Journal of Phonetics 55 (2016) 96–108

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Exposure modality, input variability and the categories of perceptual recalibration

Eva Reinisch [a,*], Holger Mitterer [b]

[a] Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich, Germany
[b] Department of Cognitive Science, University of Malta, Malta

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
12 December 2015
Accepted 22 December 2015

Keywords:
Speech perception
Perceptual learning
Audiovisual processing
Prelexical processing

ABSTRACT

Recent evidence shows that studies on perceptual recalibration and its generalization can inform us about the presence and nature of prelexical units used for speech perception. Listeners recalibrate perception when hearing an ambiguous auditory stimulus between, for example, /p/ and /t/ in unambiguous lexical context (*kee[p/t]*->/p/, *mee[p/t]*->/t/) or visual context (presence vs. absence of lip closure). A later encountered ambiguous auditory-only stimulus is then perceived in line with the previously experienced context. Unlike studies using lexical context to guide learning, experiments with the visual paradigm suggested that prelexical units are rather specific and context-dependent. However, these experiments raised doubts whether lexically-guided and visually-guided recalibration are targeting the same type of units, or whether learning in the visually-guided paradigm—with limited variability during exposure—is task-specific. The present study shows successful visually-guided learning following exposure to a variety of different learning trials. We also show that patterns of generalization found with the visually-guided paradigm can be replicated with a lexically-guided paradigm: listeners do not generalize a recalibrated stop contrast across manner of articulation. This supports suggestions that the units of perception depend on the distribution of relevant cues in the speech signal.

1. Introduction

The nature of mental representations has been a long-standing issue in cognitive science. In the study of language, this leads to the question how we store our knowledge about the words of our language(s), and through which processes they can be accessed during listening. In recent years, the field converged on the conclusion that lexical access involves at least to some extent abstraction from the acoustic form of the input and the use of some kinds of units mediating between acoustic input and lexical representations (see e.g., Goldinger & Azuma, 2003, for an overview). In order to investigate the nature of these intermediate units, a number of recent studies made use of a perceptual learning paradigm ("phonetic recalibration"), based on the assumption that patterns of generalization are indicative of the shape of these intermediate units (e.g., Mitterer, Scharenborg, & McQueen, 2013; Reinisch, Wozny, Mitterer, & Holt, 2014). Since, however, the list of proposed units is long, and previous evidence for generalization or specificity of perceptual learning is mixed, the present study addressed two related questions: First, what role does variability in the input play in the occurrence of recalibration in different perceptual learning paradigms? This is to disentangle previous conflicting results that may be attributed to the use of visual vs. lexical context information to guide recalibration. Second, what role does variability in the input play in the occurrence of *generalization* of perceptual learning? Establishing conditions in which perceptual learning does or does not occur and generalize will allow us to assess what kinds of units listeners use for speech perception.

The perceptual learning paradigms discussed here were initially used to investigate how listeners use various types of context information to adjust or "recalibrate" speech perception in adapting to unusual speakers. That is, if listeners hear a speaker produce a sound that is ambiguous between /f/ and /s/ in lexical context where the interpretation as /f/ leads to the perception of a real word but /s/ does not (e.g., *giraffe* where *gira[s]* is not a word in English), listeners tend to resolve the acoustic ambiguity via the lexical context (i.e., they interpret the ambiguous sounds as /f/; Ganong, 1980). After repeated exposure to ambiguous sounds in unambiguous

[*] Correspondence to: Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich, Schellingstr. 3, 80799 Munich, Germany. Tel.: +49 89 2180 5752.
E-mail addresses: evarei@phonetik.uni-muenchen.de (E. Reinisch), holger.mitterer@um.edu.mt (H. Mitterer).

0095-4470/$ - see front matter © 2015 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2015.12.004

lexical context by that speaker, listeners interpret ambiguous sounds in line with the previously experienced context even if the current context is ambiguous. That is, listeners who had heard ambiguous sounds in words like *giraffe* before would perceive *knife* more often than *nice* (Norris, McQueen, & Culter, 2003; Samuel & Kraljic, 2009, for an overview over similar types of studies). Similar effects have been shown with visual (lipread) context. After hearing an ambiguous auditory stimulus between /b/ and /d/ paired with the visual image of the speaker closing his/her lips (indicating the labial) or not (here: indicating the alveolar sound) an ambiguous auditory-only stimulus is perceived in line with the previously experienced visual stimulus (Bertelson, Vroomen, & deGelder, 2003). These types of perceptual learning have been termed lexically-guided and visually-guided recalibration.

1.1. The units of perceptual recalibration

As suggested above, such studies on perceptual learning not only contributed to our understanding of adaptation to unusual pronunciation variants but also to the debate about the existence of pre-lexical units. Goldinger (1998) started off this debate by suggesting that the mental lexicon contains the combined acoustic traces of the words a listener has heard in a lifetime. This questioned the existence of pre-lexical units altogether. This extreme point of view, however, turned out to be difficult to maintain, among others, in the light of experiments on perceptual learning (e.g. McQueen, Cutler & Norris, 2006; see also Mitterer, Chen, & Zhou, 2011; Sjerps & McQueen, 2010). With the current state of research there seems to be a general consensus that listeners use some form of abstract units, but are also able to store exemplars (Cutler, Eisner, McQueen, & Norris, 2010; Goldinger, 2007). The present study takes the reasoning from perceptual-learning studies one step further: if perceptual learning studies can show the existence of some type of pre-lexical units, they might also reveal the shape of these units (see Mitterer et al., 2013; Reinisch et al., 2014 for similar arguments).

Following this line of thought, recent studies on perceptual learning specifically investigated the grain-size of prelexical units that are used in speech perception. Suggestions for candidate units range from abstract phonemes (McClelland & Elman, 1986), via articulatory features that are directly perceived from the speech input (D’Ausilio et al., 2009; Galantucci, Fowler & Turvey, 2006) to context-invariant phonological features (e.g., Chomsky & Halle, 1968; Lahiri & Reetz, 2002; Marslen-Wilson & Warren, 1994). Notably, all of these units have in common that they assume independence from the surrounding context, which should allow for some generalization. Indeed, Jesse and McQueen (2011) showed that perceptual recalibration of a fricative contrast can generalize from syllable-coda to –onset position. Mitterer et al. (2011) found generalization across phonetic contexts. But the phoneme may be too big a unit for recalibration.

The “next step down“ in the size of abstract units is to assume phonological or articulatory features as the units of perception. On the one hand, abstract phonological features could simply be seen as a construct from linguistics that conveniently serves the description of language. As such, they may fall short of psychological reality. On the other hand, there seems to be (at least some) evidence in line with claims that abstract phonological features are not only useful for linguistic description but are also primary in language processing (Cornell, Lahiri, & Eulitz, 2013; Embick & Poeppel, 2014; Roon & Gafos, 2014; Scharinger, Merickel, Riley, & Idsardi, 2011). Moreover, Embick and Poeppel (2014) argue for using phonological features as the base for neurobiological investigations of language (see also e.g., Poeppel, Idsardi & van Wassenhove, 2008). Importantly, the idea of abstract features is also supported by findings on generalization of perceptual learning, for example, the generalization of a recalibrated voicing contrast from an alveolar place of articulation (i.e., /d/–/t/) to labial place of articulation (/b/–/p/; Kraljic & Samuel, 2006). That is, recalibration may not be specific to a certain exposure contrast but rather affect features or gestures, for example, for “voicing”. However, in all these cases, the acoustic variation in the implementation of the phonological contrast over contexts was rather small, so that the finding of generalization does not contradict a feature account but cannot be used to argue *for* abstract features. Abstract features, like phonemes, are supposedly independent of context and hence predict generalization over acoustically dissimilar cues.

