Congruent lip movements facilitate speech processing in a dynamic audiovisual multi-talker scenario: An ERP study with older and younger adults

Alexandra Begau*, Laura-Isabelle Klatt, Edmund Wascher, Daniel Schneider & Stephan Getzmann

Leibniz Research Centre for Working Environment and Human Factors, TU Dortmund, Germany

*Address of correspondence:
Alexandra Begau
Leibniz Research Centre for Working Environment and Human Factors
Ardeystraße 67
44139 Dortmund
Germany
Tel.: +49 / 231 1084 264
E-mail: begau@ifado.de

Abstract

In natural conversations, visible mouth and lip movements play an important role in speech comprehension. There is evidence that visual speech information improves speech comprehension, especially for older adults and under difficult listening conditions. However, the neurocognitive basis is still poorly understood. The present EEG experiment investigated the benefits of audiovisual speech in a dynamic cocktail-party scenario with 22 (aged 20 to 34 years) younger and 20 (aged 55 to 74 years) older participants. We presented three simultaneously talking faces with a varying amount of visual speech input (still faces, visually unspecific and audiovisually congruent). In a two-alternative forced-choice task, participants had to discriminate target words ("yes" or "no") among two distractors (one-digit number words). In half of the experimental blocks, the target was always presented from a central position, in the other half, occasional switches to a lateral position could occur. We investigated behavioral and electrophysiological modulations due to age, location switches and the content of visual information, analyzing response times and accuracy as well as the P1, N1, P2, N2 event-related potentials (ERPs) and the contingent negative variation (CNV) in the EEG. We found that audiovisually congruent speech information improved performance and modulated ERP amplitudes in both age groups, suggesting enhanced preparation and integration of the subsequent auditory input. However, these benefits were only observed as long as no location switches occurred. To conclude, meaningful visual information in a multi-talker setting, when presented from the expected location, is shown to be beneficial for both younger and older adults.

Keywords: audiovisual integration, multi-talker speech perception, aging, attention, event-related potentials, multisensory processing

1. Introduction

Imagine meeting for lunch with a group of people. You ask your colleague from across the table about their weekend, but you can barely understand what they are saying because next to you, other people are also chatting. It's long known that verbal conversation in loud and noisy environments and in the presence of concurrent talkers is hard work for our auditory system (Alain, McDonald, Ostroff, & Schneider, 2004). In such "cocktail-party" situations (Cherry, 1953), extracting relevant speech content from a multitude of speech streams mainly demands two cognitive processes: Firstly, the segregation of target speech against the irrelevant speech streams and secondly, the streaming of speech or the connection of speech elements across time (Bronkhorst, 2015; Hill & Miller, 2010). Such a pre-attentive semantic processing of speech, also called grouping, is required to trigger attentive processes and to guide selective auditory attention (Bronkhorst, 2015; Carlile, 2014). Attentional demands are highly modulated by the listening situation itself. For example, "homing in" auditory attention onto a talker of interest takes some time, and therefore speech comprehension is enhanced when the spatial location of the perceived target remains the same (Best, Ozmeral, Kopco, & Shinn-Cunningham, 2008). In everyday life, however, conversational situations are often highly dynamic and changing speakers pose a particular challenge for the auditory system as switches in talker locations require a continuous re-focusing of attention (Koch, Lawo, Fels, & Vorländer, 2011; Lawo, Fels, Oberem, & Koch, 2014; Lin & Carlile, 2015).

Nevertheless, speech has several advantageous characteristics that makes it distinguishable from other acoustic signals and therefore especially suitable for situations with multiple talkers. It proves resistant to background masking, making a single speech stream stand out amongst the noise. The high redundancy of speech provides a "safety net", enabling the listener to fill in the gap where information is missing due to noise or background babble (for review, see Bronkhorst, 2015). Expectations and prior knowledge about the speech

content play an important role in these situations. Also, the familiarity of a talker's voice (e.g. Johnsrude et al., 2013) and the anticipation of a target talker's location among concurrent talkers (Kidd, Arbogast, Mason, & Gallun, 2005) are used to filter out relevant information out of a mixture of auditory input.

A further important source of redundancy of speech, which is often not taken into account in studies on speech comprehension, is visual speech information. In natural speech, the expectations concerning the auditory input are modulated by complementary visual information (Lindström, 2012; van Wassenhove, Grant, & Poeppel, 2005), enhancing both audiovisual processing (Besle, Fort, Delpuech, & Giard, 2004; for review see Campbell, 2008; Tye-Murray, Spehar, Myerson, Hale, & Sommers, 2016; Winneke & Phillips, 2011) and speech intelligibility (Brault, Gilbert, Lansing, McCarley, & Kramer, 2010). For audiovisual speech integration to be successful, the typically preceding visual stimulus needs to be a valid predictor of the auditory input (Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005), in line with the analysis-by-synthesis model by van Wassenhove et al. (2005). In order to perceive auditory and visual speech as one coherent audiovisual event, it has been suggested that a general binding process of both modalities occurs first, followed by temporal, spatial and finally phonemic integration of auditory and visual speech (Baart, Stekelenburg, & Vroomen, 2014; Ganesh, Berthommier, Vilain, Sato, & Schwartz, 2014). Thus, integration of both modalities happens early in the process of speech perception, even pre-phonemically, which leads to a better performance (Besle et al., 2004; Campbell, 2008; van Wassenhove et al., 2005) possibly due to freed cognitive resources enabling a more precise interpretation of the input (Campbell, 2008). Benefits of audiovisual speech-processing are therefore especially pronounced under difficult listening conditions (Grant & Seitz, 2000; Heald & Nusbaum, 2014; Schwartz, Berthommier, & Savariaux, 2004).

Audiovisual benefits should also be evident when individual prerequisites for undisturbed speech comprehension are not fulfilled, for example, in people with hearing disorders or at an advanced age. Especially age-related declines in hearing abilities and in the processing of speech can be regarded as a prototypical example of how audiovisual input could enhance speech comprehension. Decline occurs on both the sensory and cognitive level: Hearing loss (Presbycusis) and neuronal changes, resulting in a deteriorated input of sensory information (see Peelle & Wingfield, 2016), declines in general attentional mechanisms, working memory, and processing speed (e.g. Burke & Shafto, 2015; Kropotov, Ponomarev, Tereshchenko, Müller, & Jäncke, 2016), as well as a reduced auditory spatial attention (Dai, Best, & Shinn-Cunningham, 2018) contribute to age-related hearing issues. As a consequence, older adults are usually less able to perform well in attentional highly demanding situations such as "cocktail-party" environments, where they experience difficulties with regulating their attentional focus toward the relevant stream of auditory information (Passow et al., 2012). When irregular spatial shifts of the target occur, older adults are even more affected by distractors than younger people (Volosin, Gaál, & Horváth, 2017) and need longer phases of (re-)orientation toward the relevant stimulus (Correa-Jaraba, Cid-Fernández, Lindín, & Díaz, 2016; Getzmann, Falkenstein, & Wascher, 2015).

