Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-019-01961-8

Interleaved lexical and audiovisual information can retune phoneme boundaries

Shruti Ullas[1] · Elia Formisano[1] · Frank Eisner[2] · Anne Cutler[3]

Abstract
To adapt to situations in which speech perception is difficult, listeners can adjust boundaries between phoneme categories using perceptual learning. Such adjustments can draw on lexical information in surrounding speech, or on visual cues via speech-reading. In the present study, listeners proved they were able to flexibly adjust the boundary between two plosive/stop consonants, /p/-/t/, using both lexical and speech-reading information and given the same experimental design for both cue types. Videos of a speaker pronouncing pseudo-words and audio recordings of Dutch words were presented in alternating blocks of either stimulus type. Listeners were able to switch between cues to adjust phoneme boundaries, and resulting effects were comparable to results from listeners receiving only a single source of information. Overall, audiovisual cues (i.e., the videos) produced the stronger effects, commensurate with their applicability for adapting to noisy environments. Lexical cues were able to induce effects with fewer exposure stimuli and a changing phoneme bias, in a design unlike most prior studies of lexical retuning. While lexical retuning effects were relatively weaker compared to audiovisual recalibration, this discrepancy could reflect how lexical retuning may be more suitable for adapting to speakers than to environments. Nonetheless, the presence of the lexical retuning effects suggests that it may be invoked at a faster rate than previously seen. In general, this technique has further illuminated the robustness of adaptability in speech perception, and offers the potential to enable further comparisons across differing forms of perceptual learning.

Keywords Phoneme boundary · Recalibration · Perceptual retuning · Lexical · Audiovisual

Introduction

Listeners often encounter situations where they must understand a speaker they have never heard before, and must rapidly adapt to the unique acoustic characteristics of the individual's speech. In such scenarios, information other than the auditory signal itself can be utilized to assist the listener and can influence listeners' interpretation of what they are hearing. Early studies demonstrated that knowledge of the lexicon and speech reading can create an immediate bias to what listeners perceive (Ganong, 1980; McGurk & MacDonald, 1976). More recent studies of perceptual retuning have shown that listeners can learn to disambiguate speech or speech-like sounds, by adjusting the boundary of a phoneme category and expanding the criteria used to identify a phoneme. Both lexical and speech-reading information have been established as sources that can facilitate this process, and thus enable the famously robust adaptability of human speech perception (Cutler, 2012; Vroomen & Baart, 2012).

In the initial experiments on perceptual retuning, listeners heard and viewed speech or speech-like stimuli edited to remove clear instances of a critical phoneme, which were then replaced by an ambiguous phoneme blend nearly indistinguishable from a natural version (Bertelson et al., 2003; Norris et al., 2003). In lexically guided perceptual learning, recordings of words ending in a particular phoneme (e.g., /s/, as in *carcass*) are edited to end in an ambiguous phoneme instead, such as an /s/-/f/ blend (Norris et al., 2003; Samuel & Kraljic, 2009). Following exposure to such stimuli, listeners perform a categorization task on the ambiguous token and

✉ Shruti Ullas
shruti.ullas@maastrichtuniversity.nl

[1] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6200 MD Maastricht, The Netherlands

[2] Donders Centre for Cognition, Radboud University Nijmegen, 6500 AH Nijmegen, The Netherlands

[3] MARCS Institute and ARC Centre of Excellence for the Dynamics of Language, Western Sydney University, Penrith, NSW 2751, Australia

Published online: 22 January 2020

other neighboring sounds along an /s/-/f/ continuum and are likely to report hearing more sounds in accordance with the preceding exposure stimuli (i.e., as an /s/). Listeners are also likely to perceive an /s/-/f/ blend as /f/, if they hear recordings of /f/-final words (e.g., *paragraph*) with the ambiguous token replacing the /f/. Likewise, in visually guided recalibration, participants are presented with video recordings of a speaker pronouncing a syllable (/aba/ or /ada/) paired with an audio recording of an ambiguous token (/aba/-/ada/ blend) (Bertelson et al., 2003). After sufficient exposure to these videos, participants perform a categorization task on the ambiguous token, and are also likely to report perceiving it as the phoneme it was replacing (as /aba/ if coupled with videos of /aba/, or as /ada/ with videos of /ada/). Note that we have used recalibration here to refer to the audiovisual form, retuning to refer to the lexical version, and perceptual learning when referring to both. This is in agreement with the terminology used by the researchers who have developed and deployed the two approaches, and we will maintain the distinction throughout our report for the convenience of the reader.

The lexical and visual approaches are certainly similar in that they both reveal how internal representations of speech sounds can be reshaped during perceptual experience by reference to existing knowledge. However, despite this similarity in the resulting effects, the course of the learning can vary across other dimensions, such as build-up and dissipation, or the extent to which the effects are still measureable. Lexical retuning studies typically use longer exposure phases with critical items embedded into a lexical decision task or other listening material containing filler words as well, while audiovisual recalibration studies often repeat videos of a single syllable, and eight exposure tokens can be enough to induce after-effects (see Samuel & Kraljic, 2009, for an overview). Eisner and McQueen (2006) have shown that the retuning effects from lexical information can be present up to 12 h after exposure, both during the daytime or after a night of sleep, while Baart and Vroomen (2009) noted that audiovisual recalibration effects can quickly diminish with increasing numbers of items during the follow-up categorization task, and are not observable after 24 h.