Recent studies then focused on cases in which the acoustic cues to a given distinction varied over contexts and these studies support some kind of context-sensitivity (e.g., Mitterer et al., 2013, Reinisch et al., 2014). Mitterer et al. (2013) showed that Dutch listeners recalibrate an /l/–/r/ contrast but fail to show recalibration when tested on one of the articulatorily and acoustically different allophones of /r/ (i.e., alveolar approximant vs. apical trill). The authors suggest that (context-sensitive) allophones may be the units of recalibration. Interestingly, this example also shows how the assumption of different types of features (i.e., abstract phonological features vs. gestures) results in different predictions regarding generalization. Assuming abstract features, one would have to predict that learning generalizes from one allophone to another, because all allophones of a given phoneme (in a given language) share the same feature specification (Lahiri & Reetz, 2002). If, in contrast, features were defined in terms of articulatory gestures (Galantucci et al., 2006) then the results could still be explained since the two allophones used in Mitterer et al. make use of different articulatory gestures. Finally, an account of acoustic similarity as requirement for generalization would also support the lack of generalization (see e.g., Kraljic & Samuel, 2006; Reinisch & Holt, 2014).

In order to adjudicate between abstract features, articulatory features, or (context-sensitive) allophones as the prelexical units that are relevant to perceptual recalibration, Reinisch et al. (2014) tested conditions of generalization in a tightly controlled experimental setup. They tested recalibration of a place of articulation contrast (/b/–/d/) and its generalization. A first experiment tested the generalization from "a[$^{d}/_{b}$]a" to "i[$^{d}/_{b}$]i" (where [$^{d}/_{b}$] indicates a perceptually ambiguous sound between /b/ and /d/). Via additional cue-trading experiments, Reinisch et al. established that the labial–alveolar distinction was cued mainly by formant transitions in the "a[$^{d}/_{b}$]a" case but mainly by burst spectra in the "i[$^{d}/_{b}$]i" case. The results of perceptual recalibration then showed that listeners did not generalize learning across these contexts (in either direction). While this first experiment tested the generalization to other acoustic implementations of the same phonemic contrast (i.e., comparable to allophones), the second experiment tested whether generalization occurs when the same cues (formant transitions) are used for different phoneme contrasts (/b/–/d/ versus /m/–/n/). Despite *identical* carrier stimuli cueing place of articulation by formant transitions (i.e., the same pair of vowels ("a_a") with formant transition continua from a labial to an alveolar place of articulation flanking either a (voiced) closure—leading to a percept of a stop—or a spectrally ambiguous nasal), recalibration did not generalize from "a[$^{d}/_{b}$]a" to "a[$^{n}/_{m}$]a". Finally, in order to test a case where generalization seemed most likely, Reinisch et al. tested generalization from "a[$^{d}/_{b}$]a" to "u[$^{d}/_{b}$]u", where the phoneme contrast is the same and the distinction in both cases is most strongly cued by formant transitions. Despite this minimal difference in context (same phoneme contrast, same type of cues), still no generalization occurred. In all experiments, results showed a consistent recalibration effect for the exposure contrast, but in none of the conditions did generalization occur – neither across contexts nor across manner of articulation. This led the authors to conclude that "pre-lexical processing does not make use of abstract phonological features, context-free phonemes, or speech gestures. Instead, the units of representation may vary according to the reliability and consistency across contexts of the cues for the phoneme (or allophone)" (p 104). The authors suggest that the grain-size of the units of perception may be adjusted to the extent that information about the phoneme contrast is spread out into adjacent segments, which would incorporate context-sensitivity. The present study picks up on this issue and assesses in more detail how the existence of prelexical units and a certain degree of context-sensitivity could interact in phonetic recalibration of speech perception.

1.2. Lexically vs. visually-guided recalibration

An additional goal of the present study was to assess the nature of perceptual learning experiments more generally, namely, the experimental setups and types of information (lexical vs. visual/lipread) used for recalibration. All studies on generalization but the one by Reinisch et al. discussed so far used lexically-guided recalibration to show the effect (following the paradigm first used by Norris et al., 2003). Reinisch et al. had used visually-guided recalibration using lipread context (following Bertelson et al., 2003; van Linden & Vroomen, 2007), since it would have been hard to control the exact acoustic implementation of the category contrasts in the lexical paradigm.

Interestingly, the two paradigms are often seen as interchangeable (e.g. Kleinschmidt & Jaeger, 2015), partly based on a study that indeed suggested that the two processes work similarly (van Linden & Vroomen, 2007). However, there are definitely differences between visual and lexical context effects in speech perception. Samuel and Lieblich (2014) found that lexically induced percepts, but not visually induced percepts, generate selective adaptation. Such dissociations show that the two sources of context may work differently[1]. This is supported by another set of studies that showed that, in some cases, visual context effects may be more pervasive than lexical context effects. Mitterer (2006) showed that a visually induced percept of a rounded vowel induced compensation for coarticulation in a neighboring fricative, while lexically mediated percepts of a rounded vowel failed to do so (Mitterer, 2007).

Given these differences between visually and lexically mediated context effects, it is hence an open question whether the two recalibration paradigms are as similar as often assumed. This is especially so if the question is whether learning can generalize. A number of differences between the typical lexically-guided vs. visually-guided recalibration studies has to be considered here. The typical setup of a lexically-guided recalibration study following Norris et al. (2003) consists of one long exposure block, most frequently a lexical decision task, in which ten to twenty different critical words containing the relevant ambiguous sound occur in a variety of contexts. This one exposure phase is then followed by an extended test phase in which participants typically perform phonetic categorization of the critical sound contrast or a generalization contrast. The visually-guided recalibration paradigm following Bertelson et al. (2003) typically consists of a sequence of multiple short audiovisual exposure blocks, each followed by short audio-only test blocks involving phonetic categorization of the ambiguous stimulus heard during exposure plus two adjacent steps on the critical sound continuum. Importantly, in the visual paradigm, exposure blocks consist of one video of the speaker saying one (nonce) word repeated about eight times. Given this simple psychophysics-type of design, visually-guided recalibration experiments lend themselves much better to a tight control of acoustic cues than the lexical paradigm. However, since during exposure in the visual paradigm the same video is repeated multiple times, listeners do not receive any evidence for allowable variation of the critical sound and potential contexts in which it occurs. This may explain the lack of generalization: In the visual paradigm, listeners may learn only about the exact stimuli they are hearing.

Studies on learning sound contrasts in a second language lend some credibility to such an assumption. In this field, it has been demonstrated that successful learning and generalization to new tokens or speakers depends on high-variability training (e.g., Bradlow, Pisoni, Yamada, & Tohkura, 1997; Sadakata & McQueen, 2013, and references therein). Without variation, generalization is limited (Logan, Lively & Pisoni, 1991; Lively, Logan & Pisoni, 1993). Given that studies showing some generalization of perceptual learning used the lexically-guided recalibration paradigm and the study showing the largest amount of specificity uses the paradigm typical of visually-guided recalibration studies, the question arises whether the low-variability visual paradigm in fact leads to very specific learning. It seems possible that, first, recalibration via visual/lipread context can only occur in paradigms in which participants are exposed to a very limited amount of phonetic variability and, second, generalization is found only with the high-variability lexical paradigm of recalibration.

[1] Note that perceptual learning and selective adaptation have been compared previously and have been shown to work differently (Vroomen, van Linden, de Gelder, & Bertelson, 2007). Norris et al. (2003) used control conditions to rule out that the perceptual-learning effect could be explained by selective adaptation. However, differences between visual vs. lexical context driving the effects could be shared.

1.3. The present study

Experiment 1 examined whether visually-guided recalibration occurs if presented in a high-variability paradigm matching lexically-guided recalibration studies. This issue will speak to the practice of using different perceptual learning paradigms to study phonetic recalibration and its generalization with the aim of studying the units of perception. Experiment 2 then addressed the question about the units of perception. Specifically, it tested whether a recalibrated place of articulation contrast in stops (/p/–/t/) would generalize across manner of articulation (to the nasal contrast /m/–/n/) if tested in a high-variability paradigm using lexically-guided recalibration-as had been used for all previous studies demonstrating generalization. If generalization was found here, then the abstraction of prelexical units such as phonological features as the targets of recalibration would critically depend on variability in the input. If no generalization was found (i.e., replicating results by Reinisch et al., 2014), then this may support various previous suggestions that phonological features may not be the units of perception. Alternative accounts will be discussed.