On the other hand, older people make use of several compensation strategies like the enhancement of preparatory and orienting processes (Kropotov et al., 2016) and stronger top-down regulation of the attentional focus, keeping the behavioral performance relatively stable (Alain et al., 2004). Accordingly, several studies showed an age-related posterior-to-anterior shift in neuronal activation, suggesting compensatory mechanisms based on enhanced engagement of fronto-central brain areas (Davis, Dennis, Daselaar, Fleck, & Cabeza, 2008; Kropotov et al., 2016). Moreover, audiovisual integration is still intact with age, even though visual acuity (see Owsley, 2011) as well as the processing of visual speech, that is, the ability

to lipread, appears to be diminished in older adults (Sommers, Tye-Murray, & Spehar, 2005; Tye-Murray et al., 2016; Winneke & Phillips, 2011). Still, the audiovisual performance is similar, indicating that older adults might profit even more from multimodal speech than younger adults (Winneke & Phillips, 2011). Moreover, older adults rely more on the visual information, especially when the auditory information is ambiguous (Cienkowski & Carney, 2002; Sekiyama, Soshi, & Sakamoto, 2014). At the same time, the quality of the visual content seems to gain importance. While younger adults profit from any meaningful visual content, older adult's performance is supported by high quality visual input only (Tye-Murray, Spehar, Myerson, Sommers, & Hale, 2011).

Taken together, previous research has shown that audiovisual speech perception is important, especially under difficult listening conditions and for older people. However, studies exploring audiovisual speech integration in younger and older adults in dynamic "cocktail-party" situations are scarce and have mainly focused on behavioral measures. In the present study, we investigated the underlying neuro-cognitive underpinnings of audiovisual multi-talker speech processing, using EEG-related measures in combination with a dynamic speech perception paradigm. We focused our analysis on a number of event-related potentials (ERPs) that have already been studied in relation to both audiovisual integration and multi-talker scenarios. In general, auditory ERPs like P1, N1 and P2 have been linked to sensory detection and the cognitive processing of sound (for a review, see Alain & Tremblay, 2007). The first component, the P1, a rather small positive deflection in central scalp regions is modulated by physical characteristics of the auditory stimulus (Alain & Tremblay, 2007) and is thought to reflect early processing of sensory information (Baart, 2016). In the presence of additional visual information, the P1 is usually reduced in amplitude (Stothart & Kazanina, 2016; Winneke & Phillips, 2011), possibly reflecting cross-modal sensory gating, that is, a more effective pre-attentive filtering of sensory information (Lebib, Papo, de Bode, &

Baudonnière, 2003). Accordingly, larger P1 amplitudes can be found, when easily detectable incongruent audiovisual input is presented, indicating an early mechanism to detect non-redundant audiovisual information. Larger amplitudes have also been found in older compared to younger adults, which has been linked to an age-related deficit in inhibition or gating (Friedman, 2011; Getzmann et al., 2015; Stothart & Kazanina, 2016; Winneke & Phillips, 2011).

Following the P1, there is a negative and a subsequent positive deflection called the N1-P2 complex (Alain & Tremblay, 2007). Baart (2016) investigated the influence of visual speech on N1 and P2 in a meta-analysis and found an overall suppression and speeding-up of these components throughout different studies. Van Wassenhove and colleagues (2005) argued within their analysis-by-synthesis model that auditory and visual information interact as early as at the N1-latency. The predictive value of the visual information for the auditory input seems to influence the degree of temporal facilitation in both components. That is, the more salient and predictable the visual input is, the stronger is the facilitation of the auditory processing. Accordingly, higher attentional demands seem to disrupt audiovisual integration early during processing, resulting in a decrement of the latency reduction of both components (Alsius, Möttönen, Sams, Soto-Faraco, & Tiippana, 2014). Importantly, both components also differ with respect to content-dependency. While the N1 is simply driven by anticipation of the auditory speech input due to the preceding visual input, a P2 modulation can only be observed, when the visual input reliably predicts the auditory input and speech specific binding can occur (Ganesh et al., 2014; Stekelenburg & Vroomen, 2007). It is therefore suggested, that the N1 reflects the integration of spatial and temporal audiovisual properties, while the P2 reflects the subsequent integration of phonetic properties (Baart et al., 2014; Klucharev, Möttönen, & Sams, 2003).

N1 is also thought to reflect the extent of distraction when a switch in talker location occurs. Volosin and colleagues (2017) found similar modulations in younger and older adults, arguing that older adults needed no extra time to recover from distraction. Contrarily, a study investigating distraction in a multi-talker scenario found an age-related slowing of both N1 and P2, as well as a P2 amplitude reduction with a present distractor only in younger adults (Getzmann et al., 2015). Taken together, as reviewed by Friedmann (2011), age-related modulations of N1 amplitudes are somewhat ambiguous throughout literature. Similarly, the P2 in relation to aging processes is not well assessed, although the literature also indicated an age-related enhancement and slowing of the component (Friedman, 2011).

Following the P1-N1-P2 complex, a later component, the N2 can be observed that has been linked to attentional control processes and conscious sensory discrimination (Friedman, 2011; Ritter, Simson, Vaughan, & Friedman, 1979). Processes contributing to a N2 modulation in both multimodal and multi-talker scenarios have yet to be unraveled. To date, the literature suggests a number of scarce or ambiguous findings. A larger N2 amplitude is correlated with better performance and faster reaction times in audiovisual processing (Lindström, 2012). In regard to age-related modulations of the N2 amplitude, both increases and decreases in amplitudes and consistent increases in latencies have been observed (for a review see Friedman, 2011). Several more recent studies showed a decline or absence of N2 amplitudes in older adults, in both audiovisual (Stothart & Kazanina, 2016) and multi-talker scenarios (Getzmann et al., 2015; Getzmann & Wascher, 2016), supporting the notion of an age-related inhibitory deficit. Stothart and Kazanina (2016) suggested that congruent visual information helps the elderly to process auditory input, while processing incongruent audiovisual information comes with greater conflict costs. Getzmann and Wascher (2016) found a decrease in N2 amplitude in older adults only when unexpected irrelevant information was presented, arguing that due to age, the control over irrelevant information is weakened.