Van Linden and Vroomen (2007) sought to quantify these differences between lexical and audiovisual perceptual learning by exposing participants to both forms in two separate sessions, with the categorization task immediately following each such exposure phase. Retuning effects were larger after audiovisual exposure than after lexical exposure, but could build up and dissipate in a similar fashion when the exposure and test phases were structured consistently.

What is as yet unknown is whether both forms of perceptual learning can be called upon within the same circumstances and under the same experimental constraints. Perceptual systems must be flexible so as to accommodate possible variability in speech, so listeners should be capable of switching between available contextual cues depending on the needs of the situation, but, conversely, may also find that switching between two cue types does not allow perceptual learning effects to build up sufficiently. The present study addresses this question by comparing perceptual learning effects following lexical and visual/speech-reading exposure, both within participants and within a single session. In order to compare them within a single session, the study also explored whether lexical retuning can take place under more restricted conditions, with short exposure blocks in two possible biasing directions, rather than a long exposure pointing towards only one phoneme. Following brief exposures to stimuli ending in an ambiguous phoneme (a /p/-/t/ blend), wherein the direction of the bias was changing throughout the session, participants were expected to continuously adjust the phoneme boundary between two clear phonemes, based on their responses during categorization tasks on ambiguous phoneme blends. The same procedure for both audiovisual recalibration and lexical retuning was maintained in order to compare them directly. It further allowed us to determine whether lexical retuning was possible under conditions more typical of audiovisual recalibration, by presenting only eight items per exposure block. The design, adapted from van Linden and Vroomen (2007), incorporated pseudo-words and words for audiovisual and lexical perceptual learning, respectively, by presenting interleaved exposure blocks of the two types of stimuli, each followed by test blocks containing ambiguous phonemes without context.

Methods

Participants

Sixty healthy native Dutch speakers were recruited from Maastricht University. All participants (37 female and 23 male; mean age = 22 years, standard deviation = 3 years) had normal hearing, normal or corrected-to-normal vision, and received study credits or monetary compensation for participating. The study was approved by the university ethical research board. Participants were randomly selected to be in one of three groups: exposure to audiovisual/speech-reading stimuli, to lexical stimuli, or to both.

Materials

The materials for the experiment were modeled on those used previously by van Linden and Vroomen (2007). A digital audio and video of a female native Dutch speaker was recorded in a sound-proof booth. Recordings of the syllables /op/ and /ot/ were made, as well as a set of 16 Dutch words (e.g., *siroop,* 'syrup'; or *walnoot*, 'walnut') and 16 pseudo-words (e.g., *miloop, geroot*). The words varied in number of syllables

and stress pattern and contained a range of segments, and the pseudo-words were matched in these respects to the real words, thereby creating varying input that could counteract possible selective adaptation effects from repetitive stimuli (Vroomen et al., 2007). All items were recorded with both /op/ and /ot/ endings.

A ten-step continuum ranging from clear /op/ to clear /ot/ was created using the Praat speech-editing program (Boersma & van Heuven, 2001), adapted from a procedure devised by McQueen (1991), based on earlier work by Repp (1981). The endpoints of the continuum were excised from two recordings of the Dutch pseudo-words /soop/ and /soot/ with equal durations and a sampling frequency of 44 kHz. To prepare the continuum, the durations of the consonant (plosive) bursts of /op/ and /ot/ were spliced out and equated to 186 ms, and the averaged pitch contour was calculated to replace the original. The intermediate sounds were created by concatenating the amplitudes of waveforms in 10% increments with each token after the first (e.g. 90% /op/ with 10% /ot/, etc.). The preceding vowels of the two tokens were equated to 50 ms and also interpolated using the same procedure as with the consonants. As a result, the second and third formants of the vowel were systematically decreased from the /ot/-token to the /op/-token. All items of the continuum were then spliced onto a recording of /soo/, resulting in ten items varying from /soop/ to /soot/. Multiple sets of lexical and audiovisual stimuli, or words and pseudo-words, respectively, were created with the middle steps of the continuum (steps 4, 5, 6, 7, and 8), which were most likely to be perceived as most ambiguous. These sounds were spliced into the stimuli at the zero-crossing closest to the last 50 ms of the vowel preceding the final consonant, to eliminate any co-articulatory cues from the preceding vowel. The appropriate stimuli set was individually chosen for each participant, during a categorization pre-test prior to the experiment, based on the sound perceived as /op/ or /ot/ for 50% of the responses, or as close as possible.

Lexical stimuli Lexical stimuli were 16 Dutch words with word-final voiceless stop consonants, eight ending in /op/ and eight ending in /ot/. In the edited versions, the final phoneme was replaced with the ambiguous phoneme blend. Each set of eight contained one monosyllable, three disyllables, and three trisyllables. Stimuli lasted 1,300 ms on average with a standard deviation of 160 ms. /p/-final words had an average word frequency of 421 per million, while /t/-words had an average word frequency of 367 per million.

