German words and 100 phonotactically legal nonwords were selected for the lexical decision task that served as exposure (following [7]). 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 Section 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.

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]* "bike" was also recorded as *Fahrra[k]*). Within these pairs speaking rate and pitch were closely matched.

Critical words and minimal pairs were then morphed in 11-step continua using STRAIGHT [3]. 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. 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 5 equally spaced steps around these midpoints leaving out every other step from the original eleven-step continua.

2.3. Procedure

2.3.1. Exposure

Half of the participants were randomly assigned to a /d/-bias condition, half to a /g/-bias condition. All participants were presented the same 100 words and nonwords except for the 20 critical items in which, depending on group, /d/- or /g/-words were replaced by the ambiguous morphs. Words with the other critical stop were presented in unambiguous form.

Participants were seated in a sound-proof booth and listened to the stimuli over headphones. Their task was to decide on every trial whether they heard a word or nonword 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.

2.3.2. Test

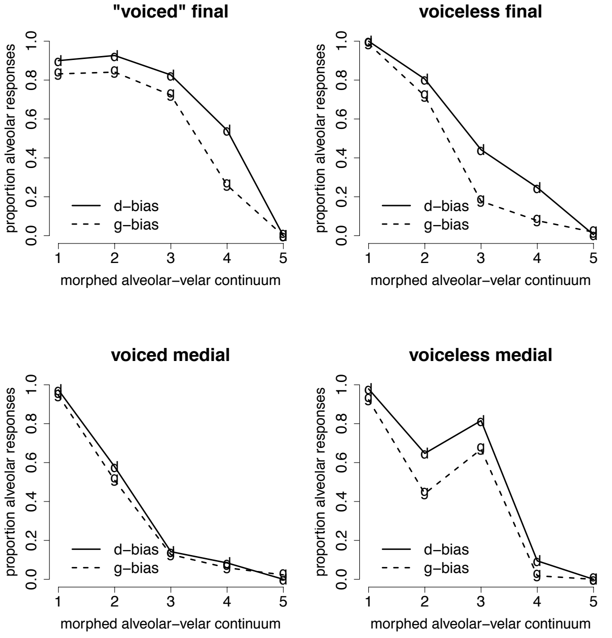
Immediately following exposure, all participants performed the same phonetic categorization task with the four minimal pair continua. On each trial, participants were first presented the upcoming pair written on the screen with the word containing /d/ or /t/ on the left. As is typical for German, phonological voicing was coded orthographically in these words. 500 ms later the 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.

3. RESULTS

3.1. Exposure

Three participants in the /d/-bias condition rejected more than 50% of the critical words 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 recalibration to occur [8]. Nonwords do not trigger recalibration [7]. Of the remaining set, 95% of the critical words were accepted as the intended words.

Figure 1: Proportion alveolar responses in the phonetic categorization task at test.



3.2. Test

Figure 1 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 1, for all but the voiced medial condition (*borden-borgen*) participants in the /d/-bias condition gave more responses favouring the alveolar stop than the /g/-bias group. Hence perceptual recalibration was found and generalized to two further conditions.

Statistical analyses were carried out using generalized linear mixed-effects models. The main model was fitted with response as the dependent variable (the word containing the alveolar stop coded as 1, the velar as 0), and 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 as fixed factors. 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 random slopes for all within-participant fixed factors (i.e., all but Group since this factor was manipulated between participants; [1]). Table 1 shows the results.

Table 1: Results of the overall analysis.

Factor	*b*	*t*	*p*
Intercept	-0.47	-2.13	<.05
Group	-1.09	-2.46	<.014
Position	0.11	0.39	0.69
Voicing	1.66	5.26	<.001
Step	-5.22	-24.2	<.001
Group*Voicing	0.52	0.91	0.36
Group*Position	-0.70	-1.11	0.27
Voicing*Position	3.59	5.50	<.001
Group*Voicing*Position	-1.41	-1.09	0.28

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 1 suggests, however, there was no interaction between Group and either Voicing or Position or a three-way interaction. To follow up on this discrepancy, linear mixed-effects models were run 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 effects of Group (in addition to an effect of continuum Step).

Table 2: Results of the analyses per condition.

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

4. 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 recalibrate 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 recalibration was found at test for the minimal pair in which the stop contrast fully matched the exposure condition (i.e., "voiced" final *Bord-borg*) and the pair that matched in phonetic realization and word position but not phonological voicing (i.e., voiceless final, *Wert-Werk*). While the status of recalibration for the voiceless medial condition (*Werte-Werke*) appeared somewhat unclear (for a discussion see below), critically, in none of the analyses we found generalization to the minimal pair that was phonologically equal but phonetically different from the realization of the exposure contrast (medially voiced *borden-borgen*). That is, phonetics rather than phonology appears to matter for generalization of perceptual learning.