In addition to these ERPs related to the onset of speech stimuli, a pronounced negative deflection immediately preceding the onset of auditory speech can be observed, which has been related to the Contingent Negative Variation (CNV, Walter, Cooper, Aldridge, McCallum, & Winter, 1964). The CNV reflects preparatory processes in expectation of an upcoming, task-relevant event. In a previous multi-talker speech perception study, for example, the CNV was reduced in older participants, which was interpreted as an age-related decline in preparatory activity of an upcoming auditory event (Getzmann, Golob, & Wascher, 2016). In the present literature concerning audiovisual processing, anticipatory slow waves such as the CNV are known, but usually not further investigated (as seen in Alsius et al., 2014; Klucharev et al., 2003). In fact, it is rather discussed as problematic for the interpretation of later ERP differences in terms of cross-modal interaction (Teder-Salejarvi, McDonald, Di Russo, & Hillyard, 2002).

To summarize, audiovisual speech perception involves a number of cognitive mechanisms resulting in an effective processing of both modalities and in several advantages compared to auditory-only speech. Here, we investigated the potential benefit of audiovisual speech information in cocktail-party scenarios, in which the listener is confronted with multiple simultaneously talking persons and occasional changes in talker locations. Specifically, three talkers simultaneously uttered short words, one containing the target information and two containing concurrent speech input. We varied the information content of visual speech information (i.e., audiovisually congruent speech input, visually unspecific speech input, no speech input) in order to assess its influence on speech comprehension. In addition, we employed two different listening scenarios, a *static* setting, in which the target talker location always remained the same, and a *dynamic* setting, in which the target talker occasionally switched position with the concurrent talkers. The *dynamic* setting was used to explore the influence of changes in talker locations on the processing of audiovisual speech

information. We analyzed both behavioral performance and electrophysiological measures of younger and older adults, expecting to see audiovisual facilitation in form of a faster and more accurate detection of the target information with congruent visual speech input, relative to visually unspecific or no visual input. We hypothesized older adults to show a worsened performance, that is, slower responses and lower accuracy, compared to younger adults. On the electrophysiological level, we expected reduced P1 amplitudes, when audiovisual congruent information is presented as well as modulations in N1 and P2 amplitudes. Here, we expected audiovisual congruency to be important only for the modulation of P2, but not for N1 amplitudes. In case of the N2, we expected larger amplitudes with audiovisually congruent information. Modulations should differ across age-groups. Even though the literature presents ambiguous results, we expected amplitude enhancements in the early P1-N1-P2 complex and a decrease or absence of N2 amplitudes in the older group due to an age-related inhibitory deficit. Finally, with regard to the *dynamic* scenario, we expected a reduced benefit of visual speech input when a change in target talker location occurred, given that the visual information presented at the expected target position cannot be used for integration with the auditory target speech information.

2. Method

2.1 Participants

50 participants were invited to participate in the study. Eight participants had to be excluded due to severe hearing problems (n = 4), visual problems (n = 1), technical issues (n = 2), and missing compliance to finish the experiment (n = 1). Thus, data of 42 participants could be analyzed, 22 of them were aged 20 to 34 years (M_{young} = 25.45, SE_{young} = 0.70; 12 female) and 20 were aged 55 to 70 years (M_{old} = 64.50, SE_{old} = 0.98; 11 female). All participants reported to be right-handed and mentally fit without any history of neurological

diseases or any psychopharmacological medication. Before starting any testing, each participant gave written informed consent. They received a payment of 10€ per hour at the end of the experiment. The experimental procedure was in accordance with the Declaration of Helsinki, approved by the Ethical Committee of the Leibniz Research Centre for Working Environment and Human Factors, Dortmund, Germany.

2.2 Sensory and cognitive abilities

Hearing performance was assessed with a pure-tone audiometry (Oscilla USB 330; Inmedico, Lystrup, Denmark) at eleven pure-tone frequencies in-between 125–8000 Hz. Participants mostly achieved a hearing threshold ≤ 30 dB below 4000 Hz, even though mild to moderate presbycusis was observed in the older group. Mild hearing loss was observed in several participants. At 2000 Hz, thresholds were ≤ 35 dB for two participants. At 3000 Hz, we observed thresholds ≤ 35 dB (n = 2), ≤ 40 dB (n = 2) and ≤ 45 dB (n = 1). Given the relatively loud stimulus presentation (75 dB), we considered this still acceptable.

Visual acuity was measured using *Landolt C optotypes* in 1.5 m distance and was at least 0.75 for both eyes in all participants, while a value of 1.0 is considered “normal” visual acuity (ISO 8596:2017(E)). Contrast sensitivity was assessed using the *Pelli-Robson Contrast Sensitivity Chart* in 1.0m distance (Pelli, Robson, & Wilkins, 1988). A score of 1.65 is recommended as lowest contrast sensitivity measure (Mäntyjärvi & Laitinen, 2001) and was achieved by all but one (older female) participant with a score of 1.5. Between-group comparisons revealed lower scores in visual acuity (V = 77.5, p <.001, r = .58) and contrast sensitivity (V = 68, p <.001, r = .65) in older compared to younger participants.

Cognitive functions were assessed using the *Montreal Cognitive Assessment* (MoCA; Nasreddine et al., 2005), a screening tool for Mild Cognitive Impairment. Older participants scored overall lower than younger ones (M_{old} = 26.40, SE_{old} = 0.62; M_{young} = 28.68, SE_{young} =0.29; V = 90.5, p < .001, r = .51), while all participants met the recommended cut-off

criterion of 17 points as an indication for dementia (Carson, Leach, & Murphy, 2018). Working memory capacity was measured using the Digit Span subtest from the German Adaptation of the *Wechsler Adult Intelligence Scale* (Hamburg-Wechsler Intelligenztest für Erwachsene – Revision 1991; HAWIE-R, Tewes, U., & Wechsler, 1991). Younger participants performed better than the older ones in forward recall (M_{young} = 9.59 digits, SE_{young} = 0.41, M_{old} = 6.90, SE_{old} = 0.46; $t(39) = -4.37$, $p < .001$, $g = -1.34$) and in backward recall (M_{young} = 8.54 digits, SE_{young} = 0.39, M_{old} = 6.84, SE_{old} = 0.45; $t(39) = -2.72$, $p = .010$, $g = -0.84$).