Audiovisual stimuli Audiovisual stimuli consisted of 16 videos of a speaker pronouncing Dutch pseudo-words, which were matched with the lexical stimuli for number of syllables. Pseudo-words were created using the program WinWordGen 1.0 for Dutch (Duyck et al., 2004), and were recorded with videos centered around the mouth of the speaker. The edited audio recordings containing the ambiguous final phoneme replaced the original audio of the video recordings**.** Based on the speaker's lip movements, eight of the videos indicated an /op/ ending, and the other eight an /ot/ ending. Each video lasted 1,400 ms on average with a standard deviation of 100 ms and no stimuli were longer than 1,500 ms. Videos were approximately 24 frames per second with 1,920 × 1,080 pixels per frame.

Procedure

Participants were individually tested in a sound-proofed room. Stimuli were delivered using Presentation software (Neurobehavioral Systems, Inc.) and sound stimuli were presented through Philips Sensimetric earphones at a comfortable listening volume. Participants first underwent a pre-test in order to determine the step of the /op/-/ot/ continuum perceived to be most ambiguous. The items of the continuum were presented 100 times in total, with more presentations of medial steps than endpoints. For each sound, participants indicated with a button press if the sound resembled /ot/ or /op/. The step of the continuum reported as /op/ or /ot/ for 50% of trials, or as close as possible, was used to determine the appropriate stimuli set to use in the exposure blocks of the experiment, as well as the sound used during the test blocks. All participants' perceived midpoints ranged between steps 4 and 8. All of the audio endings of the audiovisual and lexical stimuli would contain the individually selected ambiguous token. Individual ambiguous-token selection (as typically used for audiovisual studies since Bertelson et al., 2003) ensures that each participant will receive an equivalently effective stimulus set, but direct comparisons have shown that the perceptual learning process is unaffected by the choice between this method versus the simpler method (as typically used for audio-only studies since Norris et al., 2003) in which all participants receive the same ambiguous stimulus based on a pre-test with a separate group of listeners (Bruggeman & Cutler, 2019).

Once the appropriate midpoint and its corresponding stimuli were selected, participants began the main experiment, which consisted of 32 blocks in total, each block beginning with eight exposure stimuli, followed by six test stimuli. Four unique exposure stimuli were presented, each repeated twice, and within a block, all had either /op/ or /ot/ endings so as to induce a bias in one direction at a time. For lexical stimuli, a gray fixation cross was present on the screen, while a black screen was present between video clips during audiovisual exposure. Each exposure trial lasted 1,600 ms in total, including the sound/video presentation and a brief silence. During the test phase, the ambiguous token from the continuum and its two neighbors (one more /op/-sounding, the other closer to /ot/) were each presented twice. After each sound presentation, the participants were prompted to respond with a button

press to indicate the sound it most resembled (/op/ or /ot). Blocks were presented in pseudo-random order, where no more than two blocks with the same phoneme bias followed one another. Participants were randomly assigned to the three possible experimental conditions (lexical stimuli only, audiovisual stimuli only, or both types of stimuli). In the third group, blocks contained either lexical or audiovisual stimuli, and order was counterbalanced, such that the stimulus type changed every four blocks. For all three groups, the phoneme bias switched every one or two blocks. In total, 256 exposure trials were presented (for the third group, 128 of each exposure type), and 192 test trials. Examples of the testing procedure are shown in Fig. 1.

Blocks alternated between presenting exposure and test stimuli. Exposure blocks consisted of eight items, either audio recordings of words or videos of pseudo-words, inducing a bias towards either /op/ or /ot/ during each block. Two groups received only one of the two types of stimuli (audiovisual or lexical), while a third group was presented with both types of stimuli (changing every four blocks). Each exposure block was followed by a test block containing the most ambiguous sound along the continuum, and its two perceptual neighbors, to which listeners responded depending on whether it was perceived as /op/ or /ot/.

Results

Pre-test results

All participants underwent a pre-test to determine the most ambiguous sound along the /op/-/ot/ continuum. On average, the seventh step was closest to the 50% perceptual midpoint and was the most frequent choice across participants. Pre-test results averaged across participants over the ten steps are shown in Fig. 2.

