Journal of Phonetics 45 (2014) 91–105

Contents lists available at ScienceDirect

Journal of Phonetics

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

Phonetic category recalibration: What are the categories?

Eva Reinisch [a,b,*], David R. Wozny [b], Holger Mitterer [c], Lori L. Holt [b]

[a] Department of Phonetics and Speech Processing, Ludwig Maximilian University Munich, Schellingstr. 3, 80799 Munich, Germany
[b] Department of Psychology, and Center for the Neural Basis of Cognition, Carnegie Mellon University, USA
[c] Department of Cognitive Science, University of Malta, Malta

ARTICLE INFO

Article history:
Received 7 May 2013
Received in revised form
22 April 2014
Accepted 24 April 2014
Available online 17 May 2014

Keywords:
Speech perception
Perceptual learning
Prelexical processing
Audiovisual processing

ABSTRACT

Listeners use lexical or visual context information to recalibrate auditory speech perception. After hearing an ambiguous auditory stimulus between /aba/ and /ada/ coupled with a clear visual stimulus (e.g., lip closure in /aba/), an ambiguous auditory-only stimulus is perceived in line with the previously seen visual stimulus. What remains unclear, however, is what exactly listeners are recalibrating: phonemes, phone sequences, or acoustic cues. To address this question we tested generalization of visually-guided auditory recalibration to (1) the same phoneme contrast cued differently (i.e., /aba/-/ada/ vs. /ibi/-/idi/ where the main cues are formant transitions in the vowels vs. burst and frication of the obstruent), (2) a different phoneme contrast cued identically (/aba/-/ada/ vs. /ama/-/ana/ both cued by formant transitions in the vowels), and (3) the same phoneme contrast with the same cues in a different acoustic context (/aba/-/ada/ vs. /ubu/-/udu/). Whereas recalibration was robust for all recalibration control trials, no generalization was found in any of the experiments. This suggests that perceptual recalibration may be more specific than previously thought as it appears to be restricted to the phoneme category experienced during exposure as well as to the specific manipulated acoustic cues. We suggest that recalibration affects context-dependent sub-lexical units.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Over the last decades a large body of research has been accumulated investigating how listeners deal with understanding noncanonical pronunciation variants of new speakers they have not heard before (see, e.g., Samuel & Kraljic, 2009 for an overview). “Noncanonical pronunciation variant” hereby means that a speaker produces a certain speech sound in a way that differs from how most native speakers of this language would produce it. Specifically, in most studies these sounds were created by artificially manipulating single sounds in otherwise native-like speech (Kraljic & Samuel, 2006; Norris, McQueen, & Cutler, 2003). What has been shown is that listeners use external context such as lexical (Norris et al., 2003), visual (lipread: Bertelson, Vroomen, & de Gelder, 2003), or orthographic (Mitterer & McQueen, 2009) information to interpret these mostly ambiguous sounds and later rely on this experience when interpreting new tokens of these sounds. For example, the use of lexical context builds on the finding that listeners tend to hear real words over nonwords even when the word contains acoustically ambiguous sounds (Ganong, 1980). That is, the last sound in the word “giraffe” is likely to be interpreted as /f/ even if it is acoustically ambiguous between /f/ and /s/ since, in the lexical context of “gira_”, /f/ but not /s/ leads to the interpretation of a real word. Importantly, when such ambiguous sounds are later encountered in a neutral context (e.g., “lea_” where both “leaf” and “lease” are words), listeners still tend to interpret them in line with the previous context (McQueen, Cutler, & Norris, 2006; Mitterer, Chen, & Zhou, 2011; Sjerps & McQueen, 2010). Similar effects have also been shown with lipread context information (Bertelson et al., 2003). Listeners integrate visual (lipread) and auditory information in phoneme recognition (McGurk & MacDonald, 1976). Similar to the example above, when hearing an ambiguous auditory stimulus between “aba” and “ada” coupled with a clear visual stimulus (e.g., lip closure in “aba”), listeners are likely to interpret the sound in line with the visual information. Later, if they hear an ambiguous auditory-only stimulus, it is perceived in line with the previously seen visual stimulus (Bertelson et al., 2003; Van Linden & Vroomen, 2007; Vroomen, van Linden, de Gelder, & Bertelson, 2007). Listeners are thought to have “recalibrated” their perception and the phenomenon has been termed “phonetic category retuning”, “perceptual recalibration” or simply “perceptual learning”.

These kinds of effects have also fueled the debate between models of spoken-word recognition, for example, about the question whether mental representations of words are abstract or episodic. Using a cross-modal priming task at test, McQueen et al. (2006) tested whether listeners generalize category recalibration from one set of exposure words to a new set of test words. Their finding that generalization across words does occur argues for

[*] Corresponding author at: Ludwig Maximilian University Munich, Department of Phonetics and Speech Processing, Schellingstr. 3, 80799 Munich, Germany. Tel.: +49 89 2180 5752.
E-mail address: evarei@phonetik.uni-muenchen.de (E. Reinisch).

0095-4470/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2014.04.002

some form of sublexical representational unit. A popular choice for sublexical units in models of spoken-word recognition is the context- and position independent phoneme (McClelland & Elman, 1986; Norris & McQueen, 2008). However, despite the large body of literature showing recalibration (see Samuel & Kraljic, 2009, for an overview), what has received little attention so far is whether recalibration is indeed based on context- and position-independent types of representations (i.e., phonemes).

Most studies on category recalibration implicitly assume or explicitly suggest (e.g., Eisner & McQueen, 2005; McQueen et al., 2006) that the categories do correspond to (abstract) phonemes, and successful recalibration entails a shift in the phoneme boundary (Clarke-Davidson, Luce, & Sawusch, 2008). An account of perceptual learning based on abstract phonemes would hence predict that listeners generalize recalibration not only across words but more specifically also across syllabic positions. In line with this assumption, Jesse and McQueen (2011) showed that exposure to the ambiguous sound between /f/ and /s/ in word-final position leads to recalibration when listeners were asked to categorize the phoneme continuum in word-initial position.

One way to conceptualize phonetic recalibration of phonemes is that listeners simply learn to accept less phonetic evidence for a given phoneme category as sufficient to assume that this phoneme was intended by the speaker. If this was the case, it would not matter how phoneme identity was cued (in the same or different position). This prediction does in fact follow from theories that assume that listeners do not focus on the acoustic signal but rather on the sound-producing gestures, such as direct perception (Fowler, 1996) and motor theory (Galantucci, Fowler, & Turvey, 2006). According to these theories, listeners abstract away from the acoustic implementation in speech processing, which, as a consequence, predicts generalization across cues for a given phoneme.

Critically, however, there is evidence for recalibration involving categories that are more specific than phonemes as by definition a phoneme comprises all implementations of speech sounds or phones that group together into units that minimally distinguish the meaning of words (Hyman, 1975). Dahan and Mead (2010) investigated this issue using adaptation to noise-vocoded speech. They found that listeners recognized new noise-vocoded words better, the more similar they were to the exposure stimuli. They tested consonant-vowel-consonant (CVC) word or nonword sequences in which the match between exposure and test were (1) CV or VC sequences, (2) one of the consonants including its position in onset or coda, (3) one consonant but in different position. Results showed that consonants were easier to recognize when they had occurred in the same syllable position, and when the vowel context was identical between training and test (i.e., condition 1 was best, followed by condition 2). Accordingly, the authors argued that recalibration is context-specific. However, the range of stimuli used by Dahan and Mead was, from a phonetic point of view, a mixed bag with phonemes that are relatively context-invariant (such as voiceless fricatives – these were the categories used by Jesse & McQueen, 2011 showing position independence) and phonemes that change drastically over position (such as voiced vs. voiceless stops in American English). Generalization across contexts and positions may be "easier" the more context-invariant the acoustic cues to a phoneme are.

Mitterer, Scharenborg, and McQueen (2013) tested a more circumscribed case in which the cues to one phoneme were vastly different, namely the case of Dutch /r/. Dutch /r/ has a variety of free allophones, that is, implementations of the same phoneme that are articulatorily and acoustically distinct. Specifically, Dutch /r/ can be an apical trill, a velar fricative, or an alveolar approximant varying even within the same speaker. Mitterer et al. found that recalibration is restricted to the allophone of the phoneme heard during exposure. After recalibration of the /r/-/l/ contrast for which listeners had been exposed to poor examples of a Dutch approximant [ɹ] in word-final position, the expected category shift was found for [ɹ]-[ɫ] continua matching the allophones heard during exposure but not for continua in which the /r/ endpoint was articulated as an apical trill [r] or the /l/ was switched from "dark" slightly velarized [ɫ] to "light" non-velarized [l] (in word-initial position). Mitterer et al. thus suggest that the categories that listeners recalibrate may correspond to allophones, that is, specific implementations of a phoneme.

However, there is an alternative to the allophone as the category of recalibration. Recalibration may apply to specific acoustic cues that provide information independently of the specific phoneme contrast. Examples would be durational cues to stop voicing that are independent of place of articulation (e.g., /b/-/p/ vs. /d/-/t/) or direction of formant transitions to cue place of articulation that are independent of manner of articulation (/b/-/d/ vs. /m/-/n/). Kraljic and Samuel (2006) showed that listeners generalize recalibration of durational cues to voicing in alveolar stops (/d/-/t/) to labial stops (/b/-/p/). This suggests that listeners recalibrated phoneme-independent durational information rather than the specific phoneme contrast. That is, listeners may recalibrate specific acoustic cues.

Given this mixed evidence for the units of phonetic category recalibration, the present study set out to test the various suggested possibilities under tightly controlled experimental conditions. We tested the specificity of recalibration asking whether phonetic recalibration generalizes over acoustic cues if they cue the same phoneme contrast, whether it generalizes over phoneme contrasts if the cues were the same, and whether it generalizes over acoustic contexts. From these specific questions we sought to gain insight whether abstract phonemes, specific allophones or specific acoustic cues are likely units for recalibration. Implementing this in a lexically-guided recalibration paradigm, however, would be difficult, since lexically-guided recalibration usually requires ten to twenty words in which a given phone sequence occurs (Kraljic, Samuel, & Brennan, 2008; Poellmann, McQueen, & Mitterer, 2011). Using such a large set of target words (and additional fillers if the exposure task is lexical decision), it would be hard to control the exact acoustic implementation of a category contrast in different tokens.

