Attention, Perception, & Psychophysics (2019) 81:1006–1019
https://doi.org/10.3758/s13414-018-01651-x

PERCEPTUAL/COGNITIVE CONSTRAINTS ON THE STRUCTURE OF SPEECH COMMUNICATION: IN HONOR OF RANDY DIEHL

Attentional resources contribute to the perceptual learning of talker idiosyncrasies in audiovisual speech

Alexandra Jesse[1] · Elina Kaplan[1]

Published online: 25 January 2019

Abstract

To recognize audiovisual speech, listeners evaluate and combine information obtained from the auditory and visual modalities. Listeners also use information from one modality to adjust their phonetic categories to a talker's idiosyncrasy encountered in the other modality. In this study, we examined whether the outcome of this cross-modal recalibration relies on attentional resources. In a standard recalibration experiment in Experiment 1, participants heard an ambiguous sound, disambiguated by the accompanying visual speech as either /p/ or /t/. Participants' primary task was to attend to the audiovisual speech while either monitoring a tone sequence for a target tone or ignoring the tones. Listeners subsequently categorized the steps of an auditory /p/–/t/ continuum more often in line with their exposure. The aftereffect of phonetic recalibration was reduced, but not eliminated, by attentional load during exposure. In Experiment 2, participants saw an ambiguous visual speech gesture that was disambiguated auditorily as either /p/ or /t/. At test, listeners categorized the steps of a visual /p/–/t/ continuum more often in line with the prior exposure. Imposing load in the auditory modality during exposure did not reduce the aftereffect of this type of cross-modal phonetic recalibration. Together, these results suggest that auditory attentional resources are needed for the processing of auditory speech and/or for the shifting of auditory phonetic category boundaries. Listeners thus need to dedicate attentional resources in order to accommodate talker idiosyncrasies in audiovisual speech.

Keywords Speech perception · Perceptual learning · Multisensory processing

In face-to-face communication, listeners use and combine information obtained from hearing and seeing a talker in order to recognize speech (e.g., Massaro, 1998; Summerfield & McGrath, 1984). Having both auditory and visual speech available typically helps recognition (e.g., Jesse, Vrignaud, Cohen, & Massaro, 2000; Reisberg, McLean, & Goldfield, 1987; Sumby & Pollack, 1954), especially in adverse listening situations. Audiovisual speech is more robustly recognized than auditory speech, since the two modalities provide redundant and complementary information about speech sounds (Jesse & Massaro, 2010; Walden, Prosek, & Worthington, 1974). In addition, access to audiovisual speech provides advantages for accommodating the idiosyncratic pronunciations of a talker. Listeners can use information from one modality to disambiguate a phoneme that would be ambiguous on the basis of the information from the other modality alone. This disambiguation can guide the recalibration of phonetic categories in the ambiguous modality (Baart & Vroomen, 2010a; Bertelson, Vroomen, & de Gelder, 2003). For example, a sound between /p/ and /t/ disambiguated by seeing the talker's lips close to produce /p/ shifts the boundary of the auditory phoneme /p/ such that the sound is subsequently included in the /p/ category. Likewise, a gesture ambiguous between /p/ and /t/ that is disambiguated by hearing the talker produce /p/ expands the visual phonetic category of {p} to then include this gesture. Cross-modal phonetic recalibration thus prepares listeners to better recognize future speech from a talker. The question addressed in the present study was whether or not attentional resources are needed for listeners to accommodate to a talker through cross-modal recalibration.

Cross-modal phonetic recalibration was first demonstrated in a seminal study by Bertelson et al. (2003). In the by-now-standard paradigm, participants started by categorizing steps

✉ Alexandra Jesse
ajesse@psych.umass.edu

[1] Department of Psychological and Brain Sciences, University of Massachusetts, 135 Hicks Way, Amherst, MA 01003, USA

in an auditory /aba/–/ada/ continuum so that the perceptually most ambiguous sound A? could be determined for each participant. Participants were then exposed to either the audiovisual syllable /aba/ or /ada/, where the auditory portion of the stop consonant was replaced with their respective ambiguous sound A?. In auditory-only posttests, participants categorized steps of an auditory /aba/–/ada/ continuum more often in line with their prior audiovisual exposure. That is, participants gave more /aba/ responses at auditory posttest when the ambiguous sound A? had been disambiguated during prior exposure by seeing the talker's lips close to produce /b/ (A?Vb) than when the lips had remained open to produce /d/ (A?Vd). This aftereffect indicated the recalibration, or retuning, of these phonetic categories to include the ambiguous sound in the category intended by the talker. These aftereffects of phonetic recalibration are opposite from those of selective adaptation (Diehl, 1975; Eimas & Corbit, 1973), in which participants give *fewer* /aba/ responses after exposure to an unambiguous AbVb than to an unambiguous AdVd. The phenomenon of cross-modal phonetic recalibration has been well replicated over the past decade (Baart & Vroomen, 2010b; Baart, de Boer-Schellekens, & Vroomen, 2012; Keetels, Pecoraro, & Vroomen, 2015; Keetels, Stekelenburg, & Vroomen, 2016; van der Zande, Jesse, & Cutler, 2014; van Linden & Vroomen, 2007, 2008; Vroomen & Baart, 2009a, 2009b; Vroomen, van Linden, de Gelder, & Bertelson, 2007). Recalibration thus helps listeners adjust to ambiguous auditory speech.

Just as auditory phonetic categories can be recalibrated by visual speech information, a recent study has shown that auditory information can likewise recalibrate visual phonetic categories in response to an idiosyncrasy in visual speech (Baart & Vroomen, 2010a). In this version of the standard recalibration paradigm, participants first categorized steps in a visual /omso/–/onso/ continuum in order to determine the participants' individually perceptually most ambiguous visual step. During audiovisual exposure, this most ambiguous visual gesture was accompanied auditorily by either /omso/ (AmV?) or /onso/ (AnV?). In a visual-only phonetic categorization task at posttest, participants showed recalibration of their relevant visual phonetic categories, in which they categorized the steps of a visual continuum more often in line with their prior exposure. These aftereffects indicated a shift in the visual phonetic categories to include the ambiguous speech gesture into the category intended by the speaker. These results fit within a larger literature showing that perceivers are sensitive to visual idiosyncrasies (Heald & Nusbaum, 2014; Yakel, Rosenblum, & Fortier, 2000) and learn about them (Jesse & Bartoli, 2018; van der Zande, Jesse, & Cutler, 2013). Audiovisual speech thus allows listeners to perceive speech more reliably, in that it disambiguates the currently experienced speech but also facilitates future speech perception by recalibrating phonetic categories so that listeners can accommodate to a talker.

Cross-modal phonetic recalibration is an effective mechanism to allow listeners to flexibly cope with variation in pronunciation across talkers. Recalibration of auditory phonetic categories occurs from very little exposure (van Linden & Vroomen, 2007; Vroomen et al., 2007). Furthermore, it dissipates quickly when listeners hear the talker produce a whole spectrum of pronunciations for a phoneme, as is the case with prolonged testing (van Linden & Vroomen, 2007; Vroomen & Baart, 2009b; Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004). Recalibration is also powerful, because its aftereffects can be generalized to new instances. To ensure appropriateness, generalization is guided by bottom-up factors, such as acoustic similarity and spatial location. Aftereffects transfer to novel talkers, as long as these talkers share the same idiosyncrasy (van der Zande et al., 2014). Though the aftereffects were larger for the exposure talker than for a novel talker, seeing who the talker was during exposure did not prevent later transfer to the acoustically similar new talker. Acoustic similarity, and not talker identity, thus drives generalization. Aftereffects are also modulated by talker location. Aftereffects were larger when the talker's test sounds came from the same location as during exposure than when they came from another location (Keetels et al., 2015; Keetels et al., 2016). Listeners may thus store information about the talker's location along with the change in the phonetic category, though why this would be done remains unclear. Together, these results suggest that the bottom-up information obtained from low-level processing guides how specific recalibration is.