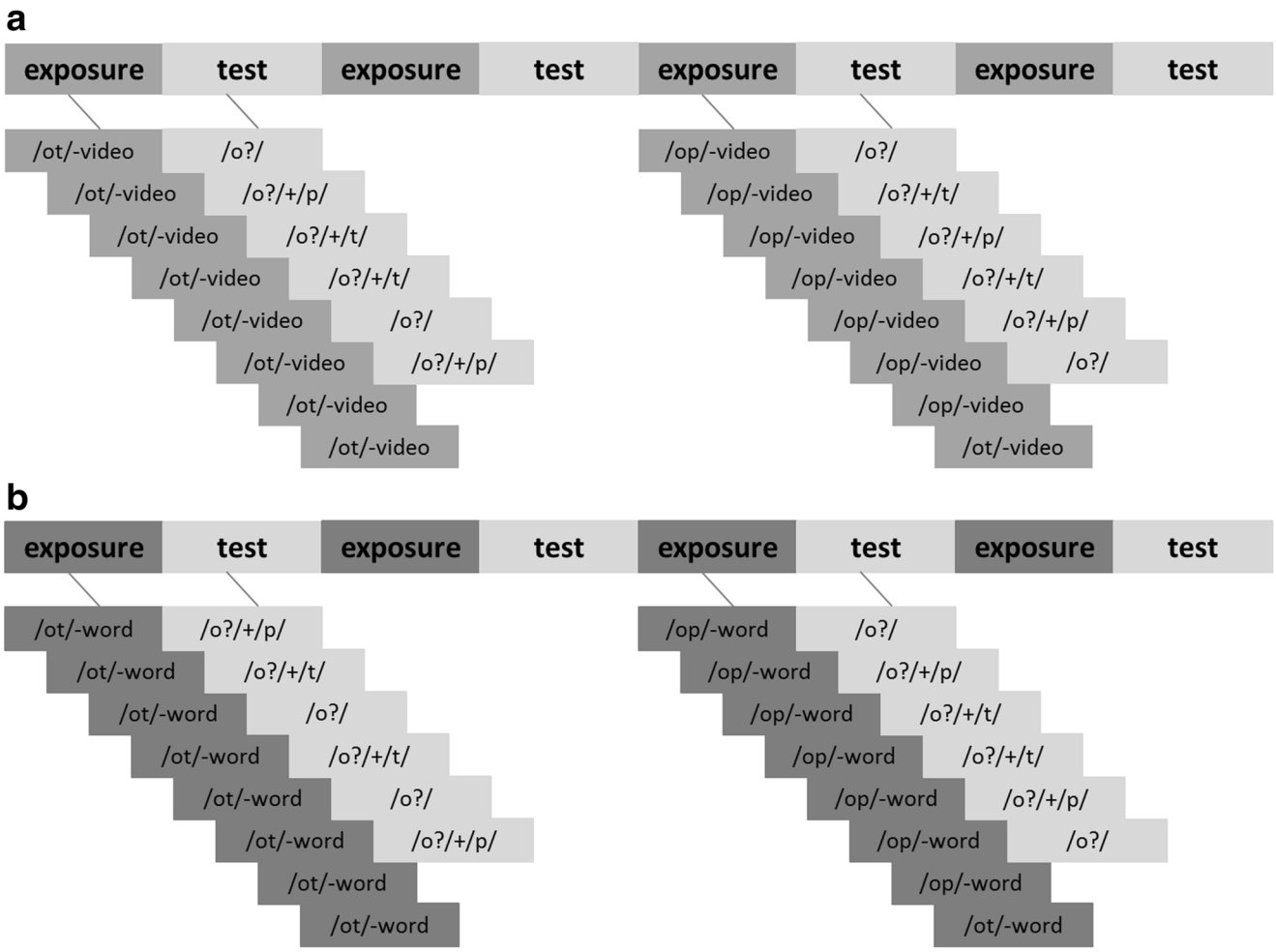

Fig. 1 Examples of the testing procedure. Participants received audiovisual (**a**), lexical (**b**), or both types of stimuli during exposure blocks, followed by test blocks. Participants who underwent single exposures would follow the procedure outlined in either panel A or B repeatedly for 32 blocks, while the third group received both A and B for the duration of the experiment. Any given exposure block aimed to elicit a bias towards either /p/ or /t/. Test items were randomized for every block

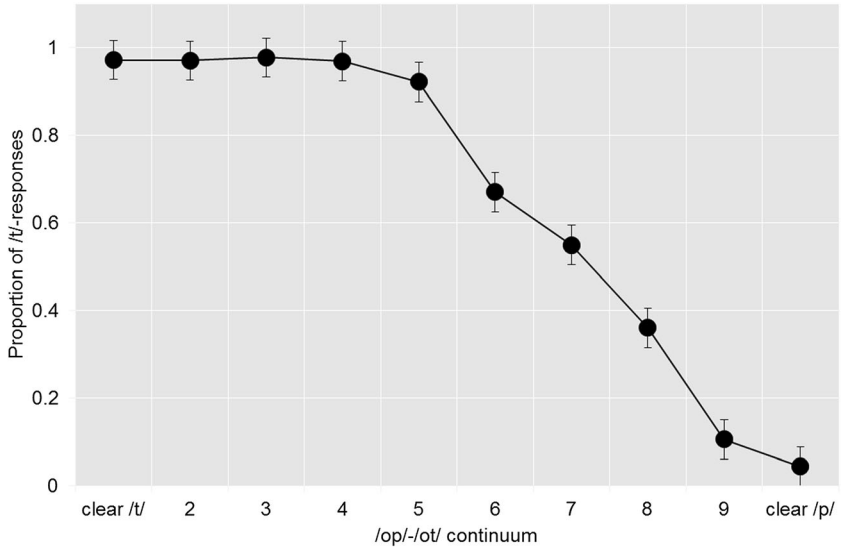

Fig. 2 Pre-test results. Proportions of /t/-responses for each of the ten continuum sounds presented during the pre-test, averaged across all participants (n=60)

Perceptual learning results

Results were analyzed using the statistical package, *R,* with the lme4 library. All variables were entered into a generalized linear mixed-effects model with a logistic linking function for a binomial distribution. Four independent variables were entered into the model. *Phoneme bias* referred to the direction of the bias induced by the stimuli, being either /op/ or /ot/, while the *conditions* were either lexical or audiovisual. One out of the three participant groups was exposed to both audiovisual and lexical stimuli, while the other two groups only underwent one form of exposure, so the model accounted for this with a variable of *switch,* by coding the two single exposure groups as one value and the third group (double exposure) as another.