Therefore, we used a paradigm in which recalibration was visually guided by means of lipread information (following the paradigm first used by Bertelson et al., 2003). The standard setup of visually-guided recalibration experiments is a pretest-exposure-posttest design. In the pretest, listeners categorize nonsense syllables along an /aba/-/ada/ continuum to establish their personal maximally ambiguous step on the continuum. During exposure, this maximally ambiguous stimulus is then paired with a video of a speaker, for example, articulating a labial (as indicated by a visible lip closure). An exposure block consists of eight such ambiguous-audio-unambiguous-video pairings, the same stimulus repeated 8 times. Each exposure block is then immediately followed by six auditory-only test trials containing the ambiguous sound heard during exposure as well as adjacent steps on the /aba/-to-/ada/ continuum. This exposure-test combination is repeated multiple times. Half of the exposure blocks show the speaker produce a labial in the video, half show the speaker produce the alveolar. A shift in the categorization functions following labial or alveolar exposure videos indicates the recalibration effect. In test trials following the video in which the speaker was seen articulating a labial, more /b/ responses are expected than when the speaker was seen producing no lip closure.

Importantly, there is experimental evidence suggesting that the underlying processes of visually-guided and lexically-guided recalibration may be equivalent (Van Linden & Vroomen, 2007). In a series of experiments Van Linden and Vroomen (2007) showed that visual disambiguation of Dutch acoustically ambiguous nonwords led to comparable effects of recalibration as did disambiguation by lexical information. Recalibration through both types of information was about equally robust, equally stable over time, and was equally affected by the presence of contrast phonemes. This led the authors to conclude that "From a functional perspective, there was no difference between bottom-up perceptual information or top-down stored lexical knowledge: Both information sources were used in the same way to adjust the boundary between two phonetic categories" (Van Linden & Vroomen, 2007, p. 1492). Hence, even though most studies addressing the units of category recalibration (McQueen et al., 2006; Mitterer et al., 2013) used lexically-guided recalibration, the suggestions about equivalence of paradigms made us chose the one allowing for a better control of the stimuli.

Our starting point was the recalibration of the place of articulation contrast /b/-/d/ in American English. A recalibration effect for this contrast has repeatedly been shown (following Bertelson et al., 2003; i.e., the paradigm described above). To address the role of the phoneme or allophone as the category for recalibration, we presented the stops in intervocalic position with two different vowel contexts: The first context was the vowel /a/, the second context was /i/. These two contexts lend themselves well for the creation of different or even complementary sets of cues in the speech signal. In the context of /a/ formant transitions in the second and third formants are falling for /b/ and rising for /d/. In this case, the cues in the surrounding vowels could provide information about place of articulation while the closure, burst and frication of the stop could be set to silence to eliminate cues in this portion of the signal. Voiced stops in American English are usually well recognizable without a burst (see Liberman, 1996). In the context of /i/, formant transitions carry relatively less information for place of articulation as they are falling for both /b/ and /d/ (Smits, Bosch, & Collier, 1996). In this case, the information about place of articulation could be restricted to the "obstruent" portion of the speech signal (closure, burst, frication) and the formant transitions in the "sonorant" (vowel) portion of the signal could be set to fully ambiguous values. If, under these conditions, listeners generalize from /aba/ or /ada/ exposure to an /ibi/-/idi/ test continuum (or the other way around), this would be evidence that listeners generalize recalibration across different acoustic cues. Here, by manipulating natural speech tokens we will be able to put the hypothesis to a stringent test as the cues will be controlled to involve only complementary portions of the speech signal. Note that if two instances of a phoneme that are cued differently are seen as two allophones of a phoneme, then finding generalization would be in conflict with the suggestions that listeners recalibrate specific allophones rather than abstract phonemes (Mitterer et al., 2013). However, the units of recalibration may not be abstract phonemes either.

To evaluate the other possibility, namely that listeners recalibrate specific acoustic cues independently of the phoneme contrast involved, we tested generalization across manner of articulation within the same place-of-articulation (i.e., labial–alveolar) contrast. That is, in the context of the vowel /a/, place of articulation can sufficiently be cued by formant transitions in the vowels for stops (/b/-/d/) as well as for nasals (i.e., /m/-/n/, see, e.g., Harding & Meyer, 2003; Kurowski & Blumstein, 1984; Repp & Svastikula, 1988 for a discussion of cue weighting in nasals). Therefore, if listeners recalibrated their perception of the formant trajectories with regard to place of articulation rather than a specific phoneme contrast, they would be expected to show recalibration of a nasal continuum after being exposed to a stop contrast, or to show recalibration of a stop continuum after exposure to an ambiguous nasal in unambiguous visual context.

2. Experiment 1

Experiment 1 investigated whether phonetic recalibration generalizes across different cues to the same phoneme. Experiment 1 can thus be termed the same-phoneme-different-cue (in different context) experiment. We tested visually-guided auditory recalibration of the labial–alveolar stop contrast in vowel–consonant–vowel nonword sequences in which the vowel contexts were either /a/ or /i/. The cues to place of articulation in the consonants were manipulated such that they were complementarily distributed across the vowel contexts. They were either the direction of the *F*2 and *F*3 formant transitions (/a/ context) or the obstruent part including closure, burst, and frication (in the /i/ context).

The critical test was whether listeners would generalize recalibration of /b/ and /d/ between vowel contexts. For example, participants were exposed to visual /aba/ and /ada/ and tested on /aba/-/ada/ to assess the basic recalibration effect and on /ibi/-/idi/ to test generalization to the same phoneme but in a different vowel context and cued by a different type of cue (i.e., formant transitions vs. consonant mix). Another group of participants was exposed to /ibi/ and /idi/ and tested on /ibi/-/idi/ for the basic effect and /aba/-/ada/ for generalization. If listeners generalize recalibration across acoustic contexts and cues, differences in the proportion labial responses at test (depending on the visual context seen during exposure) should be observed for the exposure continuum as well as the generalization continuum.

2.1. Method

2.1.1. Participants

53 participants, undergraduates at Carnegie Mellon University, took part for partial course credit. They were native speakers of American English, and reported no language or hearing disorders. 13 participated in the pretests and 40 in the recalibration experiment.

2.1.2. Materials

A male native speaker of American English was videotaped (head and shoulders) in front of a light gray background while articulating the two-syllable nonsense words /aba/, /ada/, /ibi/, /idi/, /ama/, /ana/, /ubu/, and /udu/ using a JVC Pro HD (3CCD) camera with a Fujinon zoom lens (model GJ HM 100 E; note that the /ama/-/ana/ and /ubu/-/udu/ videos were used only in Experiments 2 and 3 respectively). An audio track was recorded along with the video via a camera-mounted microphone. The nonwords were pronounced with relatively flat intonation and without specific stress on either the first or second syllable. Video tokens of each nonword were selected in which the speaker did not blink or noticeably move his head or shoulders. Additionally, care was taken that the selected video tokens with a labial vs. alveolar consonant were as closely as possible matched on duration as measured from the start of the mouth gesture of the first vowel to the end of the gesture of the second vowel. This appeared rather easy as the speaker managed to articulate the nonwords at a constant rate deviating by maximally one video frame between the selected tokens.

The same nonwords were later re-recorded in a sound-conditioned booth since high-quality recordings resulted in more naturally sounding tokens after manipulation (see below). For the high-quality recordings the speaker was prompted to articulate the tokens at approximately the same rate and with the same intonation contour as for the video recordings. Audio tokens from the high-quality recordings were selected such that the consonants and vowels matched the original audio track as closely as possible in duration. Note that an exact match in duration between an auditory and visual syllable would not be necessary, as observers consistently fail to notice such small audiovisual asynchronies (Vatakis & Spence, 2007) and tend to integrate information from asynchronous audio and video within a time window of up to 200 ms (Van Wassenhove, Grant, & Poeppel, 2007). Nevertheless, we diminished the asynchronies by adjusting the durations of the V or C segments by manually doubling or deleting single periods in the respective segments. These adjustments never amounted to more than 10 ms and were carried out at random points throughout the segments. The final durations were as follows: first vowel 169 ms, consonant 78 ms, second vowel 198 ms. The overall discrepancies between the original utterances (i.e., VCV nonwords) from the video recordings and the edited utterances from the high-quality audio recordings never exceeded the duration of one video frame. Time alignment of the soundtracks was set at the burst release of the consonant matching the original and the high-quality audio tracks and thus anchoring the most salient event in the audio and video recordings.

Table 1
Formant endpoints at the vowel–consonant interface for the VCV nonwords in all three experiments. The notation “_” indicates that the C part was set to silence; “C” indicates the presence of closure voicing, burst and frication in the C part; “N” indicates the nasal.

Experiment	Context/condition	*F*2 (Hz)		*F*3 (Hz)	
		Labial	Alveolar	Labial	Alveolar
1, 2, 3	a_a	1100	1500	2450	2700
1	iCi	2250	2250	2600	2600
2	aNa	1100	1500	2450	2700
3	u_u	980	1700	2150	2300

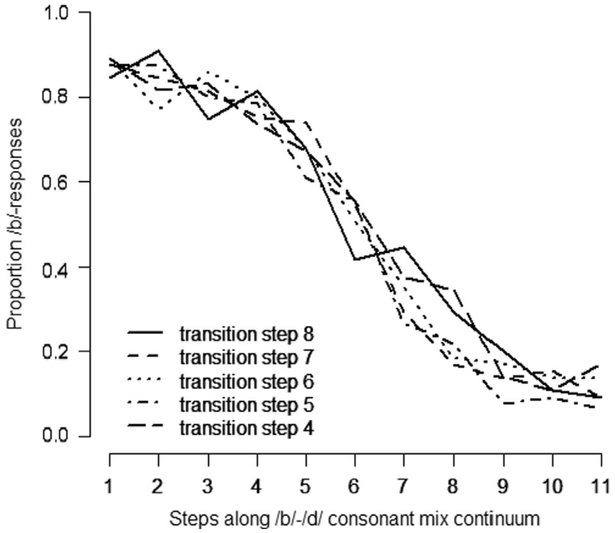

Fig. 1. Results of the pretest for the /ibi/-/idi/ continuum. The *y*-axis shows proportion /b/ responses. Steps along the /b/-/d/ consonant mix continuum are plotted on the *x*-axis. The lines represent the middle steps of the original 11-step formant transition continuum.

The high-quality audio recordings were further manipulated to create continua between labial and alveolar consonants. For the nonword pair /aba/-/ada/ the signal during the obstruent part of the nonwords (i.e., any voicing during stop closure as well as the burst and frication after the release) was set to silence. This condition will be referred to as “a_a” with the underscore indicating the silent obstruent. To create an 11-step continuum from /b/ to /d/ the endpoints of the formant transitions of the second and third formants (*F*2 and *F*3) in both vowels were interpolated linearly in Hertz. This was achieved by using LPC based source-filter separation in PRAAT (Boersma, 2001), manipulation of the filter, and subsequent recombination of the original source with the manipulated filter. Transitions in the vowels were symmetrical around the consonant and set to a duration of 70 ms. This matched the estimated transition durations of the natural /aba/ token. Endpoint values were based on the natural productions of the speaker, rounded to the next full 50 Hz.[1] These values are listed in Table 1. The intermediate continuum steps as well as the endpoints used in the experiment were resynthesized. Listeners' ability to identify the resynthesized endpoints was further confirmed in a pretest. The pretest ensured that despite the lack of the obstruent part listeners perceived a continuum from /b/ to /d/, and no confusion could be found with /m/, /n/, /ŋ/, or /g/.