In their daily lives listeners, however, often perform multiple tasks during a conversation (e.g., driving or monitoring for a child's cries) that compete for cognitive resources. As a consequence, listeners may have to reduce the cognitive resources dedicated to the processing of speech, which thus could potentially limit their ability to adjust to recent experiences with speech, such as to talker idiosyncrasies. Higher-level processing areas, such as the inferior parietal lobe (IPL) and the inferior frontal sulcus (IFS), are activated (Kilian-Hütten, Vroomen, & Formisano, 2011) during exposure to an auditory idiosyncrasy disambiguated by visual speech. The IFS and IPL have been associated with attention (Corbetta & Shulman, 2002) and working memory (Prabhakaran, Narayanan, Zhao, & Gabrieli, 2000) in other areas of research. The goal of the present study was to test whether one type of cognitive resource—namely, attentional resources—is needed for listeners to achieve effective cross-modal phonetic recalibration. Attention allows to prioritize the further processing of some stimuli over others (see, e.g., Chun, Golomb, & Turk-Browne, 2011, for an overview). Which stimuli are selected for enhanced processing can be driven by bottom-up and/or by top-down processes (e.g., Santangelo & Spence, 2007; Theeuwes, 1991; Yantis & Jonides, 1984). Attention can be thought of as a limited pool of resources (e.g., Alais, Morrone, & Burr, 2006; Arrighi, Lunardi, & Burr, 2011; Sinnett, Costa, & Soto-Faraco, 2018; Wahn & König, 2015; Wahn, Murali, Sinnett, & König, 2017). To the extent that the processing of relevant stimuli does not exhaust this pool, irrelevant stimuli can also be processed (Kahneman & Chajczyk, 1983; Lavie, 1995; Lavie & Tsal, 1994). Although resources are at least partially shared audiovisually in spatial attention, object-based attention taps into modality-specific pools of attentional resources, unless a primary task needs to be prioritized over another task in order to give a speeded response (Wahn & König, 2017).

Sharing attentional resources with other tasks thus could interfere with accommodating to a speaker. Attentional resources could be needed directly or indirectly in order for listeners to effectively adjust to talker idiosyncrasies. That is, recalibration itself could require attentional resources. Alternatively, or additionally, the evaluation and integration of auditory and visual speech, which not only occurs for speech perception but also provides information for recalibration, could rely on these resources. For example, since attention selects stimuli for further processing, reducing auditory attentional resources for speech perception can reduce the sensory encoding of auditory speech (e.g., Mattys, Barden, & Samuel, 2014; Mattys & Palmer, 2015), and as such could also diminish the effectiveness of cross-modal phonetic recalibration. Limiting attentional resources could also impact the outcomes of recalibration by affecting the integration of auditory and visual speech information. Though audiovisual integration has traditionally been regarded as a resource-free process (e.g., Colin et al., 2002; Massaro, 1987; Rosenblum & Saldana, 1996; Soto-Faraco, Navarra, & Alsius, 2004), more recent work has suggested that it may consume attentional resources (Alsius, Möttönen, Sams, Soto-Faraco, & Tiippana, 2014; Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Alsius, Navarra, & Soto-Faraco, 2007). Cross-modal recalibration seems to depend on integrating auditory and visual information into a unitary percept. When presented with sine-wave speech, which only preserves spectrotemporal information and is not spontaneously regarded as speech by naïve listeners, integration (Tuomainen, Andersen, Tiippana, & Sams, 2005) and phonetic recalibration (Vroomen & Baart, 2009a) only occur when listeners are explicitly told that what they hear is speech. These results support the idea that cross-modal recalibration relies on integration, and as such that limiting attentional resources diminishes its effectiveness. In contrast, the aftereffects of selective adaptation are not dependent on whether or not sine-wave speech is regarded as speech. Selective adaptation is, however, a lower-level, modality-specific process, occurring prior to audiovisual integration (Dias, Cook, & Rosenblum, 2016; Roberts & Summerfield, 1981; Saldana & Rosenblum, 1994; Samuel & Lieblich, 2014), that does not require attentional resources (Samuel & Kat, 1998; Sussman, 1993).

Finding that listeners need attentional resources in order for the perceptual processing and/or recalibration itself to accommodate to a speaker through cross-modal recalibration would parallel what has been observed for another type of phonetic recalibration, called *lexically guided retuning*. Lexical knowledge that disambiguates an ambiguous sound or gesture during exposure can also lead to the adjustment of auditory and visual phonetic categories (e.g., Norris, McQueen, & Cutler, 2003; van der Zande et al., 2013). For example, participants who have heard a sound A? ambiguous between /s/ and /f/ replacing the /f/ in words like "giraffe" will afterward categorize the steps on an /s/–/f/ continuum more often as /f/ than will those participants who had previously heard A? replacing the /s/ in words like "platypus." These aftereffects are reduced, however, when listeners are engaged in an auditory distractor task while being exposed to the critical speech information (Samuel, 2016), but are unaffected when participants perform a visual distractor task in which they can move their attention back and forth between the distractor task and the main task during exposure (Zhang & Samuel, 2014). Attentional resources in the modality in which the shift is to be made are therefore needed for the processes that contribute and/or are involved in the lexically guided adjustment of phonetic categories to a speaker.

The aim of the present study was to test whether taxing attentional resources can likewise affect the outcomes of cross-modal phonetic recalibration. If attentional resources are needed for any of the processes involved in cross-modal recalibration (i.e., for perceptual evaluation, integration, and/or for a separate process of recalibration), then limiting the attentional resources available would reduce the size of the observed aftereffects of cross-modal recalibration. To test this hypothesis, two experiments were conducted. In a standard cross-modal recalibration paradigm, the participants in Experiment 1 were exposed to either an audiovisual "apa" (A?Vp) or "ata" (A?Vt) within a block in which the consonant was replaced with an auditorily ambiguous sound A?. The most ambiguous sound A? was selected for each participant on the basis of the participant's results in a prior auditory-only calibration phase. During exposure, participants also always heard tone sequences on each trial. Participants either had to continuously monitor the stream for a lower-frequency target tone, requiring sustained attention, or did not perform a task on the stream. Monitoring for a target tone diverts attention to the auditory modality and away from speech. To assess the aftereffects of recalibration, the participants categorized their three most ambiguous steps from an auditory /apa/–/ata/ continuum at each test block. Recalibration would lead to more /apa/ responses after exposure to A?Vp than after exposure to A?Vt. If attentional resources are needed to effectively accommodate to a speaker, then restricting the availability of attentional resources by asking participants to perform the extrinsic auditory task during exposure should reduce the aftereffects of recalibration.

In Experiment 2, we tested the ubiquity of the need for attentional resources for cross-modal phonetic recalibration, by assessing the recalibration of visual phonetic categories through auditory speech. The paradigm remained the same as in Experiment 1, except that participants were exposed to ApV? or AtV? both before and after categorizing the steps from a visual continuum. The extrinsic task was also the same as before; that is, it continued to be in the auditory modality. The pattern of results across experiments was intended to provide insights into which of the processes directly or indirectly involved in recalibration are likely in need of attentional resources. If limiting attention reduced the aftereffects in both Experiment 1 and Experiment 2, then any of the implicated processes could require attention. In contrast, if limiting attentional resources only affected the aftereffects of recalibrating auditory phonetic categories in Experiment 1, but not the aftereffects of recalibrating visual phonetic categories in Experiment 2, then it would seem most likely, assuming that attentional resources are modality-specific, that attentional resources were needed for auditory processing and/or for the recalibration process itself in Experiment 1. That is, attentional resources might have been too limited to sufficiently process the ambiguous sound and/or to shift an auditory phonetic category boundary.

Together, the two experiments tested whether attention is important in order for listeners to be able to use cross-modal phonetic recalibration to flexibly adjust to recent experiences with speech. If any of the processes contributing to cross-modal phonetic recalibration and/or if the recalibration itself requires attentional resources, then the aftereffects of recalibration should be reduced when attention is diverted to an extrinsic auditory task. Depending on whether limiting attentional resources affects the outcomes of both or of just one type of cross-modal phonetic recalibration, the results would provide some preliminary insights into which of the involved processes likely require attentional resources.

Experiment 1

Method

Participants Twenty-four students at the University of Massachusetts Amherst (mean age = 19.58 years; 18 women, six men) contributed data to the analyses. This sample size was decided upon a priori as being appropriate (Baart & Vroomen, 2010a; van der Zande et al., 2014). Three additional participants had been excluded because they did not meet the inclusion criteria set a priori, due to their low performance in the target detection task. All of the participants were monolingual native speakers of American English with no reported hearing, vision, language, or attention deficits. An additional sample of ten participants from the same population completed a pilot experiment.

Materials A female native speaker of American English was video-recorded saying /ɑpɑ/, /ɑtɑ/, and /ɑkɑ/. Videos of the speaker's face were recorded with a SONY EVI-HD7V camera at 25 fps (1,280 × 720) using the h.264 codec. Audio was simultaneously recorded in mono with a Shure KSM44A microphone at a 48-kHz sampling rate.

An *apa* and an *ata* token with similar durations of their individual phonemes were selected to create an auditory continuum. Using Praat (Boersma & Weenink, 2016), the plosives were excised and systematically mixed in linearly increasing complementary proportions to create a 102-step continuum. An additional /ɑkɑ/ token recorded in the same session provided the vowel context for the auditory continuum. The steady-state portion of the first vowel of this /ɑkɑ/ token was cut and its duration was changed with Praat's PSOLA algorithm in order to be the same as in the first vowel in /ɑpɑ/ and /ɑtɑ/ (i.e., 253 ms). The intensity of both vowels of /ɑkɑ/ was scaled to the mean intensity of the vowels in the same position in /ɑpɑ/ and /ɑtɑ/. Linear ramps were applied to the first 75 ms of the first vowel and to the last 75 ms of the second vowel. These vowels were then concatenated with each continuum step.