What remains to be explained is the role of position. Generalization to the voiceless medial condition (*Werke-Werte*) that seems to be present in Figure 1, failed to reach significance in this condition's analysis. However, a clearer picture emerged in another analysis where only the most ambiguous sounds in each condition, (continuum steps that received responses closest to 50%), were taken into account. Here, significant effects of Group were found for all three conditions in which the stop was realized as voiceless, but again - and critically so - no group difference was found for the voiced medial condition. This suggests that the fact that the phonologically voiced stop only surfaces as phonetically voiced in word medial position is likely not the only cause for the lack of generalization found for this condition.

Nevertheless, our results have to be treated with caution, given the absence of a significant interaction in the overall analysis. This is partly due to strong inter-individual differences in the perception of the stop-continua. It is noteworthy that in previous studies on fricatives, the use of STRAIGHT [3] for morphing led to more consistency over perceivers than other methods (e.g., sample-by-sample interpolation used in [7]). This apparent advantage for STRAIGHT-generated continua appears not to apply to stops.

However, if we consider the results from the analyses per condition, we find that perceptual learning generalizes to phonetically similar tokens of the respective sound contrast but is insensitive to phonological properties. This is in line with previous findings on perceptual learning [4,6,9,10]. Note that findings that could have been ascribed to phonological features such as voicing features that generalize across place of articulation [4] can also be explained by a phonetic similarity account [9]. Cues to stop voicing in English are mainly durational in nature and as such differ little between /d/-/t/ and /b/-/p/. In contrast, generalization across phonologically defined allophones of a phoneme could not be found since they vastly differed in terms of phonetics and articulation [6].

In addition to its contribution to our understanding of perceptual learning, the present study has repercussions on another debate at the phonetics-phonology interface: the debate on incomplete neutralization of final devoicing (see e.g., [11]). That is, phonologically voiced stops in word-final position have been shown to differ phonetically from phonologically voiceless stops, mostly by duration differences in the preceding segment. Since in the present study all exposure items were of the type that are phonologically voiced, and this voicing surfaces in morphologically related forms (e.g., *Fahhra[t]–Fahrrä[d]er*; bike-bikes), listeners should have had lexical as well as phonetic cues to the phonological voicing status of our critical stops. It seems, however, that these cues were not strong enough to allow for generalization to the voiced medial condition. This indicates that, on the one hand, the differences between voiced stops and incompletely devoiced stops are too large to treat these two as the same type of phonetic category; on the other hand, differences between incompletely devoiced stops and voiceless stops appear too small for listeners to treat them as potentially different segments. This, in turn, supports suggestions that incomplete neutralization of voicing in word-final position may not be functionally relevant for perception (see also [11]) but merely surface in production.

In summary, our results reinforce the conclusions from previous studies [6,10] that listeners don't unpack the speech signal into context-free phonological/ articulatory features at a pre-lexical level. Instead, listeners appear to capitalize on segments/units/chunks of speech that are, in part, context-sensitive. Listeners seem to extract segments of sufficient acoustic similarity from the input and use them for generalization of learning in speech perception. That is, the "alphabet" of the listener in speech perception may be much larger than the number of phonemes assumed for a given language.

5. REFERENCES

[1] Barr D. J., Levy R., Scheepers C. & Tily, H. 2013. Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68, 255-278.

[2] Jesse, A., & McQueen, J. M. 2011. Positional effects in the lexical retuning of speech perception. *Psychon. B. Rev.* 18, 943–950.

[3] 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 Commun.* 27, 187–207.

[4] Kraljic, T., & Samuel, A. G. 2006. Generalization in perceptual learning for speech. *Psychon. B. Rev.* 13, 262–268.

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

[6] Mitterer, H., Scharenborg, O., & McQueen, J. M. 2013. Phonological abstraction without phonemes in speech perception. *Cognition* 129, 356–361.

[7] Norris, D., McQueen, J. M., & Cutler, A. 2003. Perceptual learning in speech. *Cognitive Psychol.* 47, 204–238.

[8] Poellmann, K., McQueen, J. M., & Mitterer, H. 2011. The time course of perceptual learning. *Proc. 16th ICPhS* Hong Kong, 1618-1621.

[9] Reinisch, E. & Holt, L. L. 2014. Lexically-guided phonetic retuning of foreign-accented speech and its generalization. *J. Exp. Psychol. Human* 40*,* 539-555.

[10] Reinisch, E., Wozny, D., Mitterer, H. & Holt, L. L. 2014. Phonetic category recalibration: What are the categories? *J. Phonetics* 45*,* 91-105.

[11] Roettger, T., Winter, B., Grawunder, S., Kirby, J. & Grice, M. 2014. Assessing incomplete neutralization of final devoicing in German. *J. Phonetics* 43, 11-25.