In the pretest 13 participants were presented all 11 steps of the continuum fifteen times in a 6-alternative forced-choice task. Listeners had to report on every trial whether the consonant they heard was /b/, /d/, /m/, /n/, /ŋ/, or /g/. Listeners logged their responses using the number keys 1 to 6 on a computer keyboard. Response options along with their key numbers were presented onscreen throughout the experiment. Response options were spelled “aba”, “ada”, “ama”, “ana”, “aga”, and “anga”. In only 54 of 2143 “valid” trials (trials with RTs below 200 ms or over 2500 ms were discarded) listeners reported sounds other than /b/ or /d/. This is a mere 2.5% of unintended identifications. The tokens reported as /b/ and /d/ formed the expected s-shaped function of a continuum reaching from 97% /b/ responses at the b-endpoint of the continuum to 0.006% /b/ responses at the /d/ end of the continuum. The middle step of the continuum with 67% /b/ responses was closest to the 50% boundary and hence perceived as the most ambiguous sound.

In the vowel context /i/, the obstruent parts in /b/ and /d/ (voicing during closure, burst, and frication) were mixed on a sample-by-sample basis in 11 steps from /b/ to /d/. This condition will be a referred to as “iCi” condition, where the “C” indicates the presence of stop voicing/closure and burst/frication. Additionally, *F*2 and *F*3 (*F*2 /b/=2100 Hz, /d/=2300 Hz; *F*3 /b/=2400 Hz, /d/=2800 Hz) were interpolated in a similar fashion as in the /a/ vowel context, again in 11 steps between the natural endpoints rounded to the next 50 Hz. Note, however, that in the context of /i/, the formant transition continuum served to find a maximally ambiguous value of this cue. Therefore, in the pretest, the 11 steps of the consonant continuum were crossed with only the 5 middle steps of the formant continuum. In a two-alternative forced-choice task between /ibi/ and /idi/ listeners clearly used the proportion of /b/ in the consonant mix to base their decisions on. Fig. 1 shows the expected s-shaped categorization function from the /b/ to the /d/ endpoint along the consonant continuum (*x*-axis). However, it also shows that formant transitions of the limited 5-step continuum appeared not to systematically influence listeners' decisions. Therefore the middle step of the formant transition continuum was chosen as the most ambiguous

[1] Note that the purpose of the experiment was to test category recalibration under tightly controlled circumstances in which the identification of the sounds involving different cues more was more relevant than the match of formant values to one speaker's specific natural productions.

transition step used for the recalibration experiment. In this way listeners could be presented with complementary cues to place of articulation in the /a/ vs. /i/ context. The cues were vowel transitions (in /a/ context) vs. information in the consonantal part of the signal (in /i/ context).

2.1.3. Procedure

For the recalibration experiment, the two 11-step continua (i.e., aba-ada, and ibi-idi) were spliced onto the respective selected videos in which the speaker produced a labial and an alveolar sound. To time-align the manipulated recordings we used the original soundtrack that was recorded with the camera-mounted microphone. The burst release was used as the main anchor point for alignment. In the recalibration experiments, the exposure videos were selected individually for each participant to contain the audio stimulus that was the most ambiguous step of the continuum for this participant.

The experiment consisted of two phases and followed the paradigm first used by Bertelson et al. (2003). In the initial "pretest" phase, participants categorized the consonants for both the a_a continuum and the iCi continuum using an auditory-only 2-alternative forced choice task. For both continua, participants responded to the full 11-step continuum 10 times totaling in 220 responses per participant. The continua were presented to participants in the same exposure-generalization order that would occur in the following "main" phase. For each participant in the pretest, a logistic function was fitted for each continuum to determine the continuum step at which the portion "labial" responses were closest to 50%. These middle steps were then used for presentation in the main phase.

The main phase of the experiment consisted of 20 exposure-test blocks. Each of these blocks consisted of 8 exposure trials and 6 test trials. In the exposure trials, participants were instructed to watch a video and listen to the speaker. For each participant, the videos for the exposure trials within each block were identical but across blocks half of them showed the speaker articulating a labial and the other half an alveolar consonant. The audio track was always the participant's individual most ambiguous sound from the continuum. To keep the listener's attention on the screen, in 10% of the exposure trials a small red dot appeared on the speaker's upper lip for 300 ms. Participants were instructed to hit the spacebar whenever the red dot appeared. The exposure trials were immediately followed by six auditory-only test trials. The test trials consisted of the individuals' most ambiguous tokens and the two adjacent steps on the original 11-step continua (i.e. ambiguous step ±1). These three steps were repeated twice resulting in 6 test trials after each exposure trial set. Half of the time, test trials following each labial or alveolar exposure set were from the same continuum (e.g., exposure=a_a and test=a_a), a recalibration control block, or from a different continuum (e.g., exposure=a_a and test=iCi), a generalization block. Sixteen participants received the a_a continuum as the recalibration control and iCi as for generalization; 24 received the iCi continuum for the recalibration control and a_a for generalization. The order of blocks was randomized separately for each participant.

2.2. Results

Analyses were conducted separately for the a_a and iCi exposure conditions to show whether recalibration can be found for both types of cues during exposure. We used linear mixed-effects models with a logistic linking function to account for the dichotomous dependent variable (labial response=1, alveolar response=0). Predictor variables were contrast coded such that effects of the factors could be interpreted as main effects and the sign of the regression weight would indicate the direction of the effect. First, overall models were fitted with Trial Type (recalibration control trials=0.5, generalization trials=−0.5), Place of Articulation during exposure "Exposure POA" (labial=0.5, alveolar=−0.5), Consonant Continuum (−1, 0, 1 where 0 is the token that was heard during exposure), and their interactions as fixed factors. Participant was entered as a random factor with additional random slopes for all possible fixed effects (see Barr, Levy, Scheepers, & Tily, 2013, for a discussion of random slopes for within-participant factors). Then, if an interaction between POA and Trial Type was found (suggesting differences in the strength of the recalibration effect between recalibration control and generalization trials) separate analyses were conducted for the two trial types.

Table 2 shows the results for these overall analyses. We found main effects of all factors, effects of all two-way interactions, and for the a_a exposure condition also a significant three-way interaction. We briefly discuss these effects before turning to the critical interaction (Exposure POA × Trial Type). The effects of Continuum suggest that even for the three most ambiguous steps of the continuum listeners gave more /b/ responses the more /b/-like the stimulus was. Main effects of Exposure POA suggest that overall listeners gave more /b/-responses following exposure to the video in which the speaker produced a labial rather than an alveolar sound (i.e., overall but disregarding the critical interaction discussed below, there was a recalibration effect). The effects of Trial Type suggest that more /b/-responses were given in the recalibration control trials than the generalization trials. The two-way interactions between Exposure POA and Continuum suggest differences in the slopes of the categorization functions for trials following labial vs. alveolar exposure. The interactions between Trial Type and Continuum suggest differences in the slopes of the continua between recalibration control and generalization trials (with steeper slopes for generalization trials when the exposure was a_a, and steeper slopes for recalibration control trials when the exposure was iCi).

Critically, for both, the a_a and iCi exposure condition an interaction between Exposure POA and Trial Type was found. This suggests differences in the strength of recalibration between recalibration control trials and generalization trials (note that the three-way interaction for the a_a context shows a further modulation of this difference depending on continuum step suggesting larger differences the "lower" the step of the continuum).

Table 2
Overall analyses in Experiment 1, the same-phoneme-different-cue experiment, for the condition in which participants saw the speaker produce the ambiguous sound in the context of a_a, and the context of iCi. The highlighted row marks the relevant interactions between Exposure POA (labial–alveolar) and Trial Type (recalibration control – generalization).

Factor	aba-ada exposure				ibi-idi exposure			
	b	*SE*	*z*	*p*	*b*	*SE*	*z*	*p*
(Intercept)	−0.19	0.23	−0.82	0.41	0.64	0.17	3.72	<0.001
Continuum	−2.71	0.23	−11.91	<0.001	−2.07	0.14	−15.13	<0.001
POA	0.37	0.18	2.11	<0.05	1.05	0.14	7.52	<0.001
TrialType	1.50	0.45	3.36	<0.001	0.74	0.33	2.27	<0.05
Continuum:POA	−0.73	0.33	−2.22	<0.05	−0.32	0.16	−1.98	<0.05
Continuum:TrialType	−1.86	0.38	−4.91	<0.001	1.68	0.27	6.17	<0.001
POA:TrialType	0.82	0.34	2.41	0.016	1.45	0.36	3.98	<0.001
Continuum:POA:TrialType	−1.63	0.62	−2.60	<0.005	0.23	0.34	0.67	0.50

Fig. 2 shows the results separately for recalibration control and generalization trials and thereby reveals the source of the interaction between Exposure POA and Trial Type. Whereas, for the recalibration control trials (left panels), the categorization functions following videos with or without lip closure are clearly different, for the generalization trials (right panels), the lines more or less overlap. Analyses reported in Table 3 confirm this. A robust recalibration effect could be found for the recalibration control trials such that more labial responses were given when participants saw the speaker articulate a labial during exposure than when they saw the speaker articulate an alveolar sound. However, listeners did not generalize this recalibration when tested on the same phoneme in a different vowel context where the phoneme was cued differently (i.e., formant transitions vs. closure, burst and frication).