2. Experiment 1

Experiment 1 set out to address the question whether lexical and visual context indeed trigger the same recalibration process as suggested by van Linden and Vroomen (2007). Van Linden and Vroomen had shown that visually and lexically-guided recalibration produce similar effects when tested in a psychophysics-type of experiment as is typical for visually-guided perceptual learning studies. In the present experiment, visual recalibration was tested in a paradigm similar to studies on lexically-guided recalibration, including one long exposure phase involving 11 critical words spread out over a lexical decision task and including many filler items. If listeners are able to recalibrate a phonetic contrast by means of visual/lipread information in this high-variability paradigm, then this would be evidence that visually-guided recalibration is not restricted to the stimuli experienced during exposure and hence may be used as a simpler variant to study the units of perception (allowing for tight control over the acoustics of the stimuli due to the smaller number of items needed in the typical visual paradigm).

Two important differences have to be noted between the present experiment using lipread information to guide recalibration and the typical lexical paradigm that will be used in Experiment 2. Given that our aim was to disentangle type of guiding information from experimental paradigm, we had to ensure that, despite the lexical decision task used for exposure in the visually-guided experiment, lexical information did not contribute to the disambiguation of the critical sounds. Therefore we opted for the use of minimal word pairs for embedding the critical sounds. Note that in this case the “misperception” of the intended sound as the other category did not change the word’s lexical status or the expected answer in the lexical decision task. Since, however, the same problem would have occurred had we used nonword pairs instead, we decided for the correct answer to be identical between the visual and lexical experiment. The other difference that we could not avoid was that participants in the visually-guided experiment were presented with both members of the minimal pairs, one with ambiguous and the other one with unambiguous audio. Since the list of minimal pairs was close to exhaustive (for details see below), using one set of minimal pairs for the ambiguous sounds and another one for the unambiguous sounds was not feasible. Importantly, during debriefing none of the participants reported noticing similar sounding words.

2.1. Methods

2.1.1. Participants

A total of 55 native speakers of German participated for a small payment. They were recruited from the student population at the University of Munich, Germany.

2.1.2. Materials

77 German words and 77 nonwords that were phonologically legal in German were used as exposure materials. The set of words consisted of 11 minimal word pairs (i.e., 22 words) as critical items and 55 filler words. Of the minimal pairs, one word ended in a labial stop (/b/ or /p/) and the other in an alveolar (/d/ or /t/; see Appendix for a full list of words). Note that German has final consonant devoicing, hence for simplicity, the stops will be referred to as /p/ and /t/, irrespective of their underlying phonological or orthographic form[2]. Except for the word-final position of these items none of the words or nonwords contained the sounds /b/, /p/, /d/, /t/ or the nasals /m/ and /n/ (which was critical for Experiment 2 where the same fillers and nonwords were used). The nonword syllables [ʔaːp]–[ʔaːt][3] were chosen for the phonetic categorization task at test. Nonword pairs rather than additional minimal pairs were chosen since – given the constraints mentioned above – the set of minimal pairs used for exposure was close to exhaustive. Moreover, since 11 critical words per sound were already fewer than in most studies, it was important to maximize the number of exposure trials (note that around 10 trials are considered the minimum number to trigger recalibration; Poellmann, McQueen & Mitterer, 2011).

[2] Note that a recent study has shown that listeners generalize a recalibrated place of articulation contrast across phonologically voiced and voiceless stops in word-final position in German (Reinisch & Mitterer, 2015).

[3] Note that German “vowel-initial” words are typically produced starting with a glottal stop, especially when spoken in isolation, as was the case here (see Mitterer & Reinisch, 2015, for discussion of the phonological status of the glottal stop in German).

All words and nonwords were recorded by a female native speaker of German in a sound-attenuated booth. Digital recordings were made in high-quality audio and video format. The video showed the speaker's head and shoulders in front of a dark gray background. Recordings were blocked by words (critical, filler) and nonwords. Care was taken that the members of the minimal pairs were spoken as similarly as possible in terms of pitch, pitch contour, and speaking rate, to facilitate the manipulation procedure described below.

Using the high-quality audio recordings, 21-step [p]–[t] audio continua (from 0–100% [p]) were created between the minimal word pairs as well as the nonword pairs to be used at test. Using the STRAIGHT algorithm (Kawahara et al., 1999) in Matlab (The MathWorks Inc.), time alignment ensured that only the same types of segments within each word pair were morphed, that is, vowels were morphed with vowels, stops with stops, etc. Whole words rather than only the stops were morphed to ensure that distributed cues such as vowel transitions were ambiguous with respect to place of articulation of the following stop.

To select the most ambiguous sounds to be used for exposure, twenty participants who did not take part in the main experiment categorized the continua in a two-alternative forced choice task (for details on this pretest see Experiment 2). For the minimal pairs the most ambiguous sounds were all close to the middle step of the continua. The average continuum step for 50% labial-responses was at 9.8 out of the range from 0 to 20.

The selected ambiguous audio tokens were then paired with both videos of the minimal pair. In addition, the endpoint tokens of the continua were paired with their respective matching videos. Audio and video-stimuli were aligned with respect to the original audio track that got replaced by the manipulated audio. All videos were cut to show the speaker with neutral mouth position before and after uttering the word. Fade-in and fad-out sequences were implemented over the first and last five frames of each video clip. It was made sure that fade-off started only after the speaker's mouth had returned to a neutral position and hence did not mask the critical word-final sound or lip movement. Videos of the filler words and nonwords were cut in a similar fashion. All video editing was done in Adobe Premiere CS4 software.

To ensure that the visual information would be sufficient for listeners to disambiguate the ambiguous audio tokens in either direction (labial or alveolar sound), 13 participants (who did not participate in the main experiments or in the other pretest mentioned above and with Experiment 2) viewed each critical video 3 times in random order and decided whether the last sound of the word was b/p or d/t. Listeners were informed that both options would form existing German words. They were asked to watch the video, and log their answer by key-press, as soon as the video was replaced by a fixation cross. Listeners responded in line with the disambiguating video on over 90% of the trials on all but one word (*Lot* "plumbline" received only 82% correct /t/-responses). Given this high level of accuracy, and errors being distributed over various items and participants, the stimuli were assessed suitable to trigger learning.

For the phonetic categorization task used at test we selected 6 continuum steps that were equally distributed across the continuum and included the endpoints on either side (from now on referred to as steps 1–6).

2.1.3. Procedure

2.1.3.1. Exposure. Approximately half of the participants were randomly assigned to what will be referred to as the p-ambiguous condition and half were assigned the t-ambiguous condition. Participants in the p-ambiguous condition were presented with the ambiguous audio tokens in videos showing the speaker close her lips. The t-ambiguous group was presented the ambiguous audio tokens in videos without word-final lip closure. The respective other tokens of the minimal pairs were presented in their unambiguous forms with the matching videos. All participants were presented with the same filler words and nonwords.

Participants were seated in a sound-proof booth in front of a computer screen wearing headphones set to a comfortable sound level. Instructions were given in written form as well as orally by the experimenter. On each trial, participants were watching a video of the speaker saying a word or nonword and had to indicate by button press whether the word was an existing German word or not. The videos were presented centered on a black screen with the speaker covering an area of about 7.5 × 10.5 cm. On the lower left of the video listeners saw the letter string *Wort* ("word") and on the lower right *kein Wort* ("not a word") displayed in white font on the black background. This was to remind participants of the assignment of keys (left-right layout) to the response options. Participants were asked to press the number key 1 if they had perceived a word and to press the number key 0 if they had perceived a nonword. Responses could be logged only after the video had faded and was replaced by a white fixation cross in the middle of the screen.