Forty-two steps of the 102-step continuum were selected to be tested in a pilot experiment. After one practice block, participants received four test blocks. Each test block consisted of three repetitions of all steps in a newly randomized order. Figure 1 shows the results of the pilot experiment with ten participants, who each categorized 23 steps of the auditory continuum (steps 0 [/t/], 47, from 51 to 90 in steps of two or three, 94, and 100 [/p/]) by button press as "apa" or "ata." On the basis of the results, we interpolated the steps in the ambiguous region in order to create steps 67, 69, and 71, and then chose steps 58, 62, 64, 66, 67, 69, 70, 71, 72, 74, 76, 79, 80, and 83, plus the two endpoints, for the calibration phase of Experiment 1. The audiovisual exposure stimuli were created by combining each of these steps with the endpoint video tokens. A fade in from and fade out to a black frame were added to the videos by repeating the first and last frames of each video, respectively, six times.

For the target detection task, nonlinguistic materials were chosen in order to avoid interference due to linguistic processing. Pure tones of 700, 750, 800, and 850 Hz served as the nontarget sounds, and a pure tone of 500 Hz was the target sound. Tones at low frequencies were chosen in order to avoid energetic masking of the critical acoustic place-of-articulation information, located in the higher frequencies of /p/ and /t/. All tones were 160 ms long, including a 10-ms-long linear fade in and fade out. The amplitude of each tone was 86 dB. In comparison, the nonsense syllables /ɑpɑ/ and /ɑtɑ/ had a mean amplitude of 76 dB.

Procedure Participants were tested individually in sound-attenuated booths. The experiment was controlled by the

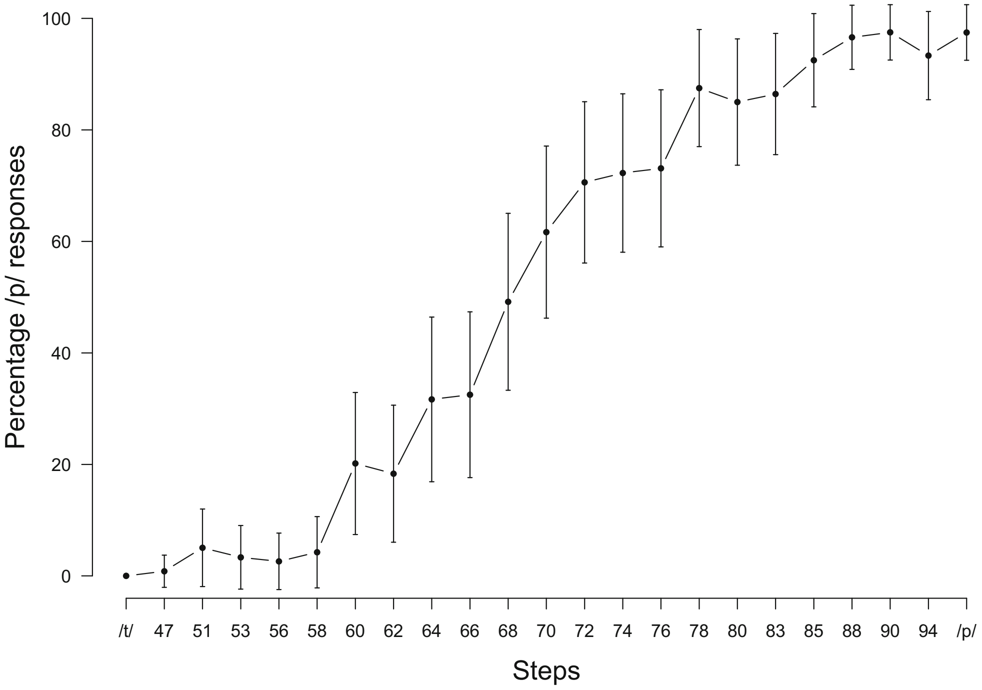

Fig. 1 Mean percentages of /p/ responses as a function of step on the auditory continuum in a pilot study for Experiment 1. Error bars show the standard errors of the means

Psychophysics Toolbox software (Brainard, 1997). Audio was presented at a comfortable, fixed level over Sennheiser HD 280 PRO headphones. The visual stimuli were presented on a 1,024 × 768 (17-in. diagonal) computer screen (60-Hz refresh rate), positioned 60 cm in front of the participant.

In the initial, auditory-only calibration phase, participants categorized 17 steps of the auditory /ɑpɑ/–/ɑtɑ/ continuum. Each trial began with a fixation cross shown for 250 ms, followed by a black screen for 200 ms before a stimulus was presented. At stimulus offset, the labels (“apa,” “ata”) appeared, starting a 3-s-long deadline for participants to respond by button press. The intertrial interval was 500 ms. Participants were instructed to respond as quickly and accurately as possible. Each participant completed one practice block and two test blocks. Each test block consisted of four repetitions of the continuum steps. Steps were presented in a newly randomized order for each repetition. On the basis of a participant’s results in this calibration phase, the step closest to the 50% cutoff point was selected as their exposure stimulus A?.

Next, participants completed 32 exposure–test sequences, in which each audiovisual exposure block was immediately followed by an auditory-only posttest (see Fig. 2 for the experimental design). During each audiovisual exposure block, participants received the ambiguous stimulus A? accompanied by the unambiguous visual token “apa” (A?Vp) or “ata” (A?Vt). The same audiovisual stimulus was presented nine times within a block. On each exposure trial, participants also always heard a seven-tone sequence. Each sequence started at the beginning of the video. On filler trials, the sequence consisted of the four nontarget tones. Three tones each occurred twice during the stream, whereas one tone occurred only once. Which of the four nontarget tones occurred only once in a sequence was counterbalanced across trials within each exposure block. Instead of that one tone, a target

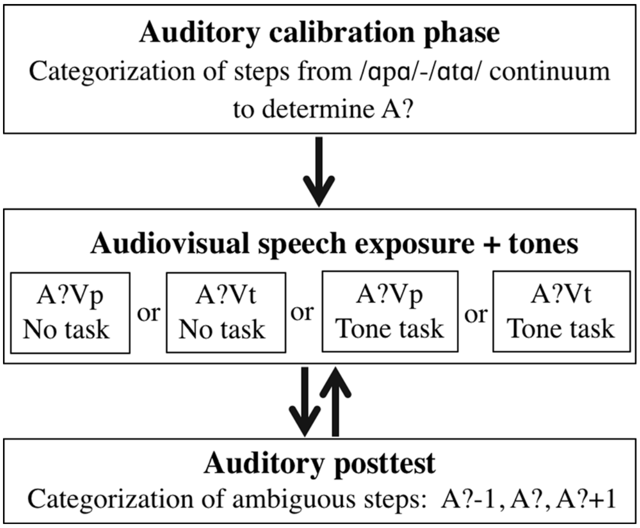
Auditory calibration phase
Categorization of steps from /ɑpɑ/-/ɑtɑ/ continuum to determine A?
Audiovisual speech exposure + tones
A?Vp No task
or
A?Vt No task
or
A?Vp Tone task
or
A?Vt Tone task
Auditory posttest
Categorization of ambiguous steps: A?-1, A?, A?+1

Fig. 2 Illustration of the experimental procedure and design of Experiment 1. During the calibration and posttest, participants categorized steps from an auditory /ɑpɑ/–/ɑtɑ/ continuum. During audiovisual exposure, participants received an ambiguous stimulus A? accompanied by the unambiguous visual token “p” (A?Vp) or “t” (A?Vt). Participants also heard tone sequences that they monitored for a lower-frequency target tone in the load condition, but not in the control condition. Participants completed a total of 32 exposure–test sequences

tone was presented in one of the seven positions on 33% of the trials (i.e., on three trials per exposure block). The order of the tones' presentation within a trial was always pseudorandom, with the constraint that no tone was immediately repeated. On target trials, the target was equally likely to occur early (i.e., as the second or third tone), in the middle (fourth or fifth tone), or late (sixth or seventh tone) during the tone sequence within each exposure block.