Note that when iCi was the exposure condition, the generalization effect was marginally significant (i.e., $p_{(Exposure\ POA)}=0.07$). This is the reason why in this condition 8 more participants were tested than in the a_a exposure condition. After 16 participants in the iCi condition the generalization was already marginally significant ($p_{(Exposure\ POA)}<0.1$ but >0.05). Since, however, after adding 8 more participants evidence for generalization was

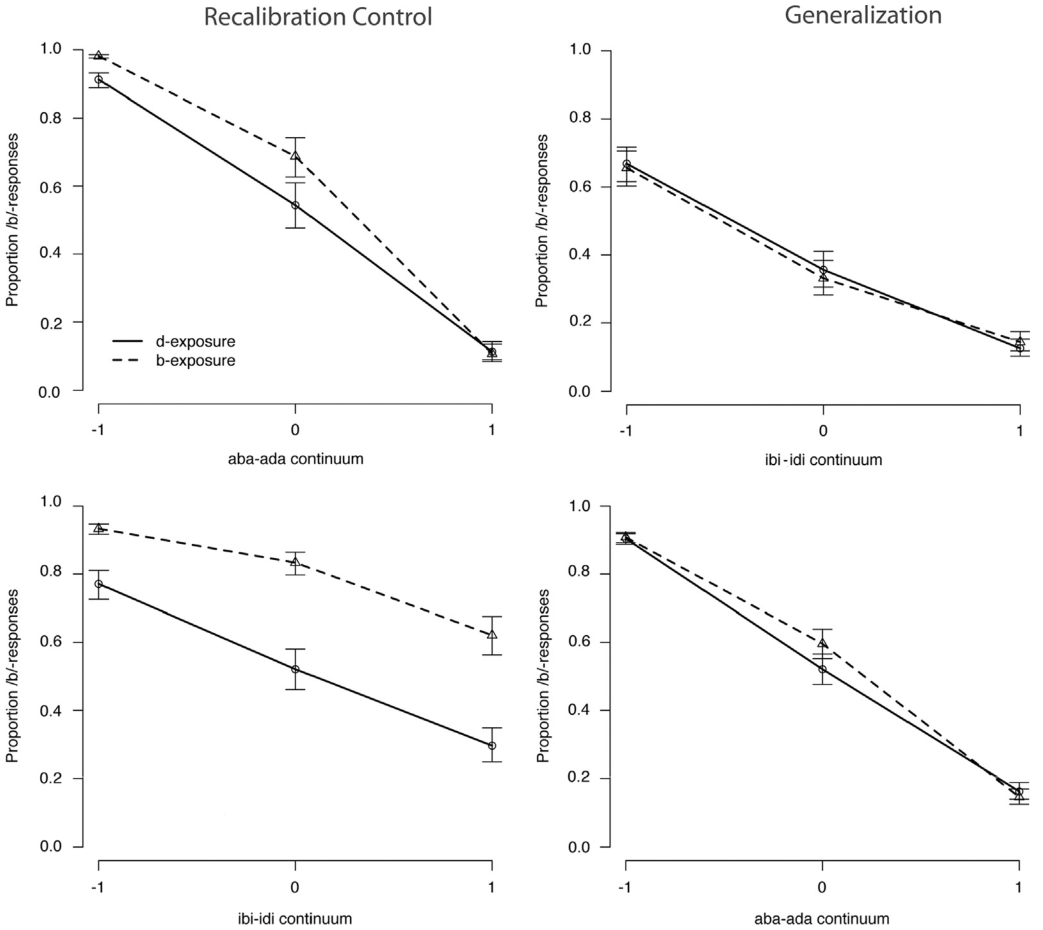

Fig. 2. Results for Experiment 1, the same-phoneme-different-cue experiment. Proportion labial responses (*y*-axis) are plotted across the test continua (*x*-axis). The dashed lines show responses following labial (i.e., /b/) exposure; the solid lines show responses following alveolar (i.e., /d/) exposure. The “error bars” are based on the Standard Error of the regression weight of Exposure POA in each of the four conditions (see Table 3). The standard error was projected back into the proportion scale, leading to larger intervals around 0.5 and asymmetric intervals at floor and ceiling. The left panels show results for the recalibration control trials and the right panels show results for the generalization trials. The upper panels show the condition in which a_a was the exposure context (i.e., formant transitions as informative cues), and the lower panels show the condition in which iCi was the exposure context (i.e., cues in closure, burst, and frication of the consonant).

Table 3
Analyses split up by recalibration control trials and generalization trials in Experiment 1, where generalization was tested to the same phoneme cued differently. Note that Generalization Trials for the a_a exposure condition are iCi trials and vice versa.

Exposure condition	Factors	Recalibration control trials				Generalization trials			
		b	*SE*	*z*	*p*	*b*	*SE*	*z*	*p*
a_a	Intercept	0.55	0.27	2.05	<0.05	−0.94	0.36	−2.62	<0.005
	Continuum	−3.52	0.35	−10.00	<0.001	−1.76	0.21	−8.48	<0.001
	ExposurePOA	0.81	0.27	3.01	<0.005	−0.03	0.23	−0.14	0.89
	Continuum*ExposurePOA	−1.68	0.52	−3.22	<0.005	0.08	0.24	0.31	0.76
iCi	Intercept	0.97	0.13	7.22	<0.001	0.27	0.31	0.88	0.38
	Continuum	−1.17	0.09	−12.42	<0.001	−2.88	0.25	−11.77	<0.001
	ExposurePOA	1.67	0.24	7.02	<0.001	0.33	0.18	1.79	0.07
	Continuum*ExposurePOA	−0.11	0.2	−0.55	0.58	−0.48	0.25	−1.94	=0.05

still not found (i.e., the effect of Exposure Place of Articulation was still marginally significant) we conclude that in line with the lack of evidence for generalization in the a_a exposure condition, generalization to the same phoneme cued differently if anything, is rather unstable (see also the interaction between Exposure POA and Continuum which along with Fig. 2 suggests that the marginal "effect" differs along the steps of the test continuum).

2.3. Discussion

Experiment 1 showed that listeners use lipread information to guide phonetic recalibration in different acoustic contexts and for different acoustic cues (cf. the two different recalibration control conditions a_a vs. iCi). However, although exposure to one specific cue in one specific vowel context robustly triggered recalibration for these recalibration control trials, recalibration was not generalized to the same phoneme in a different vowel context where different acoustic cues were relevant to determine phoneme identity. This lack of evidence for generalization suggests that abstract context-independent phonemes are likely not the categories to be recalibrated. In this case generalization would have been expected. Since we defined the stop consonants in the context of /a/ vs. /i/ by means of complementary acoustic cues, they could be interpreted as allophones, that is, different (here: acoustic) implementations of the same phonemes. The present results could thus support suggestions that recalibration is restricted to the allophone of the phoneme heard during exposure (Mitterer et al, 2013). We will return to this suggestion in Experiment 3.

However, one alternative explanation has to be considered. Since the purpose of Experiment 1 was to present listeners with complementary cues, and to make the formant transitions in the vowels the only cues to the /b/-/d/ contrast in the a_a context, the consonantal part of the signal in this context was set to silence. In contrast, in the iCi condition the consonantal part carried the crucial information to the place of articulation of the consonant while formant transitions were merely set to an ambiguous value (it would have been impossible to leave them out completely). Hence, recalibration control trials and generalization trials differed in the presence vs. absence of the consonantal portion. To exclude the possibility that acoustic coherence between trial types could be the reason for the lack of generalization, a control experiment was run with fifteen participants using the same setup and procedure as the a_a-exposure iCi-generalization condition.

The stimuli were the same as in Experiment 1 with one additional manipulation to the a_a audio continuum. The consonant in a_a was not set to silence but replaced by an ambiguous step from a /b/-to-/d/ consonant continuum. That is, the closure, burst, and frication of /b/ and /d/ in the /a/ context was interpolated to an 11-step continuum as had been done for the iCi continuum. The most ambiguous step was established in a pretest. Twelve additional listeners performed a two-alternative-forced choice task, categorizing stimuli from a continuum grid in which the 11 steps of the transition continuum were crossed with the five middle steps of the consonant continuum. The consonant mix with 30% /b/ was chosen as the most ambiguous step for use in the recalibration experiment since for this continuum step the proportion /b/ responses at the middle step of the transition continuum was closest to 50%. We could thus expect the resulting continuum to be approximately symmetrical when only formant transitions were informative cues to place of articulation.

This newly created audio continuum was then spliced onto the videos showing the speaker articulate /aba/ and /ada/ using the same procedure as before. The recalibration experiment was identical to recalibration experiment in Experiment 1 in which the /a/ context was used for exposure and as recalibration control (here with aCa audio) and iCi was used for generalization trials. Results were similar to what was found in Experiment 1, with the crucial finding[2] of an interaction between Exposure POA and Trial Type ($b_{(\mathrm{Intercept})}=-0.37$, $p<0.05$; $b_{(\mathrm{Continuum})}=-1.57$, $p<0.001$; $b_{(\mathrm{POA})}=0.18$, $p=0.21$; $b_{(\mathrm{TrialType})}=0.35$, $p<0.005$; $b_{(\mathrm{Continuum}_\ast\mathrm{POA})}=0.16$, $p=0.35$; $b_{(\mathrm{Continuum}_\ast\mathrm{TrialType})}=-0.38$, $p<0.05$; $b_{(\mathrm{POA}_\ast\mathrm{TrialType})}=1.13$, $p<0.001$; $b_{(\mathrm{Continuum}_\ast\mathrm{POA}_\ast\mathrm{TrialType})}=-0.18$, $p=0.56$). It is thus unlikely that a lack of acoustic coherence in the speech signal between exposure stimuli and stimuli used for generalization trials at test had contributed to the failure to find generalization. Replicating the interaction between Exposure POA and Trial Type with the new set of stimuli further strengthens the suggestion that the categories that are recalibrated are more specific than an abstract, context-independent phoneme. In Experiment 2 we now set out to test whether listeners may recalibrate the perception of specific acoustic cues independently of the phoneme contrast heard during exposure.

3. Experiment 2

Experiment 1 examined whether listeners can generalize perceptual recalibration from one cue to another for the same phoneme contrast. Experiment 2 tested the mirror image. Whereas Experiment 1 was coined the same-phoneme-different-cue experiment, this is a "different-phoneme-same-cue experiment": The cues are constant but the phoneme contrast in the generalization condition differs. Specifically we asked whether - if the relevant cues are the same - listeners would generalize their retuned categories across phoneme contrasts – here contrasts that differ in their manner of articulation (/b/-/d/ vs. /m/-/n/). To our knowledge, none of the previous studies on phonetic recalibration has tested generalization across manner of articulation (Kraljic & Samuel, 2006 investigated generalization of learning of the voicing contrasts across place of articulation). The experiment assessed the importance of the specific acoustic cues for the generalization of phonetic recalibration.

3.1. Method

3.1.1. Participants

Forty-six new participants were selected from the same population and according to the same criteria as in Experiment 1. All received partial course credit for their participation. Fourteen participated in the pretest and 32 in the recalibration experiment.

3.1.2. Materials and procedure

The stimuli for the a_a tokens were identical to Experiment 1. In addition to the a_a continuum, a nasal continuum from /m/ to /n/, henceforth aNa, was created to test generalization across the same acoustic cues but to a different phoneme contrast. That is, the cues to the nasal continuum were formant transitions while the nasal portion of the signal was set to an ambiguous value. To find a maximally ambiguous token for the nasal portion of the signal, a nasal continuum between /m/ and /n/ was created using recordings of /ama/ and /ana/. The nasal portions were excised and a continuum

[2] Note that other significant factors and interactions can be interpreted in a similar fashion as for Experiment 1 and will not be discussed in further detail here.

was generated by interpolating the respective nasal formants in 11 steps ($m1=295$ Hz, $n1=360$ Hz, $m2=1270$ Hz, $n2=1550$ Hz, $m3=2295$ Hz, $n3=2410$ Hz). The vowel tokens were taken from the a_a continuum. Although vowels surrounding a nasal are likely to be nasalized in natural speech, we decided to keep the context and hence the cues identical between recalibration control and generalization trials (i.e., a_a and aNa). This was to maximize chances of finding generalization. The perception of the nasal tokens as nasal despite the fully oral vowels was ensured in a pretest which also served to find the maximally ambiguous token of the nasal murmur.