To ensure that listeners were not only listening but also focusing their attention on the videos, eight filler trials (4 words and 4 nonwords) were used as "catch trials". That is, during these trials a small light-green dot appeared on the speaker's upper lip for 300 ms. If listeners saw the dot they were required to press spacebar instead of answering to the word-nonword task. Words and nonwords were presented to participants in pseudorandom order. The experiment started with at least six filler word or nonword trials before a critical word occurred. Every 55 trials participants were allowed to take a self-paced break.

2.1.3.2. Test. Immediately following exposure, all participants completed the same audio-only phonetic categorization task with the [ʔaːp]–[ʔaːt] nonword continuum. They were reminded by means of written instructions that they were going to hear nonwords and their task was to decide by button press whether the nonword ended in b/p or d/t by pressing the 1 (= labial) or 0 (= alveolar) key on the computer keyboard. Response options and their assigned keys were displayed on the screen throughout the trial and were identical across participants. Response options included letters for phonologically voiced and voiceless stops (i.e., [b/p] and [d/t]) since both graphemes can represent phonetically voiceless stops in word-final position. This also captures the fact that critical words during exposure ended in a mix of phonologically (and orthographically coded) voiced and voiceless stops (even though participants were likely not aware of the relation between critical words during exposure and test). Participants were informed that their response was logged by seeing the response option move about half a centimeter upward on the screen where it stayed for 400 ms. The next trial started automatically 500 ms later. Each of the 6 steps of the continuum was presented 14 times for a total of 84 trials. Trials were presented in random order with the restrictions that all steps of the continuum were presented before they were repeated. Halfway through the test phase participants were allowed to take a self-paced break. Exposure and test phase were implemented running E-Prime software (Psychology Tools Inc.). Completing the whole experiment took approximately 20 min.

2.2. Results

Acceptance rates of the critical items as real words during exposure was high, with an average acceptance of 93.6% for participants in the p-ambiguous group and 91.1% for participants in the t-ambiguous group. Correct performance on catch trials was 87.5%. Since during debriefing some participants reported that they had noticed the dots but continued performing lexical decision rather than pressing spacebar; we decided to include all participants in our analyses[4]. If anything, participants' failure to watch the videos would work against us in finding an effect of recalibration.

Fig. 1 shows the results of the phonetic categorization task in Experiment 1. As expected, listeners gave more p-responses towards the p-end of the continuum, and participants in the p-ambiguous group gave more p-responses (solid line) than participants in the t-ambiguous group (dashed lines). However, as the error bars suggest, there was quite some variation between participants. This observation is reflected in the statistical analyses.

For analysis, we used generalized linear-mixed effects models (lme4 package 1.1–7 in *R*-statistics Version 3.2.0) with a logistic linking function to account for the categorical dependent variable (1 = labial response, 0 = alveolar response). The predictor variables were Step and ExposureCondition. Only responses to the four middle steps of the continuum were analyzed since for the acoustically unambiguous endpoints of the continua it is unclear whether recalibration should be expected. Step was centered on zero and a negative regression weight was expected (fewer p-responses with more t-like stimuli). ExposureCondition was contrast coded such that exposure to ambiguous [p] was coded as 0.5 and exposure to ambiguous [t] was coded -0.5. With this coding, we expect a positive regression weight for ExposureCondition (more "successes", i.e., p-responses after exposure to an ambiguous labial). An interaction between ExposureCondition and Step was not specified since it did not improve the model's fit. The model included a random intercept for Participant and a random slope for Step over Participants. A random slope for ExposureCondition was not included because it is a between-participant variable (Barr, Levy, Scheepers, & Tily, 2013).

Results showed an effect of Step ($b_{(\text{Step})} = -0.99$, $z = -10.3$, $p < .001$) indicating more p-responses the more *p*-like the step of the continuum. Importantly, there was an effect of ExposureCondition ($b_{(\text{Exposure Condition})} = 0.78$, $z = 2.24$, $p < .05$). Participants in the p-ambiguous condition gave more labial responses than participants in the t-ambiguous condition.

2.3. Discussion

Experiment 1 set out to test whether listeners recalibrate a place of articulation contrast by means of visual (lipread) context even when the critical disambiguating information was presented in a set of 11 different words within a 154 trial lexical decision exposure phase – an experimental setup that previously had been used to study lexically-guided recalibration. Results showed that indeed listeners were able to extract the relevant visual information within the high-variability context and recalibrate the place of articulation contrast. Visually-guided recalibration is hence not specific to the tokens seen during exposure, which could have been hypothesized based on the traditional psychophysics-type paradigm used for visual recalibration studies (i.e., following Bertelson et al., 2003). Moreover it provides a first hint that the specificity of phonetic recalibration that has been demonstrated with a visually-guided recalibration paradigm is not a result of the exposure modality (Reinisch et al., 2014).

3. Experiment 2

Experiment 1 demonstrated visually-guided recalibration in a high-variability situation where only a small number of critical items is embedded in a larger set of words and nonwords. The purpose of Experiment 2 was twofold. First, it replicated Experiment 1 with the critical information for recalibration being provided via the lexicon. The comparison between visually-guided and lexically-guided recalibration will further qualify the results obtained in Experiment 1. Second, as discussed in the introduction, Experiment 2 set out to test the main question of the present study, that is, whether listeners would generalize a recalibrated place of articulation contrast in stops (/p/–/t/) across manner of articulation to nasals (/m/–/n/). If listeners generalized the place of articulation contrast here, then this would suggest that variability in the exposure stimuli is essential to finding generalization. This issue is well-known from the literature on second language sound acquisition but has yet not received much attention with regard to constraints on perceptual recalibration. The outcome of the generalization trials will help evaluate previous suggestions on the units that are targeted by perceptual learning. Accounts suggesting phonological features, and specifically a feature for place of articulation, would necessarily predict generalization here. If acoustic similarity between target contrasts or context sensitivity has a role to play then we may not expect generalization. The latter prediction would replicate a number of previous findings (e.g., Reinisch et al., 2014; Schuhman, 2014).

[4] 38 participants detected the dots 100% of the time. Two participants did not respond to the dots at all, three further participants missed more than two dots.

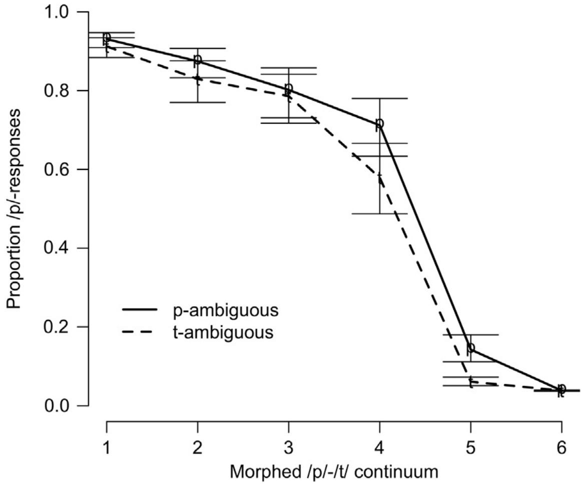

Fig. 1. Categorization results, that is, proportion labial responses for the test phase in Experiment 1. The solid lines indicate responses by the p-ambiguous group, dashed lines the t-ambiguous group. Error bars are calculated in logistic space (matching the analyses) and transformed to proportions for the plot.

3.1. Methods

3.1.1. Participants

A total of 102 native speakers of German participated for a small payment. They were recruited from the student population at the University of Munich, Germany. None had participated in Experiment 1 or in any of the pretests.

3.1.2. Materials

The same materials as in Experiment 1 were used with the exception that the critical minimal pairs were replaced by 22 German words, half ending in [p] half in [t], that did not form minimal pairs with the stops exchanged (see Appendix). They adhered to all criteria discussed before (i.e., not containing the sounds /p/, /b/, /t/, /d/, /m/ or /n/ in other than the critical word-final position). In addition, the nonwords [ʔaːm] and [ʔaːn] were chosen for testing generalization across manner of articulation. All items were recorded audiovisually in the same recording session as the other materials. In contrast to Experiment 1, here only the high-quality audio tracks were used, for critical items and fillers alike. All critical words were recorded in their correct form and with the word-final stops exchanged. For example, the word *Fahrrad* "bike" was also recorded as *Fahrra<b>* and *Aufschub* "delay" was also recorded as *Aufschu<d>*. The correct and "exchanged" versions of the critical words as well as the [ʔaːm]–[ʔaːn] nonword continuum were morphed into 21-step continua using the same procedure described in Experiment 1.