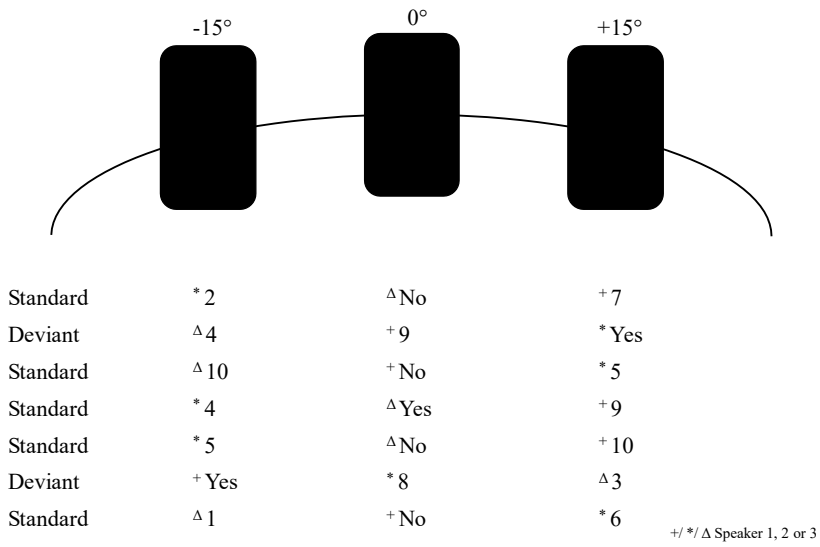
2.3 Materials and stimuli

Participants were seated in a dimly lit, sound attenuated room (5.0 × 3.3 × 2.4 m³). Attenuation was achieved through pyramid-shaped foam panels on ceiling and walls and a woolen carpet on the floor, resulting in a background noise level below 20 dB (A). For audiovisual stimulus presentation, we used three 12” vertically aligned monitors with a 1080 by 1920-pixel resolution and a 50 Hz refresh rate (Beetronics, Düsseldorf, Germany). The auditory signal was presented at about 75 dB SPL using three full range loudspeakers (SC 55.9 - 8 Ohm; Visaton, Haan, Germany), each mounted underneath a monitor. To achieve a free-field listening scenario, monitors and loudspeakers were installed on a horizontal array at -15° (left), 0° (center) and +15° (right) azimuth at 1.5m distance in approximate head height (lower frame at 1.00 m (loudspeaker) and 1.12 m (monitor), respectively, figure 1A).

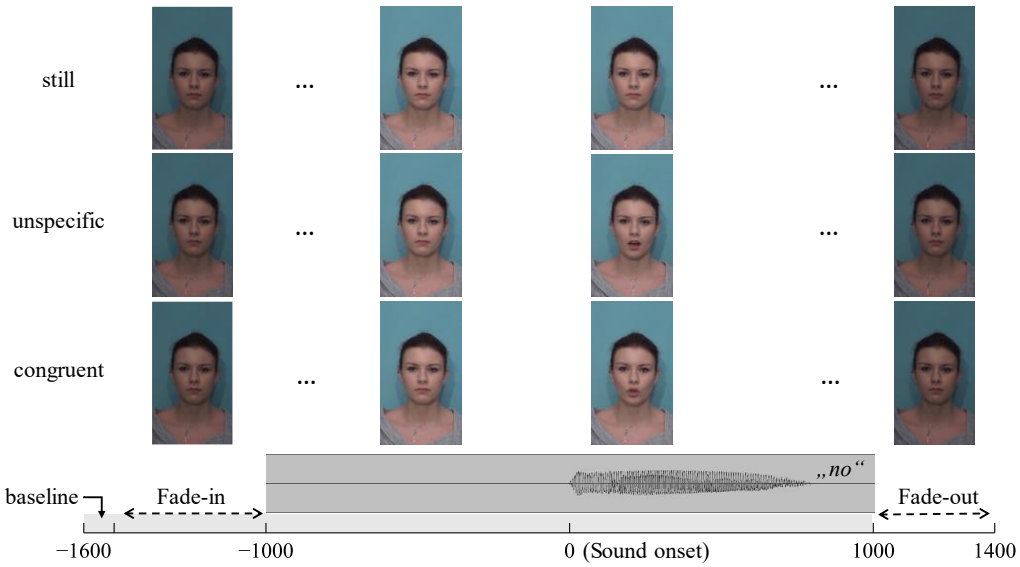
Figure 1

Experimental setup and audiovisual information

A)



B)



Note. (A) Three monitor-/loudspeaker combinations were mounted on a horizontal array, on which three talkers were shown. Per trial, each of the talkers simultaneously uttered a word, either the target words (“yes” or “no”) or concurrent words (number words 1-10). In the

dynamic setting, target words were either presented from the central position (0°, standards, 80% of trials) or from lateral positions (+/-15°, deviants, each 10% of trials). In the static setting (not shown here), target words were exclusively presented from the standard position. (B) Three different types of visual information were combined with the auditory speech information, (a) *still* faces without any motion, (b) visually *unspecific* stimuli, in which the visual speech information consisted of the person uttering the word "ba", or (c) audiovisually *congruent* information. Visual information was faded in and out, the onset of target speech sound (here "no") was exactly synchronized with the concurrent speech sounds.

As stimuli, we used short video sequences showing a talker's face and neck in front of a light-blue background (RGB values 106, 145, 161) while uttering brief words. The words were "yes" and "no" (in English) as target stimuli, and the numbers from one to ten in German language as distractor stimuli. The target words were chosen to differ clearly in visual articulation, thus providing the participants with rich visual speech information. The talkers were three young native dialect-free German females with average fundamental voice frequencies of 169 Hz (*SE* = 2.00), 213 Hz (*SE* = 7.87) and 178 Hz (*SE* = 12.22), respectively. The duration of a word was between 475 and 705 ms (*M* = 549.67, *SE* = 7.55). The talkers were instructed to pronounce the words clearly and accentuated, but to avoid unnecessary head movements.

Videos were recorded with a 1920 × 1080-pixel resolution and a 50fps frame rate under sound-attenuated conditions. While videotaping, mono audio tracks were separately recorded at 48 kHz and a 24-bit resolution using a dynamic USB-microphone (Podcaster, RØDE, Silverwater, NSW, Australia). Recording and editing was done in Audacity(R) (version 2.3.0). Audio tracks were noise reduced by 15 dB and normalized to -6 dB. Auditory speech onset was set to 1500 ms in order to ensure a synchronous onset of auditory speech in the experiment. Visual speech started and ended with the opening and closing of the talker's mouth, resulting in visual movements that lasted between 640 and 1540 ms (*M* = 1087.43 ms,

SE = 42.45). The visual speech onset preceded the auditory onset by 414.29ms on average (*range* = 80 – 860 ms, *SE* = 34.73). We added a 500-ms fade-in and a 400-ms fade-out with freeze-frames fading from and to black. Freeze frames were taken from the beginning and ending of the visual speech video showing the talkers with their mouth closed. Between fades and speech videos, we added a variable amount of freeze frames, resulting in all stimulus videos being exactly 2900 ms long (figure 1B).

We further edited the videos to vary the content of visual speech information provided to the participants. We generated video clips with audiovisually congruent, visually unspecific, and no visual speech information (still face). For visually unspecific information, we additionally recorded the talkers while uttering the word “ba” and manipulated the visual speech information. Thus, while the auditory information remained unaltered, the visual “ba” was used to replace the “yes” and “no” as well as the number words in the originally congruent videos. The resulting audiovisual videos exactly matched the timing of audiovisual congruent videos, but without containing any visual speech information that could be used for the discrimination of the target words. For the condition without visual speech information, we replaced the audiovisually congruent videos with a freeze frame of the talkers with their mouth closed, keeping the overall structure of the video unchanged (figure 1B). Videos were rendered to MP4. During the experiment, each stimulus video was followed by a fixed ISI of 1000 ms showing a black screen.