For the pretest, the 11-step vowel continuum from the a_a condition was crossed with the five middle steps of the nasal continuum and presented in a 4-alternative forced choice task. Alongside the response options /ama/ and /ana/, the response options /aba/ and /ada/ were included to ensure that listeners perceived the tokens as nasal. This is indeed what we found. Only 19 of 4064 valid trials (trials with an RT below 200 ms or above 2500 ms were excluded) led to a /b/ or /d/ response. Fig. 3 shows responses for the five steps of the nasal continuum along the 11-step formant-transition continuum. As can be seen by the s-shape of the categorization functions in the figure, listeners used the information of the formant transitions in the vowels. The five-step nasal continuum appeared to have some influence on categorization as indicated by the diversification of categorization functions. Since at the middle step of the transition continuum nasal step 4 was closest to 50% /b/ responses, Step 4 of the original 11-step nasal continuum was selected for the recalibration experiment.

The newly created aNa continuum thus contained an ambiguous nasal while place of articulation was cued by format transitions in the vowels, notably the physically identical transitions as for the a_a condition. This continuum was spliced onto videos in which the speaker articulated /ama/ and /ana/. Although the auditory vowels were taken from oral productions, the /ama/ and /ana/ videos were quite similar to the /aba/ and /ada/ videos. Manner of articulation between stops and nasals does not form a separate class of visemes (Bernstein, Demorest, & Tucker, 2000; Owens & Blazek, 1985) and our phones and videos were matched on durational properties. Procedure and analyses were identical to Experiment 1. Sixteen participants received the a_a continuum as the recalibration control and aNa for generalization; 16 received the aNa continuum for the recalibration control and a_a for generalization. The order of blocks was randomized separately for each participant.

3.2. Results

Table 4 reports the overall analyses for the data in Experiment 2, again split up by exposure condition (a_a vs. aNa). The effects of Continuum and Exposure POA suggest that listeners indeed perceived the three-step test continua as ranging from sounding more like a labial to sounding more like an alveolar sound (i.e., effect of Continuum), and that overall listeners did show recalibration (Exposure POA – for sake of description disregarding any interactions) with more labial responses following exposure to a video in which the speaker was articulating a labial. The interactions between Trial

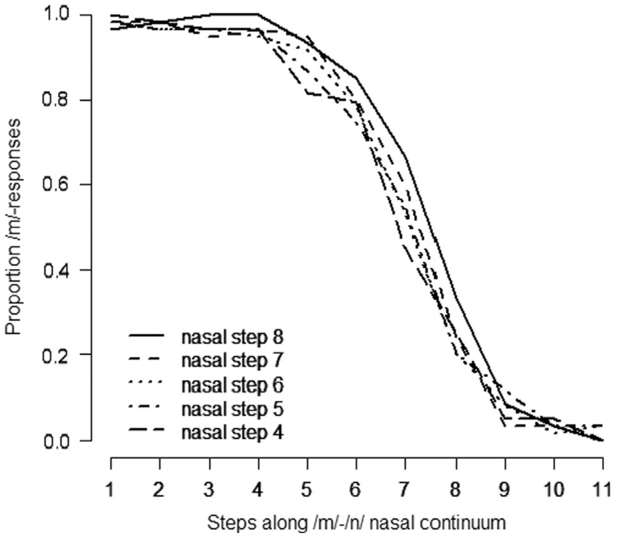

Fig. 3. Results of the pretest for the ama-ana continuum. The *y*-axis shows proportion /m/ responses. Steps along the /m/-/n/ (labial–alveolar) vowel transition continuum are plotted on the *x*-axis. The lines represent the middle steps of the original 11 step nasal continuum.

Table 4
Overall analyses in Experiment 2, the different-phoneme-same-cue experiment, for the condition in which participants saw the speaker produce the ambiguous sound in the context of a_a, and the context of aNa. The gray bar highlights the relevant interactions between Exposure place of articulation (labial–alveolar) and Trial Type (recalibration control – generalization).

Factor	aba-ada exposure				ama-ana exposure			
	b	*SE*	*z*	*p*	*b*	*SE*	*z*	*p*
(Intercept)	−0.06	0.16	−0.39	0.7	0.79	0.42	1.87	0.061
Continuum	−2.64	0.32	−8.26	<0.001	−2.83	0.29	−9.81	<0.001
POA	0.47	0.14	3.40	<0.001	0.82	0.29	2.84	<0.005
TrialType	0.5	0.41	1.20	0.23	−0.21	0.83	−0.26	0.8
Continuum:POA	−0.12	0.23	−0.51	0.61	−0.26	0.27	−0.96	0.34
Continuum:TrialType	−0.51	0.20	−2.47	<0.05	1.71	0.42	4.05	<0.001
POA:TrialType	0.7	0.30	2.32	<0.05	1.15	0.37	3.07	<0.005
Continuum:POA:TrialType	−0.76	0.56	−1.35	0.18	0.10	0.46	0.22	0.83

Type and Continuum suggest differences in the slopes of the continua between recalibration control and generalization trials (with steeper slopes for generalization trials when the exposure was a_a, and steeper slopes for recalibration control trials when the exposure was aNa). Critically, as in Experiment 1, the interactions between Exposure POA and Trial Type were significant, suggesting differences in the strength of the recalibration effect for the recalibration control and generalization trials.

Fig. 4 shows the results separately for the two trial types and thereby reveals the source of the interaction between Exposure POA and Trial Type. Whereas for the recalibration control trials (left panels) the categorization functions following videos with or without lip closure are clearly different, for the generalization trials (right panels) the lines more or less overlap. Separate analyses for recalibration control and generalization trials reported in Table 5 confirm that again, listeners showed robust recalibration for the control trials but did not generalize from either a_a to aNa or the other way around.

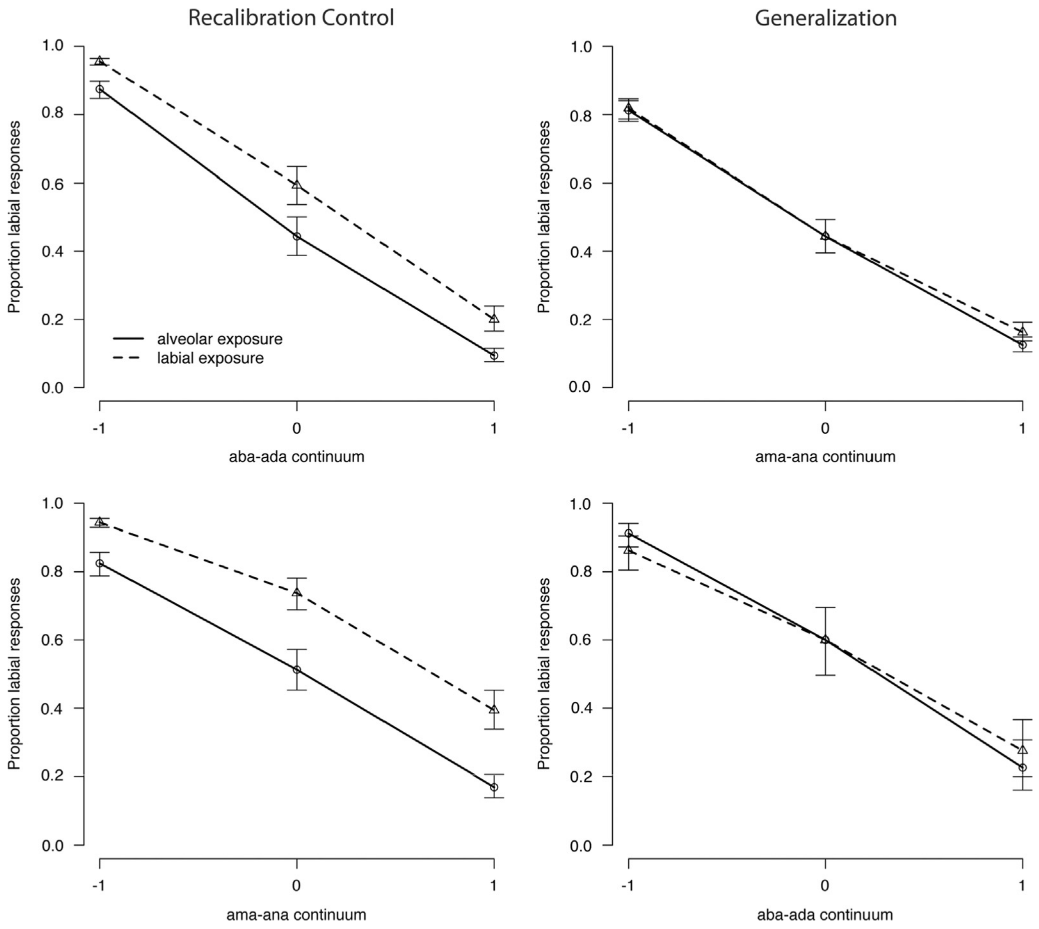

Fig. 4. Results for Experiment 2, the different-phoneme-same-cues experiment. Proportion labial responses (*y*-axis) are plotted across the test continua (*x*-axis). The dashed lines show responses following labial exposure; the solid lines show responses following alveolar exposure. The "error bars" are based on the Standard Error of the regression weight of Exposure POA in each of the four conditions (see Table 5). The standard error was projected back into the proportion scale, leading to larger intervals around 0.5 and asymmetric intervals at floor and ceiling. The left panels show results for the recalibration control trials, the right panels show results for the generalization trials. The upper panels show the condition in which the exposure stimulus was taken from the aba-ada continuum (i.e., formant transitions cue POA of stop consonants), and the lower panes show the condition in which the exposure stimulus was taken from the ama-ana continuum (i.e., formant transitions cue POA of nasal consonants).

Table 5
Analyses split up by recalibration control trials and generalization trials in Experiment 2, the different-phoneme-same-cue experiment. Note that Generalization Trials for the a_a exposure condition are aNa trials and vice versa.