During exposure, the main task for participants was to attend closely to what the talker said. Exposure was blocked by load. For half of the exposure blocks, participants had to continuously monitor the tone sequence, because they were instructed to press a button as soon as they heard the lower-frequency target tone in the sequence (*load condition*). That is, completing this task required sustained attention to the tone stream and did not allow for participants to switch their attention back and forth between the tones and speech. Working memory resources were taxed only minimally, in that a target is likely to be stored in long-term memory if it remains the same throughout the experiment (Woodman, Luck, & Schall, 2007). The target detection task was demonstrated and practiced on its own for five trials at the beginning of the experiment. For the other half of the exposure blocks, participants still heard the tone sequences during exposure but they did not perform a task on them (*control condition*). The condition order (control or load first) was counterbalanced across participants. Each half of the exposure blocks consisted of eight blocks of exposure to "apa" (A?Vp) and eight blocks of exposure to "ata" (A?Vt). The order of these blocks alternated within each half, but each type of exposure block occurred equally often as the first one across participants. This resulted in four lists (control first/load first × A?Vp /A?Vt first).

Each audiovisual exposure block was immediately followed by an auditory-only posttest phase. This test was similar to the one completed during the calibration phase, except that participants categorized only their three most ambiguous steps (A?–1, A?, A?+1) by a button press as either "apa" or "ata." Each step was presented twice per block in random order.

Results and discussion

For all statistical analyses, generalized mixed-effect models with a binomial linking function were implemented in R (R Core Team, 2014) using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). Random effects included by-participant intercepts and by-participant random slopes for repeated measures (Barr, Levy, Scheepers, & Tily, 2013). The *p* values were estimated using Satterthwaite approximation for the degrees of freedom. No overdispersion was detected. The model fits were appropriate, as checked by inspecting the distribution of the binned residuals.

Tone detection performance To ensure that the participants had been engaged in the target detection task during exposure, a priori inclusion criteria were set for the performance in that task. Participants had to have a false alarm rate of 50% or less in the load condition, and of 10% or less in the control condition, to be included in any analyses. In addition, their hit rate needed to exceed 20% and their *d'* score needed to be above 0.3. Hits were defined as responses given after target onset on target trials. With these criteria applied, the false alarm rates of the final sample of 24 participants were low in both conditions (load condition, $M = 9.31\%$, $SD = 7.53\%$; control condition, $M = 0.46\%$, $SD = 1.52\%$), and in the load condition, participants had a high hit rate ($M = 91.32\%$, $SD = 6.97\%$) and a high *d'* score ($M = 3.4$, $SD = 0.75$). A generalized mixed-effect model analyzing hit rates as a function of position (coded as a centered numerical fixed factor) showed that target detection did not vary across position;, that is, it did not vary as a function of when during the audiovisual syllable the target occurred ($\beta = -0.073$, $SE = 0.09$, $p = .39$). That is, participants did not switch their full attention from the tone sequence to the syllables at those moments when critical speech information was being provided. Rather, participants sustained their attention to the tone stream.

Phonetic recalibration Figure 3 shows the categorization of the auditory continuum at posttest as a function of step, exposure condition, and load, and their possible interactions. A generalized mixed-effect model was fit, with exposure (/p/ = –0.5, /t/ = 0.5) and load (control = –0.5, load = 0.5) as contrast-

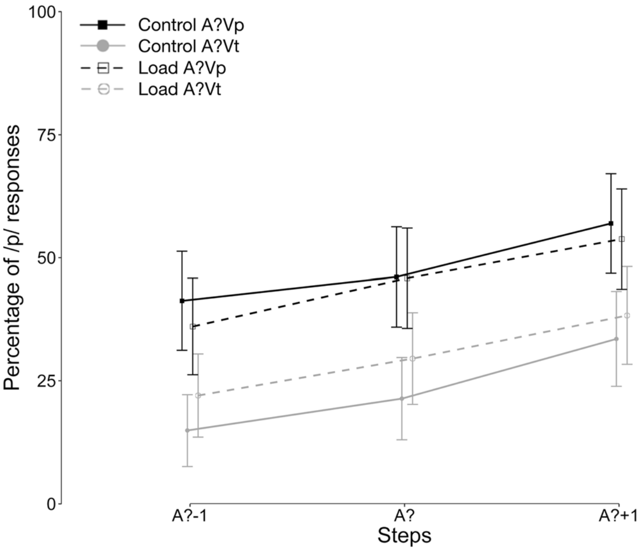

Fig. 3 Mean percentages of /p/ responses as a function of step, exposure, and load in Experiment 1. Black squares show test data for the /p/ exposure condition, and gray dots those for the /t/ exposure condition. Solid lines show the test data for the control condition, and dashed lines those for the load condition. The aftereffect of recalibration is the difference in /p/ responses after A?Vp exposure as compared to after A?Vt exposure. Standard errors of the means are shown

coded fixed factors and continuum step as a centered numerical factor. By-subjects random slope adjustments were included for load, exposure, step, and their interactions. As expected, participants gave more /p/ responses for more /p/-like steps (Step: $\beta = 0.5$, $SE = 0.01$, $p < .00001$), but this sensitivity to the continuum was not affected by load (Step × Load: $\beta = –0.08$, $SE = 0.1$, $p = .43$). Furthermore, listeners recalibrated their auditory phonetic categories in line with exposure, indicated in that more /p/ responses were given after exposure to A?Vp than after exposure to A?Vt (Exposure: $\beta = – 1.22$, $SE = 0.24$, $p < .00001$). The extent of phonetic recalibration did not vary as a function of step (Exposure × Step: $\beta = 0.18$, $SE = 0.11$, $p = .1$), and load had no effect on categorization overall (Load: $\beta = 0.23$, $SE = 0.13$, $p = .09$). However, as we predicted, the availability of attentional resources affected recalibration: The effect of phonetic recalibration was larger in the control than in the load condition (Exposure × Load: $\beta = 0.76$, $SE = 0.22$, $p < .001$). Post-hoc tests showed that recalibration occurred in both the control condition (Exposure: $\beta = – 1.59$, $SE = 0.31$, $p < .00001$) and the load condition (Exposure: $\beta = – 0.82$, $SE = 0.17$, $p < .00001$). The triple interaction between step, exposure, and load was not significant ($\beta = – 0.32$, $SE = 0.22$, $p = .15$).

Overall, these results provide evidence that the outcomes of recalibration rely on listeners' availability of attentional resources during exposure. As compared to the control condition, the aftereffects in the load condition indicate that recalibration was less complete when participants' attentional resources were taken up by the extrinsic tone detection task during exposure. In our study, performing the tone detection task did not, however, deplete the resources available for the processes contributing to recalibration, and as such did not prevent recalibration.

Experiment 2

Experiment 1 provided evidence that the successful recalibration of auditory phonetic categories through visual speech information requires that listeners dedicate attentional resources to the contributing processes. Recalibration was reduced, though not eliminated, when participants monitored a tone sequence for a lower-frequency target tone during exposure than when they just heard the tones. Cross-modal phonetic recalibration works in two directions, however: As one prior study has shown, when seeing a speech gesture that is ambiguous, listeners use auditory information to disambiguate the intended speech sound and recalibrate their visual phonetic categories to the talker (Baart & Vroomen, 2010a).

In Experiment 2, we replicated this auditorily guided recalibration of visual phonetic categories. Participants were exposed to a visually ambiguous gesture V? accompanied by a clear auditory sound (ApV?, AtV?), before being asked to categorize a visual /ɑpɑ/–/ɑtɑ/ continuum. In the case of recalibration, we expected more categorizations at test to be in line with the prior exposure. That is, more /p/ responses should be given after exposure to ApV? than after exposure to AtV?. Furthermore, we tested whether attentional resources also contribute to the outcomes of this type of cross-modal phonetic retuning. For this purpose, participants again either only listened to tones in a sequence during exposure or monitored the sequence for a target tone. If the outcomes of both types of recalibration were reduced under load, then any of the involved processes—that is, auditory/visual processing, integration, and/or recalibration—could require attentional resources in order for listeners to efficiently adjust to a talker's idiosyncrasy. In contrast, if taxing auditory attentional resources were to affect only the aftereffects of recalibrating auditory phonetic categories in Experiment 1 but not the aftereffects of recalibrating visual phonetic categories in Experiment 2, then attentional resources might be needed for auditory processing and/or for the shift of the auditory phonetic category boundary. Because attentional resources were taxed in the auditory domain, limiting these resources could affect auditory processing, and as such the outcomes of recalibration, only in Experiment 1, in which auditory speech perception was challenging because listeners had to recognize an auditorily ambiguous sound. If that were the case (and if attentional resources are not shared across modalities), auditory attentional load should not affect the auditory processing, and hence reduce recalibration, in Experiment 2, in which the auditory signal provided clear information about the speech sound's identity. Similarly, if attentional load were to reduce the outcomes of recalibration only in Experiment 1 but not in Experiment 2, then taxing auditory attentional resources could have affected shifting of the boundaries of auditory phonetic categories, but not those of visual phonetic categories. A need for attentional resources for integration, however, would be unlikely to explain this pattern of results, since the integration process should be the same across experiments.