A variable was included for the three different *sounds* used during the test phases; the most ambiguous sound (selected during the pre-test) and its two surrounding neighbors from the continuum. Finally, the serial *block position* was also included, to see whether retuning effects varied from the start to the end of the experiment. All variables were numerically coded to be centered around 0. *Phoneme bias, condition,* and *switching* were entered as fixed effects, while the within-subject factors *phoneme bias, sounds, block position,* and an additional variable of *subject* were included as random effects as well. The dependent variable was the response to the test tokens, with "0" and "1" representing /op/ and /ot/, respectively. A maximal model containing all variables was created, as well as random slopes for all within-subjects variables and their interactions. The resulting model of best fit was: Response ~ 1 + Phoneme bias * Condition * Switching * Sound * Block position + (1 + Phoneme bias * Sound * Block position || Subject). Fixed-effects correlations were checked to ensure the validity of the model, and all were less than 0.2.

The model showed a significant negative effect of the intercept, or general tendency to respond with /p/ across all test blocks. A significant main effect of *phoneme bias* and significant interactions between *phoneme bias* and *condition, block position*, and *phoneme bias,* and between *block position, phoneme bias,* and *condition*, were also found.

The main effect of *phoneme bias* revealed that more /t/ responses were seen after blocks biased towards /t/ than blocks biased towards /p/, which confirmed that listeners showed perceptual learning effects after audiovisual and lexical exposure. Bonferroni-corrected pairwise contrasts were performed on the factors in the three-way interaction, between *block position, phoneme bias,* and *condition*. Significantly more /t/-responses were found after /t/-biased blocks than /p/-biased blocks in the audiovisual condition than in the lexical condition (see Fig. 3). More specifically, significant differences between /t/-responses following /p/- and /t/-biased blocks were found across all block positions in the audiovisual condition ($p<0.002$), and for all blocks in the lexical condition ($p<0.05$), although slightly less at the first block ($p=0.06$). According to the model results, perceptual learning effects did not vary significantly across the testing session in either condition, although a statistically non-significant reduction in audiovisual recalibration was found from block 5 to block 6 (see Fig. 4). As no significant main effects were found for the remaining factors of *switching* or *sound*, we concluded that perceptual learning effects did not vary due to either of these factors.

Discussion

In the present study, listeners could adjust phoneme boundaries using both lexical and audiovisual information, and

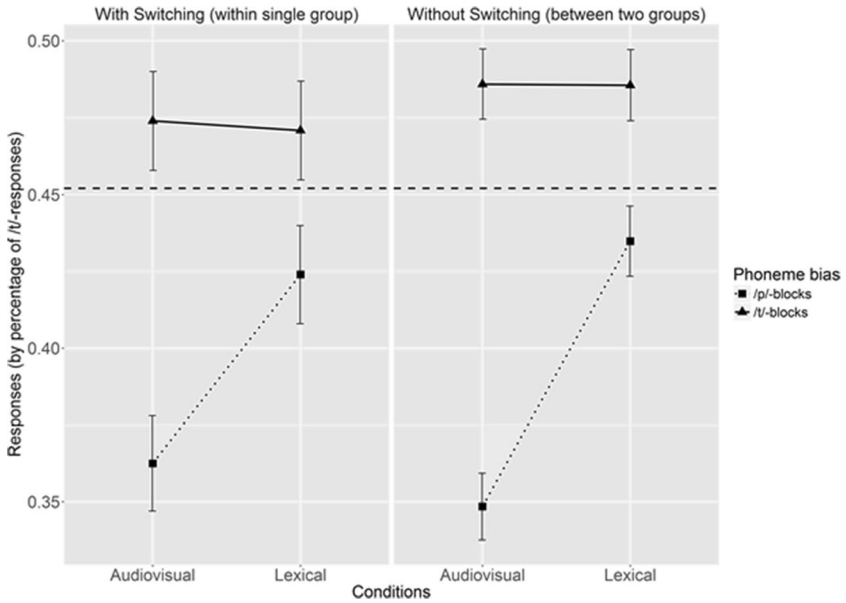

Fig. 3 Audiovisual and lexical perceptual learning effects. Proportions of /t/-responses collapsed across the three test sounds, split by group that received both exposures (**left panel**) and single-condition groups (**right panel**), and separated by exposure type. The dashed line indicates the pre-test average of /t/-responses over all participants to the individually selected midpoint (=0.4528)