2.4 Experimental Procedure

Prior to the EEG experiment, every participant went through an assessment of their sensory and cognitive abilities (see 2.2). Then, participants received written and oral instructions on the task of the EEG experiment. In a two-alternative forced-choice task, the participants had to indicate as fast and accurate as possible which target word (“yes” or “no”) they heard. For this, one of the two buttons on a keypad had to be pressed with the index or

middle finger of the dominant hand; the left button being assigned to “yes” (50 % of trials), the other one to “no”. Each trial consisted of the simultaneous audiovisual presentation of the three talkers, each uttering a word, always presenting one of the two target words among two distractor words (i.e., German numerals). Trial order, target talkers and talker locations followed a pseudo-randomized order. In total, 900 trials were presented in 6 blocks. Half of those blocks were *static* blocks consisting of standard trials, in which the target words were presented exclusively from the central location. The other half were *dynamic* blocks, consisting of 80 % standard trials and 20 % deviant trials, in which the target word was either presented from the left (10 %) or right (10 %) location. A deviant trial was always followed by at least two and at most eight (on average 3.92) standard trials. Static and dynamic blocks comprised 120 and 180 trials, respectively. Within the three static and three dynamic blocks, either audiovisually congruent, visually unspecific, and no visual speech information was presented in separate task conditions. The order of static and dynamic blocks as well as audiovisual conditions was counterbalanced across participants to avoid sequence effects. Before a new block started, the participant was informed about the upcoming audiovisual condition (audiovisually congruent, visually unspecific, or no visual speech information) and whether switches of the target talker could occur or not (static or dynamic block). Between every block, participants could have a short self-paced rest, no feedback was given during the experiment. Participants were instructed to direct their gaze at the central monitor at all times and to avoid eye movements.

2.5 Lipreading assessment

To assess the participants’ ability to lipread, we constructed a short lip-reading assessment that followed the experimental procedure. We recorded a single female talker uttering 30 randomly assembled five-word sentences in German taken from the Oldenburger Satztest (Wagener, Kühnel, & Kollmeier, 1999). A sentence was always built the same way

and consisted of a name, verb, number, adjective, and noun (e.g. “Kerstin took two old knives”). The sentences were recorded using the same equipment and setup that was used for creating the stimulus material for the EEG experiment, always starting and ending with the talker’s mouth closed. Audio tracks were removed and a 1000-ms fade with freeze frames from black was added at the beginning and end of the video using the software Shotcut. Additional freeze frames were added prior to video onset to achieve a theoretical sound onset at 2000 ms for each video. Lip movement duration was between 2940 and 4180 ms. All videos were arranged in a presentation, presented on a MacBook positioned on the table in front of the participant. The assessment started with a detailed oral and on-screen instruction and the presentation of five example videos. After each video, two tests followed – a *cued* and a *recognition* test. For the cued test, two words were displayed on the computer screen before the video started, one of which being included in the following sentence. After watching, the participants were instructed to speak out loud which one of the two words they had seen. Then, in the recognition test, a single word was displayed, and the participants had to state, if that word had been part of the sentence or not. For analysis, we calculated a sum score for each test based on correct answers, resulting in a possible score between 0 and 30 points.

Younger participants correctly answered 23.96 (*range* = 12-30, *SE* = 0.99) of the cued questions and 17.65 (*range* = 13-27, *SE* = 0.68) of the recognition questions. Older participants correctly answered an average of 23.30 cued questions (*range* = 15-29, *SE* = 0.98) and 17.35 (*range* = 12-22, *SE* = 0.60) of the recognition questions. There were no group differences in neither the cued, $V = 194$, $p = .519$, $r = .10$, nor the recognition task, $t(38.64) = -1.03$, $p = .310$, $g = -0.31$, indicating similar lipreading performance in both age groups despite worse visual abilities in the older participants. Consequently, we did not further include the test in our analysis.

2.6 EEG recording and preprocessing

For continuous EEG recordings, we used a 64 Ag/AgCl electrode cap (BrainCap; Brainvision, Gilching, Germany); the electrodes were distributed across the scalp according to the extended international 10-20 system. The signal was recorded with a sampling rate of 1000 Hz (QuickAmp DC Amplifier, Brainvision, Gilching, Germany). Impedance was kept below 10 kΩ. During the recording, the electrodes AFz and FCz were used as ground and reference electrodes, respectively.

For data preprocessing and analysis, we used MATLAB (2019a) and the open-source toolboxes EEGLAB (14-1-2b, Delorme & Makeig, 2004) and ERPLAB (v7.0, Lopez-Calderon & Luck, 2014). All data were filtered using a 0.5 Hz high-pass filter (6601-point FIR filter, 0.5 Hz transition bandwidth, 0.25 Hz cut-off frequency) and a 30 Hz low-pass filter (441-point FIR filter, 7.5 Hz transition bandwidth, cut-off frequency 33.75 Hz). The data were re-referenced to the average of all electrodes and segmented into 2600-ms epochs covering periods from -1600 to 1000 ms relative to the auditory speech onset. The first 100 ms of the epoch served as baseline period (figure 1B). For artifact rejection, we applied independent component analysis (ICA). To reduce computation time, ICA-decomposition was based on a subset of data that contained only every second trial and was down-sampled to 250 Hz. The resulting ICA weights were then transferred back onto the ‘original’ dataset, containing all trials with a sampling rate of 1000 Hz. Artifacted components containing general discontinuities, eye blinks and eye movements where identified and subtracted from the data using the automatic artifact detection algorithm ADJUST (Mognon, Jovicich, Bruzzone, & Buiatti, 2011). Moreover, dipoles were fitted to a spherical head model using EEGLAB’s function dipfit. Components with a residual variance above 40 % in the dipole solution were subtracted from the data. Finally, we additionally removed all trials with extreme fluctuations above 1000 Hz, using the built-in EEGLAB auto-rejection function (prob. threshold: 3 std.). On average, 21.43 ICs were rejected (*range* = 8 - 38) and 15.54 % of trials were excluded per

subject (*range* = 7.44 - 26.33 %). For further analyses, only trials with correct responses were included, which led to an additional exclusion of an average of 40.30 trials per subject (being 0.56 to 21.22 % of the original amount of trials). After preprocessing, the data were down-sampled to 500 Hz for further analyses.