Method

Participants Twenty-four new participants completed Experiment 2 (mean age = 20.38 years; 19 women, five men), and an additional sample of nine participants completed a pilot experiment. All of these participants were from the same population as the participants in Experiment 1. Six additional participants were excluded because they did not meet the inclusion criteria set a priori, due to their low performance in the target detection task.

Materials The materials were the same as in Experiment 1, except that a visual speech continuum was created on the basis of the same selected /ɑpɑ/ and /ɑtɑ/ tokens. To create this visual continuum (Baart & Vroomen, 2010a; van der Zande et al., 2013) in Adobe Premiere CS5 (Adobe Systems, Mountain View, CA), the video tracks of these token were overlaid and the opacity level of the /ɑpɑ/ video was systematically reduced in 5% increments from 100% to 0% in order to create 21 continuum steps. Between 30% and 60%, six additional steps at 2.5% increments were created. In total, the continuum therefore had 27 steps. The step numbers express the relative opacity of the /ɑpɑ/ video. The first and last frames of each video were duplicated six times to create a visual fade in from black and a fade out to black. The same amount of silence was added to the auditory tracks of the endpoint tokens.

In a pilot experiment, nine participants categorized 17 selected steps of the visual continuum (steps 0 [/t/], 15, 30, 33, 35, 38, 40, 43, 45, 48, 50, 53, 55, 58, 60, 85, and 100 [/p/]) by button press as “apa” or “ata.” The continuum steps were presented eight times in a newly randomized order. The procedure was the same as in the pilot experiment conducted for Experiment 1. The results in Fig. 4 show that participants were sensitive to the visual continuum and gave more /p/ responses to more /p/-like steps. The same continuum was therefore used for the calibration phase of Experiment 2. Audiovisual exposure stimuli were created by combining the steps of the visual continuum with the original audio tracks of the endpoint tokens.

Procedure The design and procedure were similar to those of Experiment 1. The only difference was that during the calibration phase, participants categorized steps from the visual continuum. During audiovisual exposure, each participant’s visual step closest to the 50% cutoff was presented accompanied by the auditory endpoint “apa” (ApV?) or “ata” (AtV?). At each visual-only posttest, participants received the ambiguous step V? and its two adjacent steps V?–1 and V?+1 for categorization, twice in a random order. Attention was taxed in the same way as in Experiment 1. Listeners either performed or did not perform a target detection task on tone sequences during this exposure. The same sequences were presented as in Experiment 1. The experimental design can be seen in Fig. 5.

Results and discussion

Tone detection performance The same inclusion criteria as in Experiment 1 were applied to the participants with regard to their performance in the target detection task. The 24 included participants had a mean false alarm rate of 11.2% (*SD* = 11.51%) in the load condition and of 0.72% (*SD* = 1.62%) in the control condition. Their hit rate in the load condition was, on average, 85.76% (*SD* = 11.49%), and their average *d′* score was 3.14 (*SD* = 0.88). The probability of a correct target tone detection did not vary across positions within the sequence ($\beta = -0.07$, *SE* = 0.09, *p* = .39), suggesting, again, that participants did not switch their full attention to the syllable at critical moments.

Phonetic recalibration Figure 6 shows the mean percentages of /p/ categorization responses to the visual continuum at posttest as a function of step, exposure, and load. The results from a generalized mixed-effect model produced evidence of recalibration: At posttest, participants categorized more visual

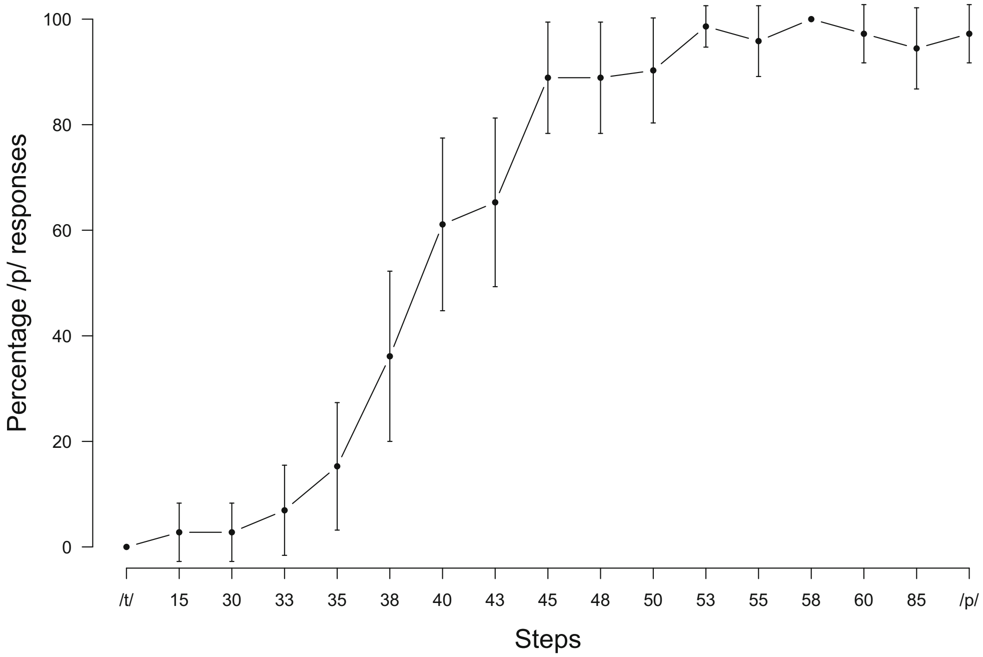

Fig. 4 Mean percentages of /p/ responses as a function of step on the visual continuum in the pilot study for Experiment 2. Step numbers express the relative opacity of the /apa/ video. Error bars show the standard errors of the means

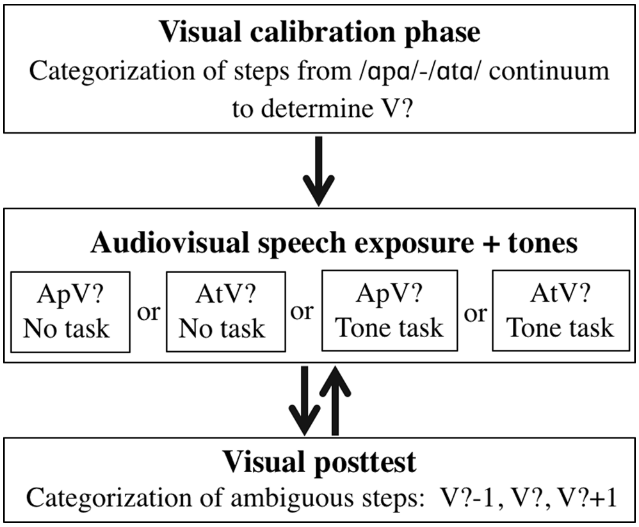
Visual calibration phase
Categorization of steps from /ɑpɑ/-/ɑtɑ/ continuum to determine V?
Audiovisual speech exposure + tones
ApV? No task
or
AtV? No task
or
ApV? Tone task
or
AtV? Tone task
Visual posttest
Categorization of ambiguous steps: V?-1, V?, V?+1

Fig. 5 Illustration of the experimental procedure and design of Experiment 2. During calibration and posttest, participants categorized steps from a visual /ɑpɑ/–/ɑtɑ/ continuum. During audiovisual exposure, participants received an ambiguous visual stimulus V? accompanied by the unambiguous auditory token "p" (ApV?) or "t" (AtV?). Participants also heard tone sequences, which they monitored for a lower-frequency target tone in the load condition but not in the control condition. Participants completed a total of 32 exposure–test sequences

steps as /p/ after exposure to ApV? than after exposure to AtV? (Exposure: $\beta = -0.86$, $SE = 0.21$, $p < .0001$). Unlike in the visually guided recalibration in Experiment 1, directing attention to an auditory task did not affect the outcomes of auditory-guided recalibration. The size of the recalibration aftereffect did not differ across attention conditions (Exposure × Load: $\beta = 0.25$, $SE = 0.22$, $p = .27$). Load also had no overall effect on categorization (Load: $\beta = 0.09$, $SE = 0.12$, $p = .46$), but participants' sensitivity to the visual continuum (Step: $\beta = 1.14$, $SE = 0.11$, $p < .00001$) became more categorical in the load than in the control condition (Step × Load: $\beta = 0.23$, $SE = 0.11$, $p < .05$). Recalibration did not vary as a function of step (Exposure × Step: $\beta = -0.03$, $SE = 0.11$, $p = .80$), and the triple interaction between step, exposure, and load was not significant (Exposure × Load × Step: $\beta = 0.10$, $SE = 0.23$, $p = .65$). Listeners thus used auditory speech information to recalibrate their visual phonetic categories, but directing attentional resources away to another auditory task did not diminish the aftereffects.