switch between these two sources of information within a session. Comparison groups that underwent only one form of exposure showed similar levels of after-effects to the group that received both exposure types. Although the interleaved exposure blocks could have potentially led to interference between the two forms of perceptual learning, no such deficit was shown. Audiovisual recalibration and lexical retuning thereby appear to be separate processes, and do not necessarily interact with each other even while being measured in alternation and with the same phoneme pair. Neither form of perceptual learning showed significant variation over the course of the experiment, with the exception of lexical retuning effects at the first test block. A reduction in audiovisual recalibration was found between the fifth and sixth test blocks (from a 15% to an 8% difference in subtracted /t/-responses), although it was not statistically significant. While audiovisual recalibration was more robust overall, it appears that the effects may not be sustained with increasing numbers of test blocks, perhaps due to fatigue with repeated testing. Vroomen et al. (2004) have also reported reductions in audiovisual recalibration with increasing numbers of test items. Nevertheless, perceptual learning effects were largely stable

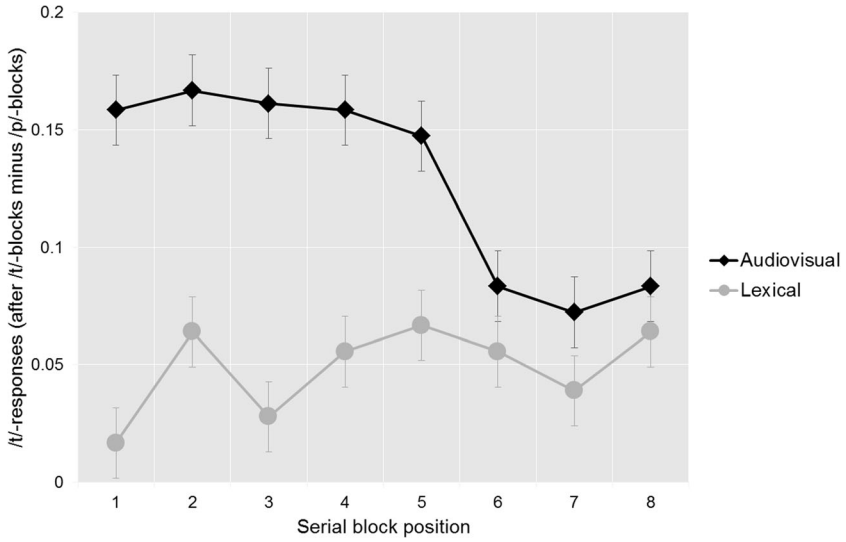

Fig. 4 Subtracted perceptual learning effects from first to last block. The subtracted difference in the percentage of /t/-responses for each block are shown (/t/-responses following /t/-biased blocks minus /t/-responses after /p/-biased blocks). Lexical- and audiovisual-only groups are collapsed every two blocks, while responses from the double-exposure group are averaged per block

throughout the testing session, and short and alternating exposures still led to observable effects on a block-to-block basis.

The experimental design used in the present study is in several ways more common in audiovisual recalibration experiments than in audio-only lexical retuning experiments. As noted, the two types of task typically differ in whether the ambiguous sound is customized to the individual participant, as in the present case, or is based on a separate pre-test with a separate participant group, as in most lexical retuning studies; but the two ambiguity determination methods have been shown to produce equivalent learning effects (Bruggeman & Cutler, 2019). In addition, lexical retuning studies commonly use longer exposure phases combined with a distractor activity, such as a lexical decision task, a counting task, or listening to a story (see Cutler et al., 2010, for an overview), and only induce a bias towards one particular phoneme, instead of repeatedly changing the bias direction (Kraljic & Samuel, 2009). Lexical retuning effects with such designs have been found to be robust and even measurable up to 12 h later (Eisner & McQueen, 2006). Lexical retuning effects in the present study may not have been as pronounced due to the experimental design, as listeners were continuously adapting the category boundary in two opposing directions. Although it is therefore arguable that such a design may be more suitable for audiovisual recalibration and may not have been optimal for inducing lexically driven retuning, perceptual shifts in all conditions were still clearly evident.

The interleaved design still allowed lexical information to adjust phoneme boundaries using the same phoneme pair in either direction, with no reduction resulting from switching between exposure types, or due to short exposure blocks (which may not have given listeners adequate time to allow effects to accumulate). Kraljic and Samuel (2007) have reported that lexical retuning can take place in a speaker-specific manner, such that one particular phoneme pair is adjusted with one speaker, and another pair with another speaker, befitting the role of retuning in social conversations with potentially many participants. Similarly, the flexibility of lexical retuning observed in the present study is consistent with the hypothesized value of lexical retuning for ensuring such adjustment to newly encountered interlocutors is rapid. Audiovisual recalibration can occur between multiple speakers (Mitchel, Gerfen, & Weiss, 2016) and even in two different directions by each ear (Keetels, Pecoraro, & Vroomen, 2015; Keetels, Stekelenburg, & Vroomen, 2016). The study design was adapted from van Linden and Vroomen (2007), who also reported that lip-reading pseudo-words led to recalibration, but in the present study, could take place while interleaved with lexical retuning. Note that the use of pseudo-words and interleaved exposure in our study may be the source of the lack of significance between test sounds (e.g., most /t/-responses for the most /t/-sounding token, etc.). Pseudo-words, rather than single syllables, were less specific to the phoneme at hand and could have led to a minor detraction in sound-specific recalibration.