Exposure condition	Factors	Recalibration control trials				Generalization trials			
		b	*SE*	*z*	*p*	*b*	*SE*	*z*	*p*
a_a	Intercept	0.16	0.24	0.67	0.5	−0.32	0.28	−1.15	0.25
	Continuum	−2.9	0.36	−8.17	<0.001	−2.41	0.35	−6.94	<0.001
	ExposurePOA	0.80	0.23	3.5	<0.001	0.11	0.2	0.54	0.59
	Continuum∗ExposurePOA	−0.09	0.40	−0.22	0.83	0.25	0.28	0.9	0.37
aNa	Intercept	0.68	0.25	2.66	<0.01	0.82	0.79	1.03	0.3
	Continuum	−1.97	0.22	−9.13	<0.001	−3.63	0.45	−8.15	<0.001
	Exposure POA	1.40	0.24	5.75	<0.001	0.18	0.42	0.43	0.67
	Continuum∗ExposurePOA	−0.21	0.31	−0.67	0.51	−0.30	0.36	−0.83	0.41

3.3. Discussion

Experiment 2 demonstrated that visually-guided phonetic recalibration is robust for different types of phonemes (i.e., for stops and nasals) but at the same time appears to be very specific when it comes to the specification of the "category" that is recalibrated. Experiment 1 showed that generalization does not occur if the same phonemes are embedded in different acoustic contexts, namely ones in which the cues to the same phoneme tend to be weighed differently; in the case of our experiment they were intentionally set to complementary values by neutralizing the respective other cue. Experiment 2 added that it is not the acoustic cues alone that listeners appear to recalibrate. Keeping the relevant cues to place of articulation and the acoustic context constant (here even physically identical) listeners still did not show shifts in their categorization functions for the generalization trials.

Note that this is in contrast to Kraljic and Samuel's (2006) findings that listeners generalize sub-phonemic cues of stop voicing across place of articulation. The authors argued that durational cues to voicing (i.e., closure duration, aspiration duration) are very similar across different places of articulation, hence listeners may be able learn general properties of duration differences and apply them to all stop voicing contrasts (even produced by different speakers; Kraljic & Samuel, 2007). The fact that durational cues are widely context independent (as presumably duration can apply to any type of segment) may thus facilitate generalization. Cues to place of articulation as used in the present study, can be realized in different ways (cf. Experiment 1) and in the case of formant transitions are distributed over segments that are adjacent to the critical segment rather than located "in" the to-be-recalibrated segment. This may "discourage" generalization. We will come back to this suggestion in Section 5.

From the present results it hence seems that cues as well as the to-be-recalibrated phoneme categories play a role in phonetic recalibration. While these findings speak against abstract context-independent phonemes or phoneme-independent cues as units of recalibration, they leave the allophone in the race as a possible candidate unit for recalibration. As indicated in Experiment 1, the way we conceptualize the allophone here is by the cues that carry the most weight for the distinction at hand. If this conceptualization of the allophone provided the basis for generalization of recalibration, we predict that we should find generalization if both, exposure and generalization trials relate to the same phoneme contrast and are cued in a similar fashion. Specifically we should find generalization if all trials during exposure and test are mainly cued by formant transitions, even if the exact formant contours naturally differ due to a different vowel context. This was tested in Experiment 3.

4. Experiment 3

In Experiment 3 we tested recalibration of the /b/-/d/ contrast in the vowel contexts of /a/ vs. /u/. In both /a/ and /u/ contexts the direction of the formant transitions of $F2$ and $F3$ can cue place of articulation and does so even when the consonantal part is set to silence (as we showed in Text). However, given the differences in formant values that cue vowel identity the exact formant transitions will differ between /a/ and /u/ contexts. In this way we should be able to tease apart whether the lack of generalization in Experiment 1 was due to differences in the context (/a/ vs. /i/) or the additional difference of the cues (formant transitions vs. consonantal portion of the signal). In keeping with the terminology for "naming" the experiments, Experiment 3 would be a same-cue-same-phoneme-different-context experiment.

4.1. Method

Fifty-nine participants who had not participated in Experiments 1, 2, or in the previous pretests took part for partial course credit, 19 in a pretest, 40 in the recalibration experiment. Participants were sampled from the same population using the same restrictions as before.

4.2. Material and procedure

Materials for the a_a condition were identical to Experiment 1. Videos with the speaker articulating the nonsense words /ubu/ and /udu/ were recorded in the same session as the previously used videos. Again, additional high-quality audio recordings for acoustic manipulation were made in a sound-conditioned booth. The editing of videos and audio was similar to Experiment 1. Video and sound/vowel duration was cut to match the a_a condition by removing sound samples from complete glottal periods at randomly selected parts throughout the vowels. To create the /ubu/ to /udu/ continuum, the natural endpoints of $F2$ and $F3$ were rounded and interpolated in 11 steps with transition durations of 70 ms. Endpoint values are given in Table 1. The part in the speech signal usually occupied by closure/voicing and burst/frication was set to silence as had been done for the a_a condition. A pretest asking participants to categorize the auditory /ubu/-/udu/ continuum in a two-alternative forced-choice task confirmed that the continuum was perceived as intended with 95% /b/ responses at the /ubu/ end and 7% /b/ responses at the /udu/ end. Procedure and analyses of the recalibration experiment were identical to the previous experiments. Sixteen participants received the a_a continuum as the recalibration control and u_u as for generalization; 24 received the u_u continuum for the recalibration control and a_a for generalization. The order of blocks was randomized separately for each participant.

4.3. Results

Overall analyses are reported in Table 6. As in previous experiments we find effects of Continuum showing that listeners perceived the three-step continua as ranging from more to less /b/ like. For the a_a context there was an effect of Exposure POA suggesting that overall (across recalibration control and generalization trials) listeners gave more /b/ responses if during exposure they had seen the speaker produce a labial than an alveolar sound. This main effect was not significant for the u_u context. Importantly, in both contexts there was an interaction between Exposure POA and Trial Type suggesting that the effect of Exposure POA was stronger for the recalibration control than the generalization trials. This can be seen in Fig. 5. For the recalibration control trials, the categorization functions after labial and alveolar exposure clearly differ, but they are close together for the generalization trials. Table 7 reports separate analyses for these two trial types. As in the previous experiments, significant effects of Exposure POA were found for the recalibration control continua, but again, no generalization could be found. For the u_u exposure an interaction between Exposure POA and Continuum step suggests that the influence of Exposure POA differed along the steps of the continuum (see Fig. 5; see also the three-way interaction in the overall analysis of the u_u context data). Overall, however, in line with the previous experiments, evidence for generalization could not be found. The "categories" that listeners recalibrate may thus be even narrower than hypothesized before.

Table 6
Overall analyses in Experiment 3, the same-phoneme-same-cue-different-context experiment where participants saw the speaker produce the ambiguous sound in the context of a_a, and the context of u_u. The gray bar highlights the relevant interactions between Exposure place of articulation (labial–alveolar) and Trial Type (recalibration control – generalization).

Factor	aba-ada exposure				ubu-udu exposure			
	b	*SE*	*z*	*p*	*b*	*SE*	*z*	*p*
(Intercept)	0.1	0.23	0.43	0.66	−0.04	0.15	−0.25	0.8
Continuum	−2.24	0.20	−11.1	<0.001	−2.34	0.17	−13.68	<0.001
POA	0.43	0.17	2.54	<0.01	0.13	0.15	0.87	0.38
TrialType	0.93	0.50	1.85	0.065	−0.84	0.34	−2.52	<0.05
Continuum:POA	0.02	0.19	0.09	0.93	0.22	0.18	1.17	0.24
Continuum:TrialType	−0.66	0.34	−1.92	0.055	0.97	0.29	3.39	<0.001
POA:TrialType	1.39	0.3	4.67	<0.001	0.7	0.31	2.24	<0.05
Continuum:POA:TrialType	−0.6	0.43	−1.39	0.17	−0.66	0.32	−2.02	<0.05

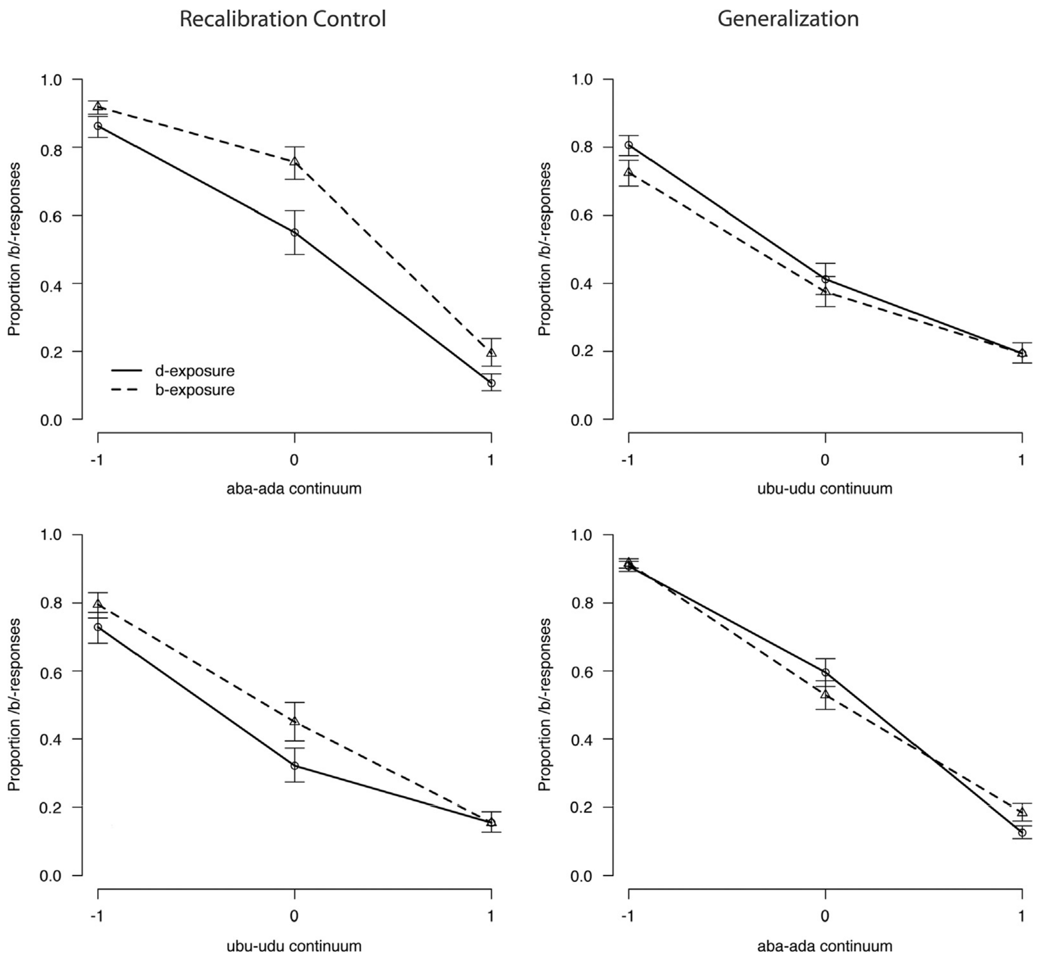

Fig. 5. Results for Experiment 3, the same-phoneme-same-cues-different-context experiment. Proportion labial responses (*y*-axis) are plotted across the test continua (*x*-axis). The dashed lines show responses following labial (i.e., /b/) exposure; the solid lines show responses following alveolar (i.e., /d/) exposure. The “error bars” are based on the Standard Error of the regression weight of Exposure POA in each of the four conditions (see Table 7). The standard error was projected back into the proportion scale, leading to larger intervals around 0.5 and asymmetric intervals at floor and ceiling. The left panels show the recalibration control trials, the right panels show the generalization trials. The upper panels show the condition in which a_a was the exposure context (i.e., formant transitions cues POA of stop consonants), and the lower panels show the condition in which u_u was the exposure context (i.e., formant transitions cue POA of stop consonants but in different vocalic context).