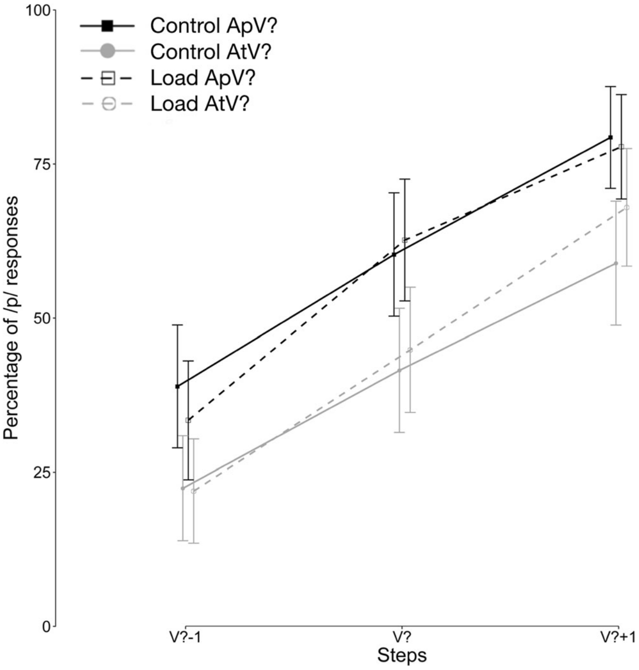

Fig. 6 Mean percentages of /p/ responses as a function of step, exposure, and load in Experiment 2. Black squares show test data for the /p/ exposure condition, and gray dots those for the /t/ exposure condition. Solid lines show the test data for the control condition, and dashed lines those for the load condition. The aftereffect of recalibration is the difference in /p/ responses after ApV? exposure than for after AtV? exposure. Standard errors of the means are shown

General discussion

The goal of this study was to test whether taxing attentional resources affects the outcomes of cross-modal phonetic recalibration. In Experiment 1, during the exposure phase in a standard cross-modal recalibration paradigm, the identity of a phoneme was auditorily ambiguous but was disambiguated by visual speech information. At test, participants categorized the steps of an auditory continuum more often in line with prior exposure, providing evidence for the recalibration of auditory phonetic categories. In Experiment 2, the visual speech information was ambiguous but was disambiguated by auditory information. Here, at test, participants categorized the steps of a visual continuum more often in line with exposure, adding to the scarce evidence that listeners also recalibrate visual phonetic categories when encountering talker idiosyncrasies in visual speech (Baart & Vroomen, 2010a; van der Zande et al., 2013). Critical to the main goal of the study, participants also always heard a tone sequence during exposure in both experiments. Participants had to either ignore this tone sequence or continuously monitor it for the occasional occurrence of a lower-frequency target tone. Performing a task on the auditory tone sequence reduced the outcome of cross-modal recalibration only in the case in which visual speech information guided the shift of auditory phonetic category boundaries. However, the effectiveness of the recalibration of visual phonetic categories was unaffected by the attention manipulation. Together, these results suggest that auditory processing and/or recalibration itself requires sufficient attentional resources for listeners to adjust to a speaker. However, although the distractor task affected the outcomes of recalibration, it did not prevent recalibration. Within the framework of perceptual-load theory (Lavie, 1995), the distractor task may not have been perceptually demanding enough to use up the available attentional resources in order to fully prevent recalibration. The unattended speech stimulus was therefore still processed to some extent, and some recalibration occurred.

The primary goal of this study was to demonstrate the importance of attentional resources for the outcomes of cross-modal phonetic recalibration. Limiting the attentional resources available in the auditory modality resulted in less effective recalibration of auditory phonetic categories, because recalibration itself and/or the processes involved in audiovisual speech perception may require such resources. One possibility is that the availability of attentional resources modulates recalibration, because attentional resources may be needed for the evaluation of auditory and visual speech information. When processing audiovisual speech, both auditory and visual speech information are evaluated and integrated into a unitary percept. Attention selects stimuli for further processing. Imposing load—that is, reducing the resources available for speech perception—can interfere with the sensory encoding of auditory speech. As a result of attentional load, listeners are less able to process acoustic detail (e.g., Mattys et al., 2014; Mattys & Palmer, 2015), and phonemes thus become less discriminable (Mattys & Wiget, 2011). Lesser matches of incoming information with mental representations suffice, then, for recognition. In the present experiments, diverting attentional resources to another auditory task could have reduced detailed processing of the auditory information. In Experiment 1, under load, participants may have had less information about the actual (auditory) idiosyncrasy, rendering recalibration more difficult. In Experiment 2, load may have had no effect on auditory processing because the processing of the clear auditory input was less demanding. Reducing the availability of attentional resources for auditory processing thus may have only affected listeners when the auditory input was more difficult to process, as in the case when the sound was rendered ambiguous by a talker idiosyncrasy. As such, the recalibration only of auditory phonetic categories, and not of visual phonetic categories, was affected.

Given our results, it seems unlikely that the load imposed by the auditory distractor task could have impacted visual processing, since it is unclear, then, why load would have affected the outcomes of recalibrating only auditory, but not visual, phonetic categories. The few prior studies examining whether auditory perceptual load can affect visual processing have shown mixed results (Berman & Colby, 2002; Houghton, Macken, & Jones, 2003; Murphy & Greene, 2017; Rees, Frith, & Lavie, 2001). Nonetheless, overall the literature suggests that while resources are at least partially shared audiovisually in spatial attention, object-based attention (which we manipulated in our study) taps into modality-specific pools of attentional resources unless a primary task needs to be prioritized over a secondary task to give a speeded response (Wahn & König, 2017). In our study, the processing of visual speech itself thus may have been unaffected by load. More research will be needed, however, on the exact circumstances under which resources are, or are not, shared cross-modally in speech perception.

It also seems unlikely that loading attentional resources could have affected the integration process that cross-modal phonetic recalibration seems to depend upon (Vroomen & Baart, 2009a). The integration of audiovisual speech has traditionally been regarded as automatic, and thereby as preattentive (e.g., Colin et al., 2002; Massaro, 1987; Rosenblum & Saldana, 1996; Soto-Faraco et al., 2004). More recent work, however, has suggested otherwise (Alsius et al., 2005; Alsius et al., 2007). In these studies, participants received McGurk stimuli in which the simultaneous presentation of a visual /k/ and an auditory /p/ resulted in reports of /t/. The proportion of these reported fusion responses was reduced when resources were taken up by a secondary task, in which participants had to detect an immediate repetition in a stream of either pictures or environmental sounds. However, traditionally, this 1-back task has been interpreted as relying largely on working memory resources (e.g., Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Wilhelm, Hildebrandt, & Oberauer, 2013). Limiting working memory may thus have affected speech perception in these studies.

Clearer evidence for a potential use of *attentional* resources in audiovisual integration has come from a replication with a tactile task (Alsius et al., 2007), in which McGurk fusion responses were less likely when participants had to detect a certain target stimulus in a stream of tactile events (0-back task). Importantly, the tactile task had no influence on recognizing unimodally presented speech, suggesting that attentional resources affected the modality-general levels of processing in audiovisual speech perception. These results were, however, only found when load was manipulated as a within-subjects variable, and not replicated in a second study with a between-subjects manipulation. Furthermore, although McGurk fusion responses were interpreted in these studies as a measure of integration, in reality they can at best show only a visual influence on auditory processing (Tiippana, 2014). Visual influence can, however, also occur on trials in which fusion responses are not reported (Brancazio & Miller, 2005). McGurk responses, therefore, do not fully capture visual influences. Load may also affect visual influences on early auditory processing. In an electrophysiological study (Alsius et al., 2014), monitoring a rapid stream of pictures for a target reduced change in the latency of the auditory-evoked N1 commonly observed in audiovisual speech when compared to auditory speech. Load may hence modulate cross-modal influences. However, visual load could have instead impacted the processing of visual speech, thus decreasing the extent to which visual information interacted with the processing of auditory speech. The prior evidence that integration in audiovisual speech perception requires attentional resources is therefore still inconclusive. In line with the current state of the literature, it seems unlikely that limiting attentional resources could have affected listeners' ability to integrate information extracted from the two modalities into a cohesive unitary percept, thereby reducing recalibration. The same integration process takes place whether or not the auditory and/or visual information is ambiguous. It is therefore unclear why limiting attentional resources would only affect the integration process involved in the recalibration of auditory phonetic categories, and not that involved in the recalibration of visual phonetic categories.