The interleaved design of the present study would also lend itself well to neuroimaging studies. With the advancement of neuroimaging techniques such as functional MRI (fMRI), this design allows for exploration of the neural underpinning of multiple phoneme percepts induced by multiple cue types, all while presenting the same acoustic token during and after various contextual conditions. The paradigm could be used to explore how other phoneme pairs may fare, and how the learning effects would vary depending on the types of phonemes being manipulated (i.e., plosives/stops vs. fricatives).

Audiovisual information proved more effective than lexical cues in inducing subsequent retuning effects, in line with prior findings (Lüttke et al., 2018; Mitterer & Reinisch, 2017; van Linden & Vroomen, 2007). This difference is predicted given the visual salience of the /p/-/t/ contrast (a bilabial vs. an alveolar plosive) compared to the subtlety of the auditory difference between the same two sounds (both voiceless, both plosive). Any potential advantage to lip-reading cues is thus tied to the phonemes at hand, as they must be visually distinguishable in order for audiovisual cues to be a source of guidance. Prior studies have noted variation in the nature of lexical retuning across phoneme pairs in audio-only presentations (Kraljic & Samuel, 2007), as well within-pair differences in shift effect size (Cutler et al., 2010); other contrasts may display varying patterns of relative effect. Notably, the difference in the magnitude of audiovisual and lexical perceptual learning effects in the present study was largely due to the difference in responses after /p/-biased blocks. The proportions of /t/-responses after audiovisual and lexical /t/-biased blocks were rather similar, whereas audiovisual /p/-blocks elicited fewer /t/-responses than lexical /p/ blocks. This strong /p/ response in the audiovisual /p/ blocks is as expected; not only is the /p-/t/ distinction visually salient, this salience is effectively carried by the /p/, so that the audiovisual contrast effectively amounts to plus versus minus lip closure. The possibility remains that the lexical information contained in the /p/-biased blocks may not have been as effective in inducing a shift in perception as the lexical information in the /t/-biased blocks; and as previously mentioned, each individual phoneme can vary in the extent that its boundary can be shifted by contextual cues. However, the reliability of the lip cues to /p/ for conversational participants is evidently the strongest effect.

The asymmetry between the sizes of the observed lexical and audiovisual retuning effects highlights how their intrinsic purposes may differ. Lexical cues can lead to retuning in response to static speaker characteristics that are unlikely to change, such as accents or idiosyncratic pronunciations unique to a particular speaker (Cutler et al., 2010). A speaker's pronunciation of a particular word is unlikely to change within a short amount of time. Lexical retuning effects may be more optimal in one particular direction, as was indeed seen in this

study. In contrast, recalibration driven by speech-reading may be particularly useful and reliable in environmental circumstances that are not tied to a specific speaker, such as the presence of noise (Macleod & Summerfield, 1987; Massaro & Jesse, 2007; Sumby & Pollack, 1954). Thereby, the retuning resulting from audiovisual cues may be more malleable and more easily reconfigured across phonemes. In real-world scenarios, this means that listeners can attend to cues according to the needs of the situation, but are capable of switching between the two if required, as is suggested by the results of this study.

As noted in the *Methods* section, the materials were designed to avoid selective adaptation effects, which typically occur when listeners have undergone repeated exposure to a clear sound, but as a result are likely to perceive similar ambiguous sounds as a contrasting phoneme to the original (Eimas & Corbit, 1973). For example, after repeated presentations of clear auditory /op/, sounds on a continuum of /op/-/ot/ are more likely to be perceived as /ot/ than as /op/, i.e., the reverse of the exposure (Kleinschmidt & Jaeger, 2016; Vroomen et al., 2004, 2007). Selective adaptation can thus be viewed as the opposite of perceptual learning effects. Interestingly, one previous study (Samuel 2001) found that listeners who underwent short exposures to ten words containing an ambiguous phoneme, similar to the design of the present study, showed selective adaptation effects during the subsequent test phases (ambiguous tokens presented without context). In this particular case, it is possible that the stimuli involved were insufficiently ambiguous, and could have been perceived as clear phonemes even when embedded in mismatching stimuli; this could potentially have induced a contrasting percept for a subsequently presented isolated sound. Importantly, the pattern of results in the current study clearly resembles perceptual learning, and not selective adaptation (which would have led to the opposite pattern of results, i.e., *fewer* /t/ responses after /t/-biased block than after /p/-biased blocks). The observed results showed significantly more /t/-responses after /t/-biased blocks and significantly fewer after /p/-biased blocks. The average proportion of /t/-responses to the individually selected midpoint (during the pre-test) was used to verify whether there were more or less /t/-responses after /t/- and /p/-biased blocks, respectively, relative to the proportion of /t/-responses during the pre-test. As shown in Fig. 3, more /t/-responses after /t/-biased blocks were seen compared to the baseline of the pre-test, and fewer /t/-responses compared to baseline were found after /p/-biased blocks as well. Therefore, it appears unlikely that listeners could have undergone selective adaptation effects, which would have been in the opposite directions compared to baseline as well. Again, Van Linden and Vroomen (2007) have also reported lexical retuning effects with short exposures containing ambiguous sounds.