2.7 Data Analysis

We performed statistical analysis using R (v3.6.3) in RStudio (RStudio Team, 2020, version 1.2.5042). For the performed analyses of variance (ANOVAs) and subsequent post-hoc pairwise *t*-tests, partial eta squared η_p^2 (Cohen, 1973) and Hedges's *g*, (i.e., corrected Cohen's *d* to account for bias due to smaller sample sizes, Fritz, Morris, & Richler, 2012) were calculated as effect size measures, respectively. Multiple comparisons were corrected using FDR (Benjamini & Hochberg, 1995). Those parametric tests were conducted, if the Shapiro-Wilk test yielded $p > .05$ and thus indicated approximately normally distributed data (Field, Miles, & Field, 2012). In case of a violation of the normality assumption, non-parametric analyses were chosen, that is, rank-based ANOVA-type statistics (as provided by the R-package nparLD; Noguchi, Gel, Brunner, & Konietschke, 2012). Note, that except for the whole-plots factor age group, only the nominator degrees of freedom will be reported in case of ANOVA-type statistics, since the denominator is set to infinite. Also note, that *F*-values are not necessarily comparable to parametric *F*-values (Noguchi et al., 2012). For non-parametric post-hoc calculations, we used the Wilcoxon signed-rank test (Wilcoxon, 1945) with *r* as an effect size measure (Fritz et al., 2012).

2.7.1 Behavioral analyses

As behavioral measures, we analyzed accuracy and response times. Response times were defined as the time interval between the end of video fade-in and a button press. Only trials with correct response were taken into the analysis, response times below three standard deviations from the mean were excluded as pre-mature responses. Separate analyses were

performed for the *static* and *dynamic* settings: For the static setting, in which only standard trials (i.e., with targets presented from the central position) were presented, a 2 (young vs. old) × 3 (congruent vs. unspecific vs. still face visual information) repeated-measures ANOVA was conducted. For the dynamic setting, with both standard and deviant trials (i.e., with targets presented from central and lateral positions), we performed a 2 (young vs. old) × 3 (congruent vs. unspecific vs. still face visual information) × 2 (standard vs. deviant target position) repeated-measures ANOVA.

2.7.2 ERP analyses

For ERP analyses, we computed grand-averages across participants, separately for the two age groups, the three audiovisual conditions, and the static and dynamic blocks. In addition, for dynamic blocks, separate averages were computed for deviant and standard trials. However, to avoid post-deviant modulations of the waveforms (e.g., Getzmann & Wascher, 2016), standard trials immediately following a deviant (i.e., the two trials post-deviant presentation) were not considered in the analysis of dynamic blocks. All data were collapsed across both target stimuli and target talkers. ERP data, that is, correct trials, were averaged across a cluster of frontocentral electrodes that included the electrodes Cz, FCz, Fz, FC1 and FC2. To analyze the amplitudes of P1, N1, P2 and N2, we determined the maximum or minimum peak in the grand-average waveforms in a previously defined time window (P1: 20 – 120 ms, N1: 75 – 175 ms, P2: 150 – 250 ms, N2: 250 – 350 ms) at the above-mentioned electrodes. Around the detected peaks, 20-ms time windows were used to compute the mean amplitude. In order to account for the broad negative, CNV-like deflection, which peaked immediately before sound onset and which appeared to differ between audiovisual conditions (cf., figures 3 and 6), we analyzed peak-to-peak measures of the P1, N1, P2 and N2 components. We calculated the absolute difference between the mean amplitudes of two respective components. For the analysis of the CNV, we calculated the mean amplitudes at

the same cluster of frontocentral electrodes within the time window 100 ms prior to the auditory speech onset. Data for the static setting were analyzed in a 2 (young vs. old) × 3 (congruent vs. unspecific vs. still face visual information) repeated-measures ANOVA. For the dynamic setting, again, we calculated repeated-measures ANOVAs with the additional factor target position (standard vs. deviant target position).

3. Results

The analysis of the behavioral and ERP data is divided into two parts: we first looked at the effects of age and visual information in the *static* setting, and then in the *dynamic* setting, where we compared standard and deviant trials.

3.1 *Static setting*

3.1.1 Behavioral data

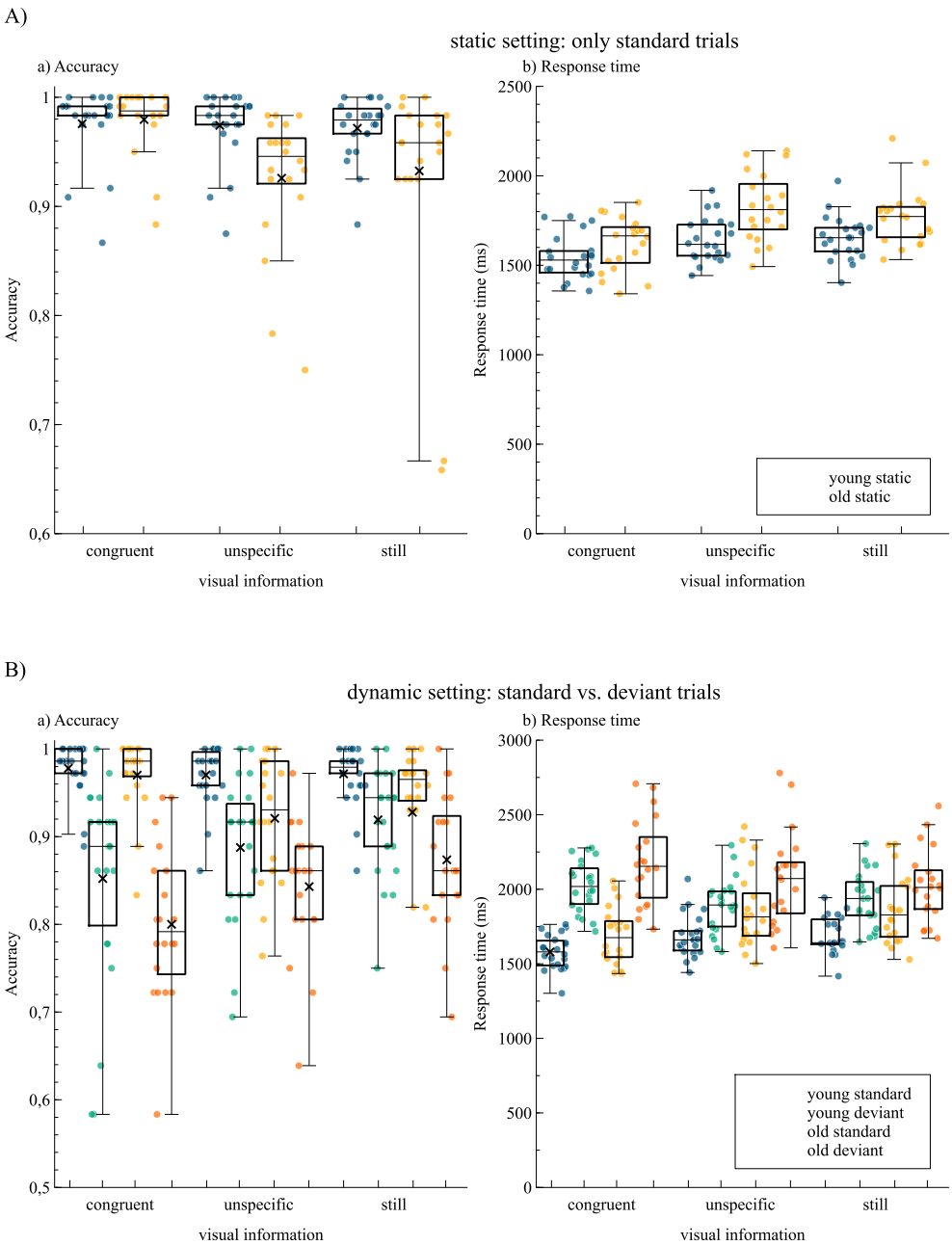
Accuracy.