In summary, load could have affected recalibration indirectly by impacting the processing of audiovisual speech. It seems unlikely, however, that taxing auditory attentional resources affected visual processing or the efficiency with which listeners combined information to recognize what was said. Rather, it remains possible that taxing auditory attentional resources limited how much information could be extracted from the auditory speech signal, thereby limiting recalibration. This restriction, however, only had an effect on recalibration when the auditory processing was already more difficult because the stimulus was ambiguous, and as such when listeners had to shift the boundaries of auditory phonetic categories. Alternatively, or additionally, recalibration itself could require attentional resources. Limiting attentional resources in the auditory modality could have affected listeners' ability to shift the boundaries of auditory phonetic categories. Both explanations predict that if load were manipulated through an extrinsic task in the visual domain, the outcomes of recalibrating visual phonetic categories, but not of recalibrating auditory phonetic categories, would be affected. Although cross-modal recalibration is a well-replicated phenomenon, its underlying mechanisms are not well understood, making it difficult to outline exactly how recalibration may rely on attention.

Attentional resources thus play a role in listeners' ability to recalibrate phonetic categories to talker idiosyncrasies. In contrast, working memory resources seem not to be involved (Baart & Vroomen, 2010b). In a recent study by Baart and Vroomen (2010b), working memory resources were taxed by asking participants to hold a letter or the spatial location of a dot in memory during a standard recalibration paradigm. Performing these tasks did not interfere with cross-modal phonetic recalibration. However, as the authors cautioned, working memory resources may not necessarily have been taxed during the critical audiovisual exposure phase. Working memory resources thus could also play a role if memory is continuously taxed, as attentional resources were taxed here.

The finding that the outcomes of cross-modal phonetic recalibration rely on the availability of attentional resources dovetails with those of recent work showing that attentional resources may also contribute to lexically guided phonetic retuning and other types of perceptual learning. The extent to which auditory phonetic categories are retuned through lexical information is reduced when listeners are engaged in an auditory distractor task during exposure (Samuel, 2016). Limiting auditory attention thus also affects at least some of the processes involved in lexically guided retuning. However, when participants can share their attention between an extrinsic visual task and speech processing during exposure, recalibration is unaffected (Zhang & Samuel, 2014). The attentional skill set of the listener is important for recalibration, even in situations in which no extrinsic task is performed: Older listeners' general ability to switch attention between stimuli, but not their ability to selectively attend to some information while ignoring other information, predicted the size of their aftereffects (Scharenborg, Weber, & Janse, 2014). More precisely, those older adults who were generally better at switching attention showed more lexically guided phonetic retuning. Taking these findings together with our results, attention contributes to phonetic recalibration, no matter whether it is guided by lexical or perceptual information.

Attention generally seems to play a role in perceptual learning, since its contribution can also be observed for other types of perceptual learning (e.g., Adank & Janse, 2010; Huyck & Johnsrude, 2012; Janse & Adank, 2012). For example, prior exposure to noise-vocoded speech only benefited listeners for recognition at test if they previously had attended to the noise-vocoded speech during exposure (Huyck & Johnsrude, 2012). If listeners' attention during exposure was diverted to a competing auditory or visual task, their performance was similar to that when no prior exposure was given. Attention to the critical speech stimuli, however, is not always necessary for learning to occur (e.g., Seitz et al., 2010; Wright, Sabin, Zhang, Marrone, & Fitzgerald, 2010). Together, these results suggest that attention is important for allowing listeners to flexibly adjust to recent experiences with speech.

In addition to demonstrating that effective outcomes from cross-modal recalibration require attentional resources, our results provide further support that listeners also recalibrate their visual phonetic categories (Baart & Vroomen, 2010a; van der Zande et al., 2013). Talker variability affects speech in both modalities. Listeners are sensitive to the variability in production across talkers and to the consistency within talkers in both modalities (e.g., Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007; Wong, Nusbaum, & Small, 2004; Yakel et al., 2000). Listeners can use auditory information to disambiguate visual speech in cases in which an idiosyncrasy rendered the visual speech information ambiguous. Listeners can furthermore also use lexical information in a similar vein (van der Zande et al., 2013). The use of lexical information becomes particularly relevant when both the auditory and visual speech are ambiguous, due to an idiosyncratic pronunciation. Lexical information, however, recalibrates visual phonetic categories directly, and not indirectly via the recalibration of auditory categories (van der Zande et al., 2013). Listeners also use talkers' idiosyncratic realizations of visual speech to form representations of these talkers' identities (Jesse & Bartoli, 2018). These representations allow listeners to recognize talkers even from new utterances. Overall, listeners are sensitive to a talker's idiosyncratic way of speaking in both modalities and adjust their representations in order to better recognize the talker's speech, in addition to building identity representations so as to recognize the talkers themselves. Recalibrating visual phonetic categories helps listeners keep their audiovisual speech perception system optimized to their conversational partner. Audiovisual speech provides an important interface for more reliable and efficient recognition of speech. By flexibly accommodating to talkers, listeners can ensure that visual speech information can optimally aid recognition.

Conclusions

During our everyday conversations, people often perform multiple tasks at the same time, and thus have to distribute attentional resources across these tasks. Our results show that focusing on another auditory task reduces listeners' ability to process critical auditory information and/or to recalibrate auditory phonetic categories. Listeners' ability to accommodate a talker idiosyncrasy in the auditory modality was hence negatively impacted. The outcomes of cross-modal recalibration thus rely on the availability of sufficient attentional resources in the modality for processing and/or recalibration in that same modality. Cross-modal recalibration is an efficient, powerful mechanism that optimizes listeners' speech perception system to new experiences, if sufficient attentional resources are available to the listener.

Author note The authors thank David Kajander for his help with this work, and Ashley Boulanger, Jessica Bowe, Emmaline Brown, Joanna Kalucki, Bridget North, Alexander Roll, Anna Paterson, Megan Stoutenburgh, and Gabrielle St. Pierre for their assistance with the data collection. Parts of this work were presented at the 57th Annual Meeting of the Psychonomic Society, Boston, Massachusetts (Kajander, Kaplan, & Jesse, 2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychology and Aging*, *25*, 736–740. https://doi.org/10.1037/a0020054

Alais, D., Morrone, C., & Burr, D. (2006). Separate attentional resources for vision and audition. *Proceedings of the Royal Society B*, *273*, 1339–1345. https://doi.org/10.1098/rspb.2005.3420

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: Evidence from ERPs. *Frontiers in Psychology*, *5*, 727. https://doi.org/10.3389/fpsyg.2014.00727

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, *15*, 839–843. https://doi.org/10.1016/j.cub.2005.03.046

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research*, *183*, 399–404. https://doi.org/10.1007/s00221-007-1110-1

Arrighi, R., Lunardi, R., & Burr, D. (2011). Vision and audition do not share attentional resources in sustained tasks. *Frontiers in Psychology*, *2*, 56. https://doi.org/10.3389/fpsyg.2011.00056

Baart, M., de Boer-Schellekens, L., & Vroomen, J. (2012). Lipread-induced phonetic recalibration in dyslexia. *Acta Psychologica*, *140*, 91–95. https://doi.org/10.1016/j.actpsy.2012.03.003

Baart, M., & Vroomen, J. (2010a). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters*, *471*, 100–103. https://doi.org/10.1016/j.neulet.2010.01.019

Baart, M., & Vroomen, J. (2010b). Phonetic recalibration does not depend on working memory. *Experimental Brain Research*, *203*, 575–582. https://doi.org/10.1007/s00221-010-2264-9

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Berman, R. A., & Colby, C. L. (2002). Auditory and visual attention modulate motion processing in area MT. *Neuropsychologia*, *14*, 64–74. https://doi.org/10.1016/s0926-6410(02)00061-7

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

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer (Version 6.0.19) [Computer program]. Retrieved from https://dx.www.praat.org/

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Brancazio, L., & Miller, J. L. (2005). Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception & Psychophysics*, *67*, 759–769. https://doi.org/10.3758/BF03193531

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73–101. https://doi.org/10.1146/annurev.psych.093008.100427

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology*, *113*, 495–506. https://doi.org/10.1016/s1388-2457(02)00024-x

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews*, *3*, 201–215. https://doi.org/10.1038/nrn755

Dias, J. W., Cook, T. C., & Rosenblum, L. D. (2016). Influences of selective adaptation on perception of audiovisual speech. *Journal of Phonetics*, *56*, 75–84. https://doi.org/10.1016/j.wocn.2016.02.004

Diehl, R. L. (1975). The effect of selective adaptation on the identification of speech sounds. *Perception & Psychophysics*, *17*, 48–52. https://doi.org/10.3758/BF03203996

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*, 99–109. https://doi.org/10.1016/0010-0285(73)90006-6

Heald, S. L. M., & Nusbaum, H. C. (2014). Talker variability in audio–visual speech perception. *Frontiers in Psychology*, *5*, 698. https://doi.org/10.3389/fpsyg.2014.00698