To assess the most ambiguous sounds of the critical words to be used for exposure, 20 participants who did not take part in the main experiments categorized the last sounds of all word–nonword continua as well as the minimal pairs used in Experiment 1 (steps 0, 2, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 18, 20 of the original 21-step continuum). To select the most ambiguous steps of the critical items, sigmoid functions were fit over categorization responses. Continuum steps at which answers were closest to 50% labial responses were taken as a starting point for further qualitative assessment. Two trained phoneticians listened to these steps to determine whether the tokens sounded "normal", that is, whether in the lexical context the ambiguous stop was recognizable as the intended phone. If a token sounded "wrong" and thus like a nonword, adjacent steps along the continuum toward the intended endpoint were considered until consensus was reached that the words would likely be identified as intended. Note that it is crucial that the critical words could be identified as existing words in order to trigger learning. Had we chosen words that were not sufficiently ambiguous, learning could not have occurred. For the generalization continuum an equally spaced 6-step [ʔaːm]–[ʔaːn] continuum was chosen spanning the two continuum endpoints; similar to the [ʔaːp]–[ʔaːt] test continuum that had already been used in Experiment 1.

3.1.3. Design and procedure

The design and procedure were identical to Experiment 1 with the only exception that listeners received disambiguation of the acoustically ambiguous sound by means of lexical information. That is, participants in the p-ambiguous condition were presented the ambiguous sound in words where only /p/ formed a real word. Words ending in /t/ were presented with unambiguous sounds (i.e., the endpoints of the continua). The opposite was the case for the t-ambiguous group. All participants were presented with the same filler words and nonwords. Critically, as in most previous lexcially-guided recalibration studies, all stimuli were presented audio only. Listeners performed the lexical decision task with only their response options (*Wort* "word", *kein Wort* "not a word") and the respective key labels (1 and 0) displayed on the screen.

The test phase consisted of two parts the order of which was counterbalanced across participants. For one group of participants ($N=49$) the first part was identical to Experiment 1. Participants categorized the 6-step [ʔaːp]–[ʔaːt] nonword continuum 14 times in random order. This set of responses will be used for comparison to the results from visually-guided recalibration in Experiment 1. Following these 84 trials testing the basic recalibration effect, listeners immediately proceeded to the generalization phase where they were asked to categorize another 84 trials, now of the [ʔaːm]–[ʔaːn] generalization continuum (again 6 steps repeated 14 times). The other group (53 participants) first categorized the [ʔaːm]–[ʔaːn] generalization continuum followed by the [ʔaːp]–[ʔaːt] continuum that tested the basic recalibration effect. In both generalization groups about half of the participants were in the p-ambiguous condition, and half in the t-ambiguous condition.

3.2. Results

Given that previous studies using lexical information to guide learning had shown that about ten critical exposure trials are necessary in order to trigger learning (Poellmann et al., 2011), participants in our study were required to reject not more than three critical words as nonwords (note that this would result in a minimum of only 8 accepted critical trials, however, since the total number of critical words was only 11, we decided to allow for some misses). Nine participants were excluded for not meeting this criterion. In the final set of participants the average acceptance rate of critical items in the p-ambiguous condition was 95.8% and in the t-ambiguous condition 94.2%.

3.2.1. Generalization across manner of articulation in lexically-guided recalibration

Fig. 2 illustrates the categorization results. The left panels of Fig. 2 show results from the first part of the test phase that immediately followed exposure; the right panels show results for the block that followed. The upper panels show results for participants who first categorized the [ʔaːp]–[ʔaːt] continuum followed by the [ʔaːm]–[ʔaːm] continuum; the lower panels show results for the group that categorized [ʔaːm]–[ʔaːm] first. As can be seen in the figure, we found the basic recalibration effect for the [ʔaːp]–[ʔaːt] continuum such that participants in the p-ambiguous group gave more p-responses than participants in the t-ambiguous group. This appears independent of whether the [ʔaːp]–[ʔaːt] continuum was tested first or second. As for the [ʔaːm]–[ʔaːn] generalization continuum, neither block order appeared to lead to substantial differences in proportion p-responses between participants in the p-ambiguous and t-ambiguous groups. Only Step 4 of the [ʔaːm]–[ʔaːn] continuum that immediately followed exposure showed some difference between exposure conditions. However, even here error bars suggest large variation between participants.

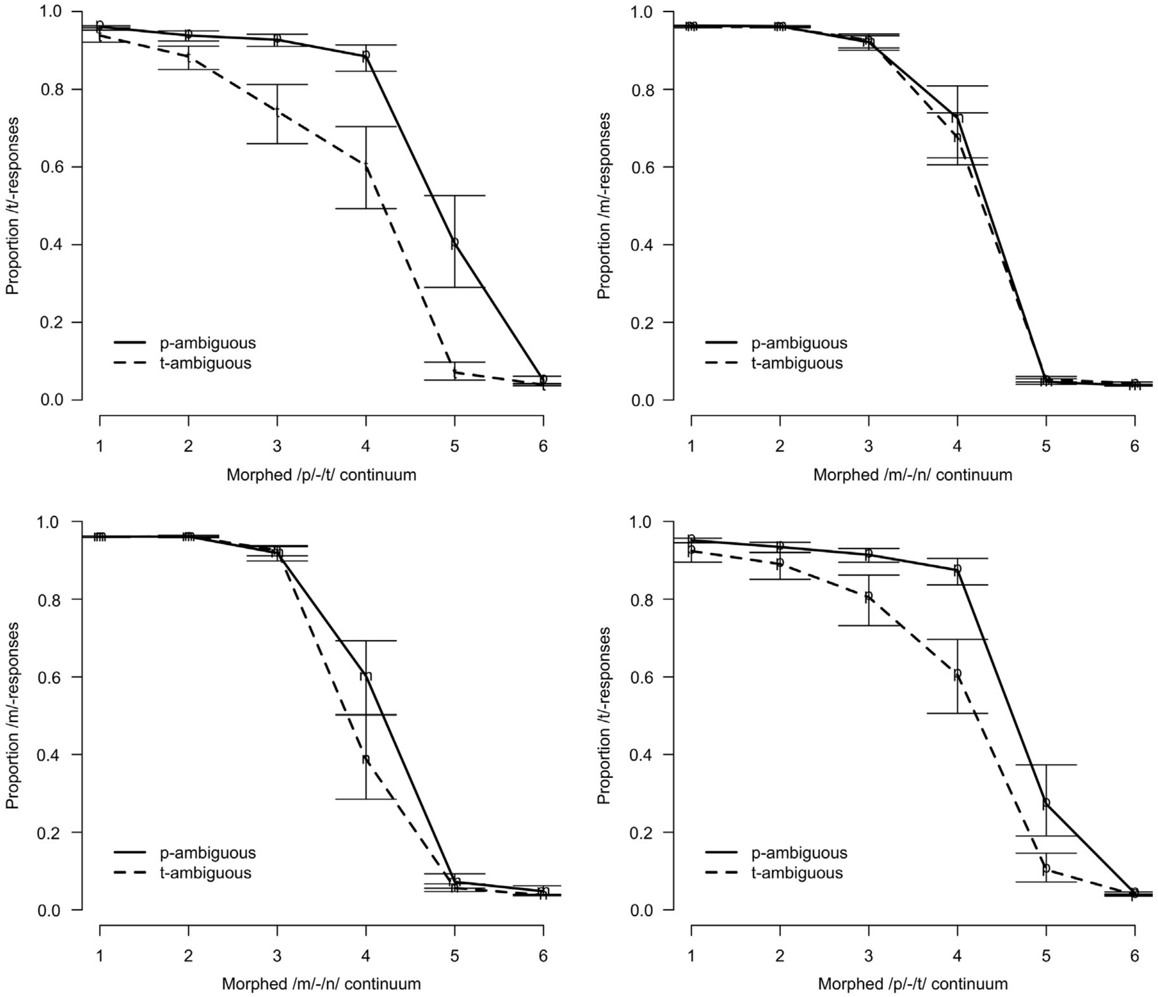

Fig. 2. Categorization results, that is, proportion labial responses for the test phase in Experiment 2. The left upper and right lower panels show results for the stop contrast testing the basic recalibration effect; the upper right and lower left panels show results for the nasal generalization contrast. The solid lines indicate responses by the p-ambiguous group, dashed lines the t-ambiguous group. Error bars are calculated in logistic space (matching the analyses) and transformed to proportions for the plot.