Overall, the results of the present study suggest that it is possible to compare audiovisual and lexical retuning under similar constraints and that listeners are capable of using both sources of information within a short period of time to adjust phoneme boundaries. While audiovisual cues were, as expected, able to elicit larger recalibration effects, our results indicate that lexical retuning may be flexible in a manner not previously shown, using short exposures to create shifts in two opposing directions, all within a single session. Both lexical and audiovisual perceptual learning were achieved with interleaved exposure blocks and, consequently, we suggest that phoneme boundary retuning can be utilized as a short-term solution for listeners' perceptual difficulties, and can be updated rapidly in accordance with the available contextual cues. The robustness of adaptability in speech perception becomes more apparent with every new investigative technique. In conclusion, the present technique would allow itself to be deployed in the future to explore the neural underpinnings of perceptual retuning, and to investigate potential differences in the multiple percepts induced by lexical and visual/speech-reading information.

Open Practices The data and materials for all experiments are available at https://hdl.handle.net/10411/RWVUTN. None of the experiments were pre-registered.

Acknowledgements We acknowledge financial support from the Netherlands Organization for Scientific Research gravity program Language in Interaction.

References

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science*, *14*(6), 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9/10), 341–347. https://doi.org/10.1097/AUD.0b013e31821473f7

Bruggeman, L. & Cutler, A. (2019). No L1 privilege in talker adaptation. *Bilingualism, Language and Cognition*. https://doi.org/10.1017/S1366728919000646

Cutler, A. (2012). Native listening: the flexibility dimension. *Dutch Journal of Applied Linguistics*, *1*(2), 169–187. https://doi.org/10.1075/dujal.1.2.02cut

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory Phonology, 10*, 91–111. https://doi.org/10.1017/CBO9781107415324.004

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: a tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *36*(3), 488–499. https://doi.org/10.3758/BF03195595

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*. https://doi.org/10.1016/0010-0285(73)90006-6

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

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110–125. https://doi.org/10.1037/0096-1523.6.1.110

Keetels, M., Pecoraro, M., & Vroomen, J. (2015). Recalibration of auditory phonemes by lipread speech is ear-specific. *Cognition*, *141*, 121–126. https://doi.org/10.1016/j.cognition.2015.04.019

Keetels, M., Stekelenburg, J. J., & Vroomen, J. (2016). A spatial gradient in phonetic recalibration by lipread speech. *Journal of Phonetics*, *56*, 124–130. https://doi.org/10.1016/j.wocn.2016.02.005

Kleinschmidt, D. F., & Jaeger, T. F. (2016). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203. https://doi.org/10.1037/a0038695

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

Kraljic, T., & Samuel, A. G. (2009). Perceptual learning for speech. *Attention, Perception & Psychophysics*, *71*(3), 481–489. https://doi.org/10.3758/APP

Lüttke, C. S., Pérez-Bellido, A., & de Lange, F. P. (2018). Rapid recalibration of speech perception after experiencing the McGurk illusion. *Royal Society Open Science*, *5*(3), 170909. https://doi.org/10.1098/rsos.170909

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131-141. https://doi.org/10.3109/03005368709077786

Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word recognition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics (pp.* 19-35). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568971.013.0002

McGurk, H., & MacDonald, M. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 433–443. https://doi.org/10.1037/0096-1523.17.2.433

Mitchel, A. D., Gerfen, C., & Weiss, D. J. (2016). Audiovisual perceptual learning with multiple speakers. *Journal of Phonetics*, *56*, 66–74. https://doi.org/10.1016/j.wocn.2016.02.003

Mitterer, H., & Reinisch, E. (2017). Visual speech influences speech perception immediately but not automatically. *Attention, Perception & Psychophysics*, *79*(2), 660–678. https://doi.org/10.3758/s13414-016-1249-6

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Repp, B. H. (1981). Perceptual equivalence of two kinds of ambiguous speech stimuli. *Bulletin of the Psychonomic Society*, *18*(1), 12-14. https://doi.org/10.3758/BF03333556

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*(4), 348–351. https://doi.org/10.1111/1467-9280.00364

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2), 212-215. https://doi.org/10.1121/1.1907309

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*(6), 1483–1494. https://doi.org/10.1037/0096-1523.33.6.1483

Vroomen, J. & Baart, M. (2009). Recalibration of phonetic categories by lipread speech: measuring aftereffects after a 24-hour delay. *Language and Speech, 52*(2-3), 341-350. https://doi.org/10.1177/0023830909103178

Vroomen, J., & Baart, M. (2012). Phonetic recalibration in audiovisual speech. In M. M. Murray and M. T. Wallace (Eds.) *The Neural Bases of Multisensory Processes.* (pp. 363–379). Boca raton (FL): CRC Press.

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*(3), 572–577. https://doi.org/10.1016/j.neuropsychologia.2006.01.031

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, *44*(1-4), 55–61. https://doi.org/10.1016/j.specom.2004.03.009