Houghton, R. J., Macken, W. J., & Jones, D. M. (2003). Attentional modulation of the visual motion aftereffect has a central cognitive locus: Evidence of interference by the postcategorical on the precategorical. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 731–740. https://doi.org/10.1037/0096-1523.29.4.731

Huyck, J. J., & Johnsrude, I. S. (2012). Rapid perceptual learning of noise-vocoded speech requires attention. *Journal of the Acoustical Society of America*, *131*, EL236–EL242. https://doi.org/10.1121/1.3685511

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the *N*-back task as a working memory measure. *Memory*, *18*, 394–412. https://doi.org/10.1080/09658211003702171

Janse, E., & Adank, P. (2012). Predicting foreign-accent adaptation in older adults. *Quarterly Journal of Experimental Psychology*, *65*, 1563–1585. https://doi.org/10.1080/17470218.2012.658822

Jesse, A., & Bartoli, M. (2018). Learning to recognize unfamiliar talkers: Listeners rapidly form representations of facial dynamic signatures. *Cognition*, *176*, 195–208. https://doi.org/10.1016/j.cognition.2018.03.018

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics*, *72*, 209–225. https://doi.org/10.3758/APP.72.1.209

Jesse, A., Vrignaud, N., Cohen, M. A., & Massaro, D. W. (2000). The processing of information from multiple sources in simultaneous interpreting. *Interpreting*, *5*, 95–115. https://doi.org/10.1075/intp.5.2.04jes

Kahneman, D., & Chajczyk, D. (1983). Tests of the automaticity of reading: dilution of Stroop effects by color-irrelevant stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 497–509. https://doi.org/10.1037/0096-1523.9.4.497

Kajander, D., Kaplan, E., & Jesse, A. (2016). Attention modulates cross-modal retuning of phonetic categories to speakers. *Abstracts of the Psychonomic Society*, *21*, 114.

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

Kilian-Hütten, N., Vroomen, J., & Formisano, E. (2011). Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. *NeuroImage*, *57*, 1601–1607. https://doi.org/10.1016/j.neuroimage.2011.05.043

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 451–468. https://doi.org/10.1037/0096-1523.21.3.451

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, *56*, 183–197. https://doi.org/10.3758/BF03213897

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 391–409. https://doi.org/10.1037/0096-1523.33.2.391

Massaro, D. W. (1987). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale: Erlbaum.

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

Mattys, S. L., Barden, K., & Samuel, A. G. (2014). Extrinsic cognitive load impairs low-level speech perception. *Psychonomic Bulletin & Review*, *21*, 748–754. https://doi.org/10.3758/s13423-013-0544-7

Mattys, S. L., & Palmer, S. D. (2015). Divided attention disrupts perceptual encoding during speech recognition. *Journal of the Acoustical Society of America*, *137*, 1464–1472. https://doi.org/10.1121/1.4913507

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *65*, 145–160. https://doi.org/10.1016/j.jml.2011.04.004

Murphy, G., & Greene, C. M. (2017). The elephant in the road: Auditory perceptual load affects driver perception and awareness. *Applied Cognitive Psychology*, *31*, 258–263. https://doi.org/10.1002/acp.3311

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

Prabhakaran, V., Narayanan, K., Zhao, Z., & Gabrieli, J. D. E. (2000). Integration of diverse information in working memory within the frontal lobe. *Nature Reviews*, *3*, 85–90. https://doi.org/10.1038/71156

R Core Team. (2014). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from https://dx.www.R-project.org/

Rees, G., Frith, C., & Lavie, N. (2001). Processing of irrelevant visual motion during performance of an auditory attention task. *Neuropsychologia*, *39*, 937–949. https://doi.org/10.1016/s0028-3932(01)00016-1

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), Hearing by eye: The psychology of lipreading (pp. 97–113). Hillsdale: Erlbaum.

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics*, *30*, 309–314. https://doi.org/10.3758/BF03206144

Rosenblum, L. D., & Saldana, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 318–331. https://doi.org/10.1037/0096-1523.22.2.318

Saldana, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *Journal of the Acoustical Society of America*, *95*, 3658–3661. https://doi.org/10.1121/1.409935

Samuel, A. G. (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. *Cognitive Psychology*, *88*, 88–114. https://doi.org/10.1016/j.cogpsych.2016.06.007

Samuel, A. G., & Kat, D. (1998). Adaptation is automatic. *Perception & Psychophysics*, *60*, 503–510. https://doi.org/10.3758/bf03206870

Samuel, A. G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 1479–1490. https://doi.org/10.1037/a0036656

Santangelo, V., & Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1311–1321. https://doi.org/10.1037/0096-1523.33.6.1311

Scharenborg, O., Weber, A., & Janse, E. (2014). The role of attentional abilities in lexically guided perceptual learning by older listeners. *Attention, Perception, & Psychophysics*, *77*, 493–507. https://doi.org/10.3758/s13414-014-0792-2

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S., & Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition*, *115*, 435–443. https://doi.org/10.1016/j.cognition.2010.03.004

Sinnett, S., Costa, A., & Soto-Faraco, S. (2018). Manipulating inattentional blindness within and across sensory modalities. *Quarterly Journal of Experimental Psychology*, *59*, 1425–1442. https://doi.org/10.1080/17470210500298948

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition*, *92*, B13–23. https://doi.org/10.1016/j.cognition.2003.10.005

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

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, *36A*, 51–74. https://doi.org/10.1080/14640748408401503

Sussman, J. E. (1993). Focused attention during selective adaptation along a place of articulation continuum. *The Journal of the Acoustical Society of America*, *93*, 488–498. https://doi.org/10.1121/1.405629

Theeuwes, J. (1991). Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Perception & Psychophysics*, *49*, 83–90. https://doi.org/10.3758/BF03211619

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology*, *5*, 725. https://doi.org/10.3389/fpsyg.2014.00725

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio–visual speech perception is special. *Cognition*, *96*, B13–B22. https://doi.org/10.1016/j.cognition.2004.10.004

van der Zande, P., Jesse, A., & Cutler, A. (2014). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics*, *43*, 38–46. https://doi.org/10.1016/j.wocn.2014.01.003

van der Zande, P., Jesse, A., & Cutler, A. (2013). Lexically guided retuning of visual phonetic categories. *Journal of the Acoustical Society of America*, *134*, 562–571. https://doi.org/10.1121/1.4807814

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. https://doi.org/10.1037/0096-1523.33.6.1483

van Linden, S., & Vroomen, J. (2008). Audiovisual speech recalibration in children. *Journal of Child Language*, *35*, 809–814. https://doi.org/10.1017/S0305000908008817

Vroomen, J., & Baart, M. (2009a). Phonetic recalibration only occurs in speech mode. *Cognition*, *110*, 254–259. https://doi.org/10.1016/j.cognition.2008.10.015

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

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. 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*, 55–61. https://doi.org/10.1016/j.specom.2004.03.009

Wahn, B., & König, P. (2015). Audition and vision share spatial attentional resources, yet attentional load does not disrupt audiovisual integration. *Frontiers in Psychology*, *6*, 14608. https://doi.org/10.3389/fpsyg.2015.01084

Wahn, B., & König, P. (2017). Is attentional resource allocation across sensory modalities task-dependent? *Advances in Cognitive Psychology*, *13*, 83–96. https://doi.org/10.5709/acp-0209-2

Wahn, B., Murali, S., Sinnett, S., & König, P. (2017). Auditory stimulus detection partially depends on visuospatial attentional resources. *I-Perception*, *8*, 204166951668802. https://doi.org/10.1177/2041669516688026

Walden, B. E., Prosek, R. A., & Worthington, D. W. (1974). Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech, Language, and Hearing Research*, *17*, 270–278. https://doi.org/10.1044/jshr.1702.270

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, 433. https://doi.org/10.3389/fpsyg.2013.00433/abstract

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, *16*, 1173–1184. https://doi.org/10.1162/0898929041920522

Woodman, G. F., Luck, S. J., & Schall, J. D. (2007). The role of working memory representations in the control of attention. *Cerebral Cortex*, *17*(Supp. 1), i118–i124. https://doi.org/10.1093/cercor/bhm065

Wright, B. A., Sabin, A. T., Zhang, Y., Marrone, N., & Fitzgerald, M. B. (2010). Enhancing perceptual learning by combining practice with periods of additional sensory stimulation. *Journal of Neuroscience*, *30*, 12868–12877. https://doi.org/10.1523/JNEUROSCI.0487-10.2010

Yakel, D. A., Rosenblum, L. D., & Fortier, M. A. (2000). Effects of talker variability on speech-reading. *Perception & Psychophysics*, *62*, 1405–1412. https://doi.org/10.3758/BF03212142

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 601–621. https://doi.org/10.1037/0096-1523.10.5.601

Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 200–217. https://doi.org/10.1037/a0033182