Analyses using generalized linear-mixed effects models confirmed these visual impressions. Response was coded as labial = 1 and alveolar = 0 (for both, the [ʔaːp]–[ʔaːt] continuum and the [ʔaːm]–[ʔaːn] continuum). Step was centered on zero, with only the middle four steps entering analyses since for the acoustically unambiguous endpoints of the continua that received close to 0 and 100% labial responses in all conditions an additional effect of recalibration is unlikely to occur. ExposureCondition was contrast coded with ambiguous = /p/ mapped on 0.5. In addition the factors Block (first, directly following exposure coded as 0.5; second coded −0.5), Contrast (stops coded as 0.5, nasals as −0.5) and their interactions with ExposureCondition and with each other were entered as fixed factors. Given this coding, positive regression weights can be expected if more labial responses are given by the p-ambiguous group, for the first block and the [ʔaːp]–[ʔaːt] contrast. Participant was entered as a random factor with random slopes of Block and Contrast over participants. Table 1 shows the results.

There was a significant effect of Step suggesting that more labial responses were given the closer the sound was to the labial end point of the continua. The effects of ExposureCondition (more labial responses for the p-ambiguous group) and Contrast (more labial responses for the [ʔaːp]–[ʔaːt] contrast) were modulated by an interaction between these factors. That is, recalibration did occur but was stronger for the [ʔaːp]–[ʔaːt] than the [ʔaːm]–[ʔaːn] contrast. Critically, this difference was not modulated by Block, that is, whether the given contrast was categorized first or second during test did not influence the results.

To follow up on the interaction between ExposureCondition and Contrast, that is, to test whether recalibration can be found for stop and nasal contrasts, separate generalized mixed models were run for the two Contrasts but including all other factors as described above. Results are given in Table 2. For the stop contrast significant effects of Step and ExposureCondition were found, both in the expected direction. For the nasal contrast, however, only the effect of Step was significant. That is, a recalibrated place of articulation contrast is not generalized from stops to nasals. The lack of effect for Block suggests that the results were stable over the whole test phase.

Fig. 2, left bottom panel, suggests that there may be a generalization effect to the nasal contrast on Step 4 specifically. Therefore, we tested the effect of ExposureCondition for this one step only. Even though this test is anti-conservative (since it is based on visual inspection of the data), it still did not reveal a significant effect ($b = 0.94$, $z = 1.42$, $p > 0.1$).

3.2.2. *Lexically-guided vs. visually-guided recalibration*

The other goal of Experiment 2 was to provide a comparison to the visually-guided recalibration effect in Experiment 1. To do so, a subset of data from Experiment 2 was selected such as to match the condition from Experiment 1. That is, only categorization responses from the [ʔaːp]–[ʔaːt] continuum were used from the group of participants ($N = 47$) who responded to this continuum immediately following exposure (what is illustrated in the upper right panel of Fig. 2). The cross-experiment comparison had response as the dependent variable (/p/ = 1, /t/ = 0), and Step (centered on zero, only the middle four steps were entered in the analysis for the reasons given above), ExposureCondition (p-ambiguous = 0.5, t-ambiguous = 0.5), Modality (lexical = 0.5, visual = −0.5), and the interactions between the latter two factors as fixed factors. Participant was entered as a random factor with a random slope for Step over participants. This was the maximum random effects structure (Barr et al., 2013). Table 3 shows the results.

Significant effects of ExposureCondition, Modality, and Step suggest that across both experiments listeners gave more p-responses the more p-like the stimulus was (Step), participants gave more p-responses in the lexical than the visual experiment (Modality) and participants in the p-ambiguous exposure group gave more p-responses than participants in the t-ambiguous group (ExposureCondition). The effect of ExposureCondition was modulated by an interaction with Modality suggesting that the recalibration effect was stronger for the lexical than the visual exposure (i.e., stronger in Experiment 2 than in Experiment 1; but note that the effect of ExposureCondition was significant in Experiment 1).

3.3. *Discussion*

The purpose of Experiment 2 was twofold. Fist, the analysis of a subset of the data allowed us to further interpret the results from Experiment 1. Experiment 1 had demonstrated that listeners can recalibrate their /p/–/t/ categories through visual information even if only eleven critical words occur in a 154-word lexical decision task, as is typical for lexically-guided recalibration experiments. The comparison of these results to a similar data set from Experiment 2 where lexical information guided recalibration revealed that the recalibration effect was stronger if guided by lexical rather than visual information. This would suggest that lexical context effects are somewhat stronger than visual context effects in such a high-variability situation. Note that this is in contrast to a previous study using

Table 1
Results of the generalized mixed model comparing the recalibration effects for stops (basic recalibration effect) and nasals (generalization) in Experiment 2.

	B	*Z*	*p*
Intercept	1.54	0.48	<.001
ExposureCondition	1.29	4.02	<.001
Contrast	0.56	1.95	= .05
Block	−0.07	−0.26	.79
Step	−1.17	−47.72	<.001
ExposureCondition:Contrast	2.15	3.75	<.001
ExposureCondition:Block	0.59	1.03	.30
Contrast:Block	0.34	0.53	.59
ExposureCondition:Contrast:Block	0.39	0.31	.75

Table 2
Separate analyses for the stop and nasal continua.

	Stops			Nasals		
	b	*z*	*P*	*B*	*z*	*p*
Intercept	1.45	6.52	<.001	1.84	8.93	<.001
ExposureCondition	1.92	4.33	<.001	0.29	0.75	.46
Block	0.09	0.21	.84	−0.36	−0.92	.36
Step	−0.84	−31.37	<.001	−1.78	−31.45	<.001
ExposureCondition:Block	0.63	0.71	.48	0.56	0.71	.48

Table 3
Results of the generalized mixed model comparing the recalibration effects following visual exposure in Experiment 1 and lexical exposure in Experiment 2.

	b	*Z*	*p*
Intercept	1.64	6.01	<.001
ExposureCondition	1.51	4.40	<.001
Modality	0.73	2.13	<.05
Step	−1.06	−14.7	<.001
ExposureCondition:Modality	1.37	2.00	<.05

a psychophysics-type experiment (van Linden & Vroomen, 2007). There the effect of visual context appeared stronger than lexical context, highlighting the importance of the experimental setup and listening situation. One possible explanation for our smaller effect of visual than lexical context might be that perceivers are less able to focus their attention on the relevant part of the visual signal in a high-variability paradigm, especially with a lexical decision task. As briefly discussed above, in the visual-context experiment, the answer for lexical decision would have been the same no matter which of the two sounds had been perceived – a confound that could not be avoided. This lowering of attention to the visual information could explain the difference in strength of the effects for lexical and visual recalibration as well as with the previous study that compared the two types of context. There is evidence to suggest that audio-visual integration in speech perception is dependent on the allocation of attention (e.g., Alsius, Navarra, & Soto-Faraco, 2007; Navarra, Alsius, Soto-Faraco, & Spence, 2010). Note however, that even under presumably lowered attention, visual context in Experiment 1 did trigger phonetic recalibration.