4.4. Discussion

Experiment 3 demonstrated further specificity of phonetic recalibration. We hypothesized that if the categories that listeners recalibrate are something like a particular allophone of a phoneme in the sense that recalibration is cue and phoneme specific, we should find generalization for the /b/-/d/ phoneme contrast in a_a and u_u context where in both cases formant transitions were made the only informative cues to the phoneme contrast. However, listeners did not show category recalibration for generalization trials. There was no indication of generalization to a different vowel context. In contrast, effects for recalibration control trials, as in the previous experiments, were robust despite a somewhat weaker effect of the u_u exposure. Whereas the regression weight of Exposure POA for the a_a condition was 1.11 (t=4.27) the regression weight for the u_u context was

Table 7
Analyses split up by recalibration control trials and generalization trials in Experiment 3, the same-phoneme-same-cue-different-context experiment. Note that Generalization Trials for the a_a exposure condition are u_u trials and vice versa.

Exposure condition	Factors	Recalibration control trials				Generalization trials			
		b	*SE*	*z*	*p*	*b*	*SE*	*z*	*p*
a_a	Intercept	0.59	0.23	2.53	<0.05	−0.37	0.43	−0.87	0.39
	Continuum	−2.64	0.29	−9.09	<0.001	−1.92	0.25	−7.75	<0.001
	ExposurePOA	1.11	0.26	4.27	<0.001	−0.15	0.19	−0.84	0.40
	Continuum*ExposurePOA	−0.2	0.35	−0.56	0.57	0.36	0.24	1.49	0.14
u_u	Intercept	−0.45	0.15	−2.94	<0.005	0.39	0.28	1.41	0.16
	Continuum	−1.84	0.21	−8.8	<0.001	−2.79	0.22	−12.81	<0.001
	Exposure POA	0.47	0.23	2.02	<0.05	−0.24	0.17	−1.42	0.15
	Continuum*ExposurePOA	−0.13	0.19	−0.66	0.51	0.69	0.29	2.35	<0.05

only 0.47 ($t=2.02$).[3] However, this weaker effect of u_u can be explained by the fact that the lip rounding in /u/ somewhat obstructs the visibility of the lip closure and hence diminishes the impact of the information guiding recalibration. This is in line with earlier reports that vowels with lip-rounding (e.g., /u/ and /y/) are themselves visually salient, but reduce the visual salience of the surrounding consonants (Benoit, Mohamadi, & Kandel, 1994; Owens & Blazek, 1985). Nevertheless, the present results once more demonstrate robust recalibration for the recalibration control trials. However, the lack of generalization suggests that perceptual recalibration may be even more specific than previously thought.

5. General discussion

In three experiments, we demonstrated robust effects of visually-guided phonetic recalibration of a place-of-articulation contrast (labial v. alveolar) by replicating the effect with the same (a_a) as well as with different phoneme contrasts, cues, and contexts (iCi, aNa, u_u). To address the question about the units that listeners recalibrate, we tested generalization of recalibration in three different conditions: in Experiment 1 the phonemes were the same but cues were different, in Experiment 2 the phonemes were different but the cues were the same, and in Experiment 3 the phonemes and the cues were the same but the vowel context differed. In all three experiments, we replicated robust effects for the recalibration control trials but no generalization. This suggests that visually-guided phonetic recalibration is quite specific.

Throughout the discussions of Experiments 1 and 2, we suggested that the present data are in line with previous suggestions that the category for recalibration roughly matches an allophone, that is, a specific acoustic implementation of a phoneme (Mitterer et al., 2013). In Experiment 1 listeners did not generalize across different cues to the same phoneme contrast and in Experiment 2 they did not generalize across phoneme contrasts that were cued identically. From these results we predicted that listeners may generalize recalibration if the phoneme contrast and the cues were the same, that is, if the implementation of the phoneme (i.e., the allophone) matched between exposure and (generalization) test trials. Therefore, in Experiment 3, phonemes and cues were made the same (i.e., formant transitions only) and only the acoustic context differed (a_a vs. u_u). However, contrary to our predictions, again no generalization could be found. Hence, phonetic recalibration appears more specific than applying to an allophone of a phoneme as it seems to be restricted to the exposure context.

It might be argued that the lack of generalization cannot be accepted as it is a null-finding (no significant effect of exposure in the generalization trials). Across experiments, however, this seems unlikely. The mean learning effect, calculated as the mean regression weight in the recalibration control trials, is one logit unit. This constitutes a strong effect as logit units are related to Cohen's *d* (Cohen, 1992). In most of our anlayses, even half of that would have led to a significant generalization effect, but none was found. Moreover, the mean generalization effect is 0.03. This all suggests that phonetic recalibration may indeed be constrained by the category contrast and the specific cues providing category information including context.

So what are the categories that listeners recalibrate? On first sight it might seem that it is only the specific ambiguous token heard during exposure that listeners later categorize in line with the previously experienced visual context. However, if this was the case, effects for the recalibration control trials should have been restricted to the middle step of the test continua, that is, the exact tokens heard during exposure. However, this is not what we found. Only in the a_a exposure condition in Experiment 1 there was an interaction between Exposure Place of Articulation and Continuum for the recalibration control trials suggesting significant differences in the recalibration effect (i.e., Exposure POA) for the different steps of the test continuum. In the other experiments, the effect generalized to the other tokens (i.e., the adjacent steps on the continuum) used during the test phase without significantly decreasing in size. This suggests that there is room for at least some variation between the tokens heard during exposure and the ones to be categorized at test. Phonetic recalibration is hence not token specific. Also, the nature of generalization continua could not have been the problem as generalization was tested in both directions and whatever continuum was presented during exposure produced the expected effect for the recalibration control trials.

The apparent context-specificity of recalibration in our study seems puzzling given that a rather large number of previous studies have shown generalization of recalibration across the lexicon (McQueen et al., 2006; Mitterer et al., 2011; Sjerps & McQueen, 2010) or across position in the word (Jesse & McQueen, 2011); conditions in which the phonetic context necessarily differs between exposure and test. The most obvious difference between previous studies finding generalization and the present study is the paradigm used to induce recalibration. Whereas almost all work on generalization (except for generalization across speakers) was based on lexically-guided phonetic recalibration, we used a visually-guided recalibration paradigm to be better able to control for the cues and context that we manipulated. This choice was based on previous evidence that effects of visual and lexical context are similar and serve an equal role in guiding recalibration (Van Linden & Vroomen, 2007). Van Linden and Vroomen compared the two types of recalibration in a paradigm that matched visually-guided recalibration experiments. Whether similar effects could be shown if visual information was used to disambiguate minimal word pairs within a longer list of target and filler words, the classic paradigm of

[3] Note that 8 more participants were tested in the u_u exposure condition than in the a_a exposure condition as after 16 participants the effect for u_u was marginally significant. Having tested 24 participants, however, the effect was significant.

lexically-guided recalibration studies, is the subject of ongoing research. As for now, we have to keep in mind that the monotony of the exposure blocks in visually-guided recalibration paradigms could be the reason for the specificity of the effect. It is not unprecedented in the literature, for example, on learning new phonemes in a second language, that variability during training facilitates later identification and production the trained categories (Logan, Lively, & Pisoni, 1991). Note, however, that paradigm and variability of exposure stimuli tend to be confounded in the literature on phonetic recalibration and until now cross-context generalization has been tested only in lexically-guided recalibration. It therefore could be the case that basic recalibration effect does not appear to depend on variability during exposure (as shown by the vast literature on visually-guided recalibration) but generalization across contexts does. According to this interpretation, the current data would force a re-interpretation of the current thinking, as the two forms of phonetic category recalibration (lexically guided and visually guided) – despite being tested in different experimental paradigms – tend to be viewed as equivalent.

It is nevertheless possible to resolve the apparent conflict between the current finding of specific learning and earlier reports of generalization without assuming that the audio-visual paradigm is to blame. On close inspection, previous reports of generalization include cases of acoustic cues that are highly consistent between exposure and test, such that that learning in these cases may be more specific than previously assumed. Generalization has been reported across speakers and in perceptual adaptation to noise vocoded speech. For the question about generalization of lexically-guided recalibration across speakers, results on whether or not generalization can be found have been mixed (e.g., Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007; Reinisch & Holt, 2014). Whereas recalibration of contrasts in stop voicing (/d/-/t/) appears to generalize across speakers (Kraljic & Samuel, 2007), recalibration of place of articulation in fricatives (/s/ vs. /f/ or /s/ vs. /ʃ/) appears to be more specific (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007). The reason for this discrepancy between generalization behavior across phoneme contrasts and types of cues has been claimed to be the degree to which these cues vary between speakers and hence convey information about the speaker (e.g., Kraljic & Samuel, 2005, 2007). For the stop-voicing contrast the recalibrated cues are widely speaker and context-independent. In addition, the cues of stop closure and aspiration duration could be seen as acoustic units of their own, facilitating generalization. Fricatives, in contrast, tend to vary systematically between different speakers hence they also serve to identify the speaker making generalization less useful (Kraljic & Samuel, 2007). Critically, in the case of fricatives, cross-speaker generalization appears to depend on the sampling of the fricatives in acoustic-perceptual space during exposure and test (i.e., between speakers; Kraljic & Samuel, 2005; Reinisch & Holt, 2014). If the fricatives heard during exposure and test (as well as between the exposure and generalization speaker at test) are perceived to be sufficiently similar then cross-speaker generalization does occur. Note that the type of category contrast used as well as the degree of acoustic similarity are likely an issue in explaining the present results.