In figure 2A, behavioral data are visualized, showing accuracy and response times depending on age group and visual information. Older participants reacted less accurate on the task (M_{old} = 94.60 %, SE_{old} = 0.93; M_{young} = 97.39 %, SE_{young} = 0.39; figure 2A; table 1). The rate of correct responses also depended on the visual information, an effect which was modulated by an interaction with age (cf. figure 2A). Post-hoc comparisons showed that older participants performed more accurately with audiovisually congruent stimuli compared to visually unspecific information, V = 176, p_{adj} = .006, r = .74, and still face stimuli, V = 111, p_{adj} = .012, r = .70; performance in the last two conditions did not differ, V = 110, p_{adj} = .576, r = .25. However in younger participants, no differences in accuracy between the content of visual information was shown (congruent vs. still $t(21)$ = 0.72, p_{adj} = .576, g = 0.15; congruent

vs. unspecific, $V = 66$, $p_{adj} = .754$, $r = .0176$; still vs unspecific, $t(21) = -0.70$, $p_{adj} = .576$, $g = -0.14$).

Figure 2

Behavioral results for the (A) static and (B) dynamic setting



Note. Response accuracy (rate of correct responses) and speed (response times) are shown separately for the younger and older group, for congruent, unspecific, and still visual

information, and for standard and deviant trials in the dynamic setting. Box plot whiskers mark 1.5 IQR, mean is indicated by the cross.

Response times.

In the older group, higher response times were found compared to the younger group (M_{old} = 1738.22 ms, SE_{old} = 24.06, M_{young} = 1612.83 ms, SE_{young} = 16.01; figure 2A; table 1). Additionally, the analysis revealed a main effect for visual information, which interacted with the age group. However, post-hoc tests indicated that for both the younger and older group, response times following audiovisually congruent information were lowest compared to visually unspecific information, old: $t(19) = -6.80$, $p_{adj} < .001$, $g = -1.46$, young: $t(21) = -5.70$, $p_{adj} < .001$, $g = -1.17$, and still faces, old: $V = 0$, $p_{adj} < .001$, $r = .88$, young: $V = 16$, $p_{adj} < .001$, $r = .77$. In both groups, there was no difference between visually unspecific and still face stimuli, old: $t(19) = -1.75$, $p_{adj} = .115$, $g = -0.36$,young: $t(21) = 0.85$, $p_{adj} = .403$, $g = 0.18$.

3.1.2 ERP data

Looking at the grand-averaged waveforms at frontocentral scalp locations, the onset of the (still) faces of the talkers (at -1500 ms, figure 3) elicited a sequence of deflections which was related to genuine visual processing and not analyzed here. The onset of the auditory information (at 0 ms, i.e., the time-locking event) elicited a prominent pattern of deflections reflecting the P1, N1, P2, and N2 components, which was preceded by a broad, CNV-like negativity. This series of deflections will be further analyzed and described below.

Table 1

2×3 ANOVA results for trials in static setting

	Accuracy[1]	Response time[2]		P1-N1[1]	N1-P2[2]		P2-N2[1]	CNV[1]
conditions	*F*	*F*	η_p^2	*F*	*F*	η_p^2	*F*	*F*
age	**4.83***	**9.32****	**.19**	**4.25***	1.60	.04	**21.38*****	.01
visual information	**14.26*****	**54.10*****	**.57**	**10.01*****	0.80	.02	**24.97*****	**16.65*****
age × visual information	**9.39*****	**5.69****	**.12**	**3.22***	**3.59***	.08	**4.89***	1.32

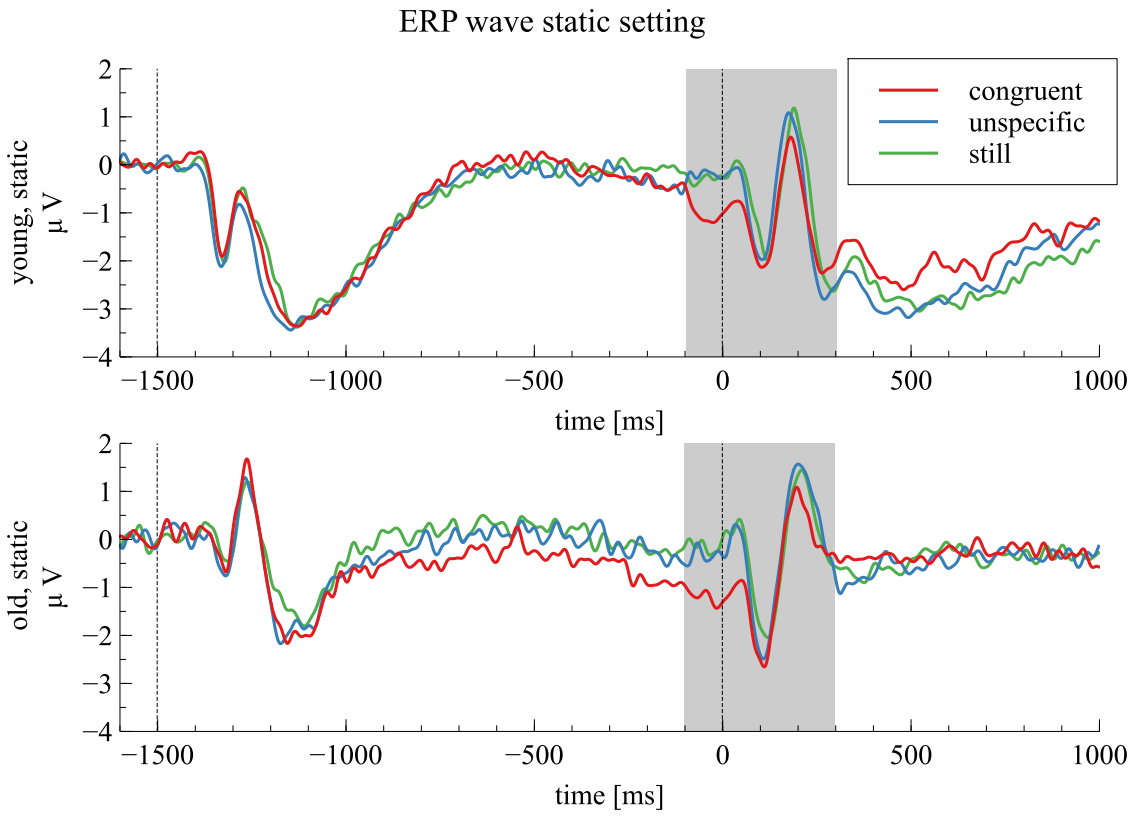
Notes. [1] ANOVA-type statistic; except for whole-plots factors, denominator degrees of freedom are infinite, thus reporting only nominator degrees of freedom. Degrees of freedom: visual information, age × visual information = 1.95 (accuracy), 1.92 (P1-N1), 1.75 (P2-N2), 2.00 (CNV); age = 1, 39.94 (accuracy), 1, 36.42 (P1-N1), 1, 36.40 (P2-N2), 1, 40.00 (CNV)