The second question that Experiment 2 addressed was whether recalibration of a place of articulation contrast would generalize across manner of articulation if presented in a high-variability lexical-recalibration paradigm. Here the answer is clear: generalization does not occur. Even for the group of participants where generalization was tested immediately following exposure, no difference between exposure groups could be found (not even when trying to cherry-pick and analyze only Step 4 for which Fig. 1 suggests a small difference between exposure groups). This “null” effect is unlikely due to a lack of power. In the most frequently cited study where generalization across sound contrasts has been found (Kraljic & Samuel, 2006), the effect in the generalization condition (/b/–/p/) was slightly but not significantly larger than in the basic recalibration condition (where exposure words contained ambiguous sounds between /d/ and /t/). Based on this equal effect between basic recalibration and generalization conditions in this previous study (and a study by Jesse & McQueen, 2011, that found a similar effect for recalibration and generalization of a fricative contrast across word position), we can estimate the expected effect size for generalization in our study to also equal the effect in the baseline condition. Since the average logOdds of labial responses in in our basic recalibration condition (i.e., exposure and test on stops) is a large one (mean difference = 1.275, pooled SD = 1.386, leading to *d* = 0.92), we have a power of more 99.2% to detect an effect in the generalization condition (based on the function *pwr.t2n.test* from the R package *pwr*).

Note also that the interaction provides the critical result of showing a significant difference between the learning effects of the two sound contrasts (i.e., stops vs. nasals). Importantly, group differences for the stop contrast were found independent of a delay in testing. This suggests that even with only 11 critical words during exposure, listeners can use lexical information to robustly recalibrate a place of articulation contrast, but generalization across manner of articulation does not occur. Implications of differences between modalities and sound contrasts (basic recalibration vs. generalization) will be discussed in Section 4.

4. General discussion

The present study tested two questions about the role of input variability and specificity of perceptual recalibration: First, what role does variability in the input play in the occurrence of recalibration in different perceptual learning paradigms (i.e., visual vs. lexical)? Second, what role does variability in the input play in the occurrence of *generalization* of perceptual learning? Specifically, we asked whether listeners would generalize a recalibrated place of articulation contrast in stops ([p]–[t]) across manner of articulation to nasals ([m]–[n]). This test of generalization informs us about the prelexical units that are targeted by recalibration and contributes to a current debate about whether abstract features should be taken as units for speech processing (e.g., Embick & Poeppel, 2014). The lack of generalization in the present experiment (and the significant interaction between ExposureCondition and Contrast) is in line with other recent perceptual learning studies that countered feature accounts and supported units such as allophones as the most likely prelexical units (e.g., Mitterer et al., 2013; Reinisch et al., 2014). Before we turn to this issue in more detail we discuss the first question about exposure modality.

Experiment 1 speaks to the mechanism of perceptual recalibration, showing that listeners use not only top-down lexical information but also bottom-up visual/lipread information to recalibrate phonetic categories. Recalibration occurs even if the relevant disambiguating cues are not as salient as in a typical setup for visually-guided recalibration where the same exposure video is presented several times with no or little interruption. We showed that visually-guided recalibration is effective even when the majority of tokens during exposure is fully congruent between the audio and video signal. In our experiment, only 11 out of 154 tokens were edited such that the auditory input was ambiguous and hence not fully matched the visual input. That is, visual recalibration is not specific to the tokens presented during exposure. With regard to our second main question about generalization of recalibrated place of articulation contrasts we can therefore assume that the lack of generalization in a previous study using visual recalibration (i.e., Reinisch et al., 2014) cannot be solely due to the use of visually guided exposure.

This conclusion is reinforced by the results of Experiment 2, which tested whether perceptual learning about place of articulation generalizes from stops to nasals. The results of Experiment 2 revealed strong effects of recalibration that were stable even following a delay in testing (i.e., for the group of participants that categorized the generalization contrast first). Generalization across manner of articulation from stops to nasals could not be found. This seems to clash with the previously found generalization for voicing in Kraljic and Samuel (2006). However, one way to resolve this apparent conflict is to take into account the acoustic similarity of the critical cues in the case of Kraljic and Samuel, that is, a long vs. short closure duration and the presence vs. absence of aspiration. Note that the need for acoustic similarity between exposure and generalization contrasts has also been proposed with regard to generalization of recalibration across speakers (Kraljic & Samuel, 2007; Reinisch & Holt, 2014) and has recently been confirmed by another study testing generalization of a recalibrated place of articulation contrast across manner of articulation (Schuhman, 2014). Specifically, Schuhman showed that native English listeners generalize a recalibrated /f/–/s/ contrast to their voiced counterparts /v/–/z/ but not to the closest stop contrast /p/–/t/. Arguably the stop contrast may not have matching place features in all types of linguistic theories that assume features (especially articulatory phonology), but again the feature “voice” allowed for generalization. As such, the data of previous studies that do show generalization are not in conflict with feature accounts but fall short as strong evidence *for* abstract phonological features. Abstract phonological features would require generalization across conditions with more diverse cues such as different phonetic contexts as, for instance, tested in Reinisch et al. (2014), and also across manner of articulation as tested in the present experiments. The failure to find generalization of phonological features here resonates with the “old” problem of abstract features (Klatt, 1989): Their acoustic implementation varies too strongly to make them useful in perception.

This leaves alternative accounts, for example, that perceptual recalibration is to some extent context specific and/or may depend of the distribution of the relevant cues in the adjacent segments. Context-independence as evidenced by generalization across syllable-positions (/f/–/s/ contrast) and generalization across place of articulation (duration of stop closure and burst/VOT for the voicing distinction in stops) has been shown for sound contrasts where the relevant information was relatively "constrained to the critical sounds", that is, the main cue to the contrast was conveyed on the critical segments (as opposed to neighboring segments) such as the frequency distribution of spectral energy in fricatives and durational properties of the stops. Contrasts for which generalization of recalibration could not be found had more distributed cues: liquids in the study by Mitterer et al. (2013) and formant transitions as cues to place of articulation in stops and nasals in Reinisch et al. (2014) and in the present study. Note that in Reinisch et al. (2014) even physically identical cues by means of formant transitions in “aba”–“ada” vs. “ama”–“ana” did not allow for generalization (in either direction).

From a learning perspective, a further comparison can be drawn to the acquisition of second language sound categories. That is, if learners of a second language are trained on a new sound contrast, would they generalize their newly acquired ability to discriminate this contrast across sound contrasts differing in manner of articulation? Similarly, does the presence of a certain place of articulation contrast in a learner's first language help them to learn a similar second language contrast with the same place distinction but a different manner of articulation? With regard to the first issue, native Dutch listeners seem able to generalize a trained duration distinction between Japanese singleton and geminate fricatives to stops and affricates (Sadakata & McQueen, 2013). Pajak (2010) found evidence of generalization from long and short vowels to singleton and geminate consonants. This, together with the recalibration results by Kraljic and Samuel (2006) suggests that duration cues can flexibly be transferred across place and manner of articulation. The latter further demonstrates that generalization across manner of articulation per se is not problematic. However, the type of generalization condition tested in the present study (i.e., generalization of a recalibrated place of articulation contrast in stops across manner of articulation) has led to similar results in a number of studies of second language contrasts. These studies (see below for details) tested whether the presence of a certain place of articulation contrast in a learner’s first language helps them to perceive or learn a similar second language contrast with the same place distinction but a different manner of articulation. For example, Polka (1992) tested whether native speakers of Farsi are better at discriminating a velar-uvular place contrast for Salish glottalized stops than native English listeners. Farsi has a phonemic velar-uvular place contrast in voiced stops whereas English does not have any velar-uvular place distinction. However, the expected advantage for Farsi listeners was not found. Moreover, Gogoi (2010) failed to find an advantage for Bengali-English bilinguals over Spanish-English bilinguals in learning to distinguish a number of Malayalam dental-retroflex contrasts (e.g., in nasals, laterals) despite the presence of such a place distinction for stops in Bengali but not in Spanish. Similarly, English speakers do not seem able to make use of their native tense-lax feature in making a distinction between non-native front-rounded vowels (Ettlinger & Johnson, 2009).