The second field of perceptual adaptation in which the degree of specificity has been an issue is adaptation to vocoded speech. That is, listeners are trained to understand a signal in which the original speech signal has been divided into a number of frequency bands in which the amplitude envelope of the signal in these bands is used to modulate a source signal, usually noise (i.e., noise-vocoded speech). Depending on the number of frequency bands used for vocoding, the resulting signal is moderately to severely degraded, compared to the original speech signal. As discussed in Section 1, Dahan and Mead (2010) found that listeners are able to adapt to noise-vocoded speech and generalize adaptation across words. Nevertheless, there was also evidence for specificity of learning in these results. A word at test was recognized better if it had the same diphone as one of the exposure stimuli. That is, the word *boys* was recognized better if participants had heard an exposure stimulus with the same onset and nucleus (e.g., the nonword *baish*). This points towards some sort of specificity or context-dependency. Hervais-Adelman, Davis, Johnsrude, Taylor, and Carlyon (2011) tested generalization across specific properties of the vocoded signal such as frequency range and type of the source signal used for vocoding. They found that listeners generalize adaptation across the frequency range of the signal, that is, from a low-pass filtered signal during training to a high-pass filtered signal at test and the other way around. This suggests that listeners can abstract away from the specific acoustic properties of the signal at least to some extent. Limited generalization across different types of source signal, however, constrains the claim for a general application of adaptation. As with cross-speaker generalization in lexically-guided recalibration, the requirement for generalization of learning for vocoded speech appears to be some sort of acoustic coherence. Together with the specificity of adaptation in Dahan and Mead (2010) the results suggest that there may be two parts to learning to recognize noise-vocoded speech. There may be a general adaptation mechanism through which listeners learn how to deal with this form of input (explaining the generalization reported in Hervais-Adelman et al., 2011). This type of learning may be similar to the perceptual switch that occurs as listeners suddenly perceive sine-wave speech as speech and not as an electronic signal (Remez, Rubin, Berns, Pardo, & Lang, 1994). Next to this general mechanism, another specific learning process may link certain acoustic patterns to specific phone sequences (explaining the specificity in the data of Dahan & Mead, 2010). In studies using edited natural speech, such as the lexically-guided recalibration studies discussed above as well as the present study, it is likely that the latter of these processes constrains generalization. Given the relatively small deviations from fully natural speech it is unlikely that listeners have to adjust to our input in general but recalibration involves learning of specific acoustic patterns.

Specifically, in the present study, lack of generalization might result from a combination of factors that have been discussed for other types of generalization. First, the formant transitions in the vowels are spectral cues that indeed encode extra information about the speaker through their “relative location” (in terms of both a spectrogram and the basilar membrane). So generalization of cues to place of articulation might be more restricted, similar to restricted cross-speaker generalization of fricatives and unlike durational cues to stop voicing in lexically-guided category recalibration. Second, with regard to studies on generalization of recalibration across words and position in the words, for fricatives it has to be noted that the cues to fricatives are mainly located in the fricatives themselves. Even though fricatives differ in their formant transitions, listeners most often rely on the fricative spectrum itself (Wagner, Ernestus, & Cutler, 2006). In many studies on perceptual recalibration, it was even the physically same fricative token that had been spliced into all critical words (as was the case for cross-position generalization in Jesse & McQueen, 2011). Therefore the fricative tokens appeared similar across positions and acoustic contexts. The cues to place of articulation in stops and nasals, in contrast, instantiate themselves in the form of coarticulatory vowel transitions. In addition, the obstruent part includes frequency information in the burst and frication (or the nasal formants. e.g., Repp & Svastikula, 1988). It may thus be the case that phonemes with more distributed cues are stored and recalibrated in a context-sensitive manner. In this sense, the observed pattern of recalibration effects could be explained by recalibration of diphones or triphones (Massaro, 1998; Wickelgren, 1969). Note that this would also explain the specificity of recalibration in Mitterer et al. (2013) as the liquids they used comprise not only articulatorily different types of allophones but the cues are just as distributed (or even more so) as in the segments under investigation here.

With these remarks, we have entered an old debate about the “grain size” of units in speech perception (e.g., Marslen-Wilson & Warren, 1994). McQueen et al. (2006) already argued that the perceptual recalibration paradigm may be used to investigate the nature of pre-lexical representations. There now seems to be accumulating evidence that the grain size of the recalibration units may depend on the distribution of cues across the signal (e.g., in Mitterer et al., 2013, and in the present results; see also the discussion in Kraljic & Samuel, 2007). Given that in our case context-sensitive allophones or n-grams may explain the results, they question theories that assume that the units of speech perception already abstract away from acoustic characteristics at the pre-lexical level. Two theories in particular are questioned by the current findings: theories that assume that the basic unit is the context-invariant phonological feature (Chomsky & Halle, 1968; Lahiri & Reetz, 2002) and theories that assume that not the acoustic but the articulatory features are primary in speech perception (D'Ausilio et al., 2009; Galantucci et al., 2006). According to motor-based theories of speech perception, the primary objects of perception are speech gestures. This means that listeners should learn about the lip closing gestures in our experiments, and a different acoustic implementation of these gestures (e.g., stop vs. nasal) should not matter for perceptual recalibration. The current data hence fit better with the assumption that acoustic/auditory characteristics of the speech signal are primary in speech perception.

A similar case can be made about theories that assume lexical access to be achieved by extracting context-invariant features. Such features had been proposed in Phonology (e.g., Chomsky & Halle, 1968) and are part of a contemporary model of spoken-word recognition (the Featurally-Underspecified Lexicon (FUL) model by Lahiri & Reetz, 2002). For Experiment 1, the argument could be made that the feature-detectors simply learn a specific acoustic implementation of the feature [LABIAL] during exposure. Failure to generalize from the a_a to the iCi context could then be explained by the specific associations between certain acoustic cues and the feature. However, this explanation does predict generalization from a_a to aNa in Experiment 2 (and potentially also Experiment 3), because in this experiment the same acoustic cues (formant transitions) were used to cue labiality, and only the manner of articulation (or the specific acoustic context) differed. The FUL model explicitly argues that features are context-independent (Lahiri & Reetz, 2002) and hence necessarily predicts generalization here. So, indeed the perceptual recalibration paradigm can inform us about what the categories of speech perception can or cannot be.

This conclusion rests, however, on the assumption that the visually-guided recalibration reflects a general speech-perception mechanism (an assumption empirically supported by Van Linden & Vroomen, 2007). The specificity of learning in our results may also cast doubt on the widely-held belief that recalibration is a general processing mechanism that does its work wherever the knowledge comes from (see, e.g., Kraljic & Samuel, 2007, 2: “Visual context can also serve to constrain the interpretation of phonemes, and therefore results in perceptual learning that is comparable to Norris, et al.'s original finding”) and whatever the modality of recalibration is (Mitterer & de Ruiter, 2008). Our results show that it may be useful to test this assumption more thoroughly.

To summarize, in a series of perceptual recalibration experiments in which acoustic cues and contexts were tightly controlled we showed that listeners recalibrate perception of the exposure contrast for stops and nasals involving various types of acoustic cues. We demonstrated that visually-guided perceptual recalibration is specific to the exposure phonemes, cues, and context. From a methodological point of view, this encourages further research into the mechanisms of category recalibration using different types of context information to drive the learning. This would constitute a further test whether this specificity is limited to visually-guided recalibration. From a theoretical perspective the results of the present study suggest 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). Moreover, their grain size (segmental, di- or triphones) may be adjusted to the extent that information about the phoneme contrast is spread out into adjacent segments.

Acknowledgments

The main part of this research was conducted while the first author was working at the Department of Psychology and associated with the Center for the Neural Basis of Cognition at Carnegie Mellon University. She is now at the Department of Phonetics and Speech Processing at the University of Munich (LMU). This research was supported through grants R01 DC004674 (NIH) and 0746067 (NSF) to Lori L. Holt. Thanks go to Patrick van der Zande and Dan Acheson for help with the video recordings, and the research assistants of the Holt lab at Carnegie Mellon University for help with testing participants.

References

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.
Benoit, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research*, *37*(5), 1195–1203.
Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*, 233–252.
Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*, 592–597.
Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.
Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NJ: Harper & Row.
Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics*, *70*(4), 604–618, http://dx.doi.org/10.3758/pp.70.4.604.
Cohen, J (1992). A power primer. *Psychological Bulletin*, *112*, 155–159, http://dx.doi.org/10.1037/0033-2909.112.1.155.
Dahan, D., & Mead, R. L. (2010). Context-conditioned generalization in adaptation to distorted speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 704–728, http://dx.doi.org/10.1037/a0017449.
D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, *19*(5), 381–385.
Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238, http://dx.doi.org/10.3758/BF03206487.
Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, *99*, 1730–1741, http://dx.doi.org/10.1121/1.415237.
Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin and Review*, *13*(3), 361–377, http://dx.doi.org/10.3758/BF03193857.
Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.
Harding, S., & Meyer, G. (2003). Changes in the perception of synthetic nasal consonants as a result of vowel formant manipulations. *Speech Communication*, *39*(3), 173–189.
Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 283–295.
Hyman, L. M. (1975). *Phonology*. New York: Holt, Reinhart, and Winston.
Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, *18*, 943–950, http://dx.doi.org/10.3758/s13423-011-0129-2.
Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal?. *Cognitive Psychology*, *51*, 141–178, http://dx.doi.org/10.1016/j.cogpsych.2005.05.001.
Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, *13*, 262–268, http://dx.doi.org/10.3758/BF03193841.
Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*, 1–15, http://dx.doi.org/10.1016/j.jml.2006.07.010.
Kraljic, T., Samuel, A. G., & Brennan, S. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, *19*, 332–338, http://dx.doi.org/10.1111/j.1467-9280.2008.02090.x.
Kurowski, K., & Blumstein, S. E. (1984). Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. *The Journal of the Acoustical Society of America*, *76*, 383.
Lahiri, A., & Reetz, H. (2002). Underspecified recognition. In: C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology*, 7 (pp. 637–676). Berlin: Mouton de Gruyter.
Liberman, A. M. (1996). *Speech: A special code*. Cambridge, Mass: MIT Press.

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*(2), 874–886.

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

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86, http://dx.doi.org/10.1016/0010-0285(86)90015-0.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

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

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

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

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

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361, http://dx.doi.org/10.1016/j.cognition.2013.07.011.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian Model of Continuous Speech Recognition. *Psychological Review*, *115*, 357–395, http://dx.doi.org/10.1037/0033-295X.115.2.357.

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

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech, Language and Hearing Research*, *28*(3), 381.

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

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

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*, 129–156.

Repp, B. H., & Svastikula, K. (1988). The perception of the [m]-[n] distinction in VC syllables. *Journal of the Acoustical Society of America*, *83*, 237–247.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*, 1207–1218, http://dx.doi.org/10.3758/APP.71.6.1207.

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, http://dx.doi.org/10.1037/a0016803.

Smits, R., Bosch, L. ten, & Collier, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment. *The Journal of the Acoustical Society of America*, *100*(6), 3852–3864, http://dx.doi.org/10.1121/1.417241.

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.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory–visual speech perception. *Neuropsychologia*, *45*(3), 598–607.

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the “unity assumption” using audiovisual speech stimuli. *Perception & Psychophysics*, *69*, 744–756.

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.

Wagner, A., Ernestus, M., & Cutler, A. (2006). Fricative inventory and the relevance of formant transitions for fricative identification. *Journal of the Acoustic Society of America*, *120*, 2267–2277.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, *76*, 1–15.