So how can these findings be resolved with regard to the question about the units of perception? One possible solution would be to assume that temporally defined cues such as aspiration as a cue for voicing *are* pre-lexical units, but place of articulation or manner of articulation is not. The suggestion here is that pre-lexical units may not be constrained to one grain size, but rather to acoustically distinct properties of the input, which in some cases may be small (in the case of aspiration) but sometimes need to be large (disyllables – e.g., in the case of distributed cues to place of articulation) to achieve some degree of context independence (Poellmann, Bosker, McQueen, & Mitterer, 2014). Current evidence has started to converge on certain patterns of generalization of perceptual recalibration or the lack thereof. The replication of the latter for different sound contrasts with different paradigms and in different labs (e.g., Mitterer et al., 2013; Reinisch et al., 2014; Schuhman, 2014) also suggests that it cannot be discarded as a simple null effect anymore. Additional research on different sound contrasts, and different acoustic implementations of the same contrast in different languages will further help to understand the patterns of generalization and the pre-lexical units that can support such suggestions.

In summary, the present results showed that listeners recalibrate a place of articulation contrast in stops by means of visual/lipread and lexical information. This learning, however, did not generalize across manner of articulation to a nasal contrast. Results are not compatible with the concept of abstract phonological features that would predict generalization here. Similarly, accounts of speech perception that make articulatory features central are challenged by these findings. Rather the results are in line with previous studies suggesting that the grain-size of prelexical units used during perception may depend on the nature of cues that are involved. This might include smaller (aspiration) or larger units (diphones) than phonemes depending on their distribution in the speech signal.

Acknowldegments

We would like to thank Klaus Jänsch for technical assistance with the video recordings, Veronika Wimmer for help with video editing and with running the experiments, and Marthese Borg for comments on a previous version of the paper. This research was supported by a grant from the Alexander von Humboldt Foundation to the first author. Parts of the results were presented at the 2014 meeting “Architectures and Mechanisms for Language Processing” (AMLaP), September 2014, Edinburgh, UK.

Appendix

Words used in Experiments 1 and 2. Frequency measures are taken from Wortschatz Leipzig (http://wortschatz.uni-leipzig.de/ -last viewed 22 May 2015). Note that due to the constraints on the word sets explained in the text word pairs were not matched on frequency. For the audiovisual items the higher frequency of the t-items may have been of an advantage since visual /t/ is confusable with many more sounds than visual /p/ which is clearly visible due to the lip closure and hence confusable only with a small number of other labial sounds.

Lexical items (frequency) *translation*		Visual items (frequency) *translation*	
b/p	d/t	b/p	d/t
Aufschub (899) *delay*	Fahrrad (8027) *bike*	Alp (698) *alp*	alt (42035) *old*
Erwerb (3451) *acquisition*	gelehrt (1592) *scholarly*	gelb (2354) *yellow*	Geld (155526) *money*
Galopp (301) *gallop*	Salat (3384) *lettuce*	Grab (4742) *grave*	Grad (36599) *degree*
grob (2605) *rough*	Eid (957) *oath*	halb (8740) *half*	Halt (71483) *stop*
Hieb (141) *hit/blow*	fad (372) *bored*	Kalb (872) *calf*	kalt (9509) *cold*
Klub (7821) *club*	Kleid (3739) *dress*	Korb (3059) *bask*et	Kord (19) *cord*
Raub (2034) *robbery*	Fett (4561) *fat*	Laub (1892) *foliage*	laut (111764) *noisy*
schlapp (643) *flabby*	Flut (3852) *flood*	Leib (5437) *body*	Leid (6384) *suffering*
Schub (2038) *push*	rot (6159) *red*	lieb (3138) *dear*	Lied (10932) *song*
Sirup (279) *sirup*	akut (1404) *acute*	Lob (8274) *praise*	Lot (806) *plumb-line*
Urlaub (21386) *holiday*	Klugheit (415) *wisdom*	Weib (360) *woman (old, possibly negative term)*	weit (83853) *wide*

References

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research*, *183*, 399–404.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. (2013). Random-effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk after effect. *Psychological Science*, *14*, 592–597.

Bradlow, A., Pisoni, D., Yamada, R. A., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*, 2299–2310.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NJ: Harper & Row.

Cornell, S. A., Lahiri, A., & Eulitz, C. (2013). Inequality across consonantal contrasts in speech perception: evidence from mismatch negativity. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 757–772.

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.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, *19*, 381–385.

Embick, D., & Poeppel, D. (2014). Towards a computational(ist) neurobiology of language: correlational, integrated and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, *30*, 1–10.

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

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin and Review*, *13*, 361–377.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Gogoi, D. V. (2010). *Acquisition of novel perceptual categories in a third language: the role of metalinguistic awareness and feature generalization*. Gainsville, FL, USA: University of Florida.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

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). Pirrot Dudweiler, Germany

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics*, *31*, 305–320.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin Review*, *18*, 943–950.

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.

Klatt, D. (1989). Review of selected models of speech perception. In W. D. Marslen- Wilson (Ed.), *Lexical representation and process* (pp. 169–226). Cambridge, MA: MIT Press.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*, 148–203.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, *13*, 262–268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*, 1–15.

Lahiri, A., & Reetz, H. (2002). Underspecified recognition. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology*, *7* (pp. 637–676). Berlin: Mouton de Gruyter.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: a first report. *The Journal of the Acoustical Society of America*, *89*, 874–886.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: the role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242–1255.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological Review*, *101*, 653–675.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113–1126.

Mitterer, H. (2006). On the causes of compensation for coarticulation: evidence for phonological mediation. *Perception Psychophysics*, *68*, 1227–1240.

Mitterer, H. (2007) Top-down effects on compensation for coarticulation are not replicable. In: *Proceedings of interspeech 2007* (pp. 1601–1604) Causal Productions, Adelaide

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.

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: evidence from a learning paradigm. *Cognitive Science*, *35*, 184–197.

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

Navarra, J., Alsius, A., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion*, *11*, 4–11.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Pajak B. (2010). Perceptual advantage from generalized linguistic knowledge. In: S. Ohlsson, R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 369–374). Austin, TX

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* [ICPhS XVII] (pp. 1618–1621). Department of Chinese, Translation and Linguistics, City University of Hong Kong: Hong Kong

Poellmann, K., Bosker, H. R., McQueen, J. M., & Mitterer, H. (2014). Perceptual adaptation to segmental and syllabic reductions in continuous spoken Dutch. *Journal of Phonetics*, *46*, 101–127.

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Proceedings of the Royal Society of London*

Polka, L. (1992). Characterizing the influence of native language experience on adult speech perception. *Perception Psychophysics*, *52*, 37–52.

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*, 539–555.

Reinisch, E. & Mitterer, H. (2015). Perceptual learning in speech is phonetic, not phonological: evidence from final consonant devoicing. *Proceedings of the18th International Congress of Phonetic Sciences*. Glasgow, UK.

Reinisch, E., Wozny, D., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: what are the categories?. *Journal of Phonetics*, *45*, 91–105.

Roon, K. D., & Gafos, A. I. (2014). Perceptuo-motor effects of response-distractor compatibility in speech: beyond phonemic identity. *Psychonomic Bulletin Review*, *22*, 242–250.

Sadakata, M., & McQueen, J. M. (2013). High stimulus variability in nonnative speech learning supports formation of abstract categories: evidence from Japanese geminates. *Journal of the Acoustical Society of America*, *134*, 1324–1335.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, Psychophysics*, *71*, 1207–1218.

Samuel, A. G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 1479–1490.

Scharinger, M., Merickel, J., Riley, J., & Idsardi, W. J. (2011). Neuromagnetic evidence for a featural distinction of English consonants: sensor- and source-space data. *Brain and Language*, *116*, 71–82.

Schuhman, K. S. (2014). *Perceptual learning in second language learners*. Stony Brook University.

Sjerps, M. J., & McQueen, J. M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 195–211.

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1483–1494.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: contrasting build-up courses. *Neuropsychologia*, *45*, 572–577.