[2] Degrees of freedom: Age = 1, 40; visual information, age × visual information = 2, 80. For N1-P2 and P2-N2 Mauchly p <.05, p-values were Greenhouse Geisser corrected.

* p < .05, ** p < .01, *** p < .001.

Figure 3

ERP waveforms at electrode cluster Cz, FCz, Fz, FC1 and FC2 in static setting



Note. Waveforms are shown separately for the younger and older group and for audiovisually congruent, visually unspecific, and still visual information. The time interval shows length of one epoch, baseline interval was -1600 to -1500, i.e. prior to stimulus onset, sound onset was at 0 ms. The grey area marks the analyzed part of the ERP wave.

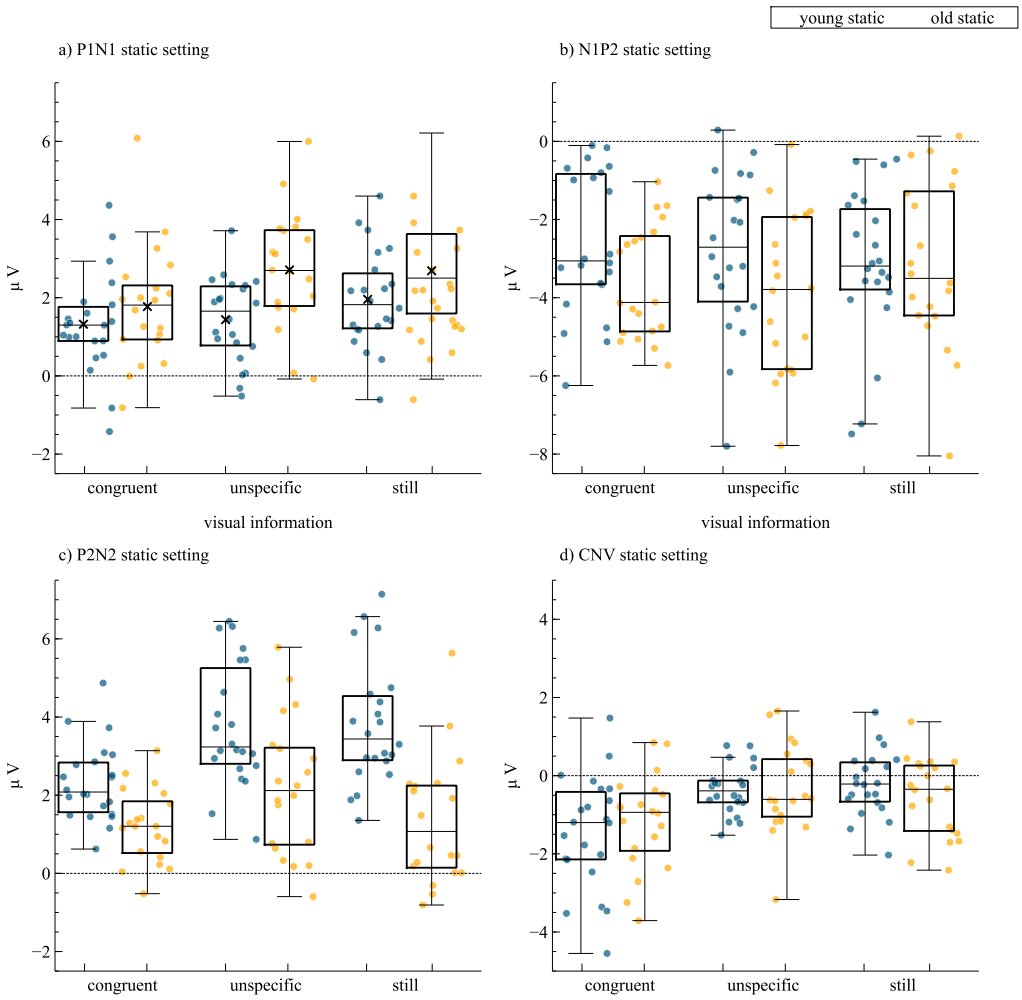
P1-N1.

Older participants had larger P1-N1 peak-to-peak amplitudes than younger participants (M_{old} = 2.47 µV, SE_{old} = 0.20; M_{young} = 1.68 µV, SE_{young} = 0.13; table 1; figure 4a). A main effect of visual information was found that interacted significantly with age group. Post-hoc comparisons revealed that for older adults, audiovisually congruent stimuli elicited smaller peak-to-peak amplitudes compared to visually unspecific stimuli, V = 26, p_{adj} = .006, r = .66 and still faces, t(19) = -4.12, p_{adj} = .003, g = -0.89, with no differences between

the last two, $t(19) = -0.115$, $p_{adj} = .910$, $g = -0.03$. For the younger group, there was a trend for smaller peak-to-peak amplitudes for audiovisually congruent stimuli compared to still faces, $t(21) = -2.27$, $p_{adj} = .051$, $g = -0.47$, while there was no statistically significant difference between audiovisually congruent and visually unspecific speech information, $t(21) = -0.130$, $p_{adj} = .910$, $g = -0.03$. However, visually unspecific speech stimuli elicited smaller peak-to-peak amplitudes than still faces $t(21) = -2.49$, $p_{adj} = .042$, $g = -0.51$.

Figure 4

ERP amplitudes in static setting



Note. Mean amplitudes of a) P1-N1, b) N1-P2 c) P2-N2 absolute difference values as well as d) CNV shown separately for the younger and older group, and for audiovisually congruent, unspecific, and still visual information. Box plot whiskers mark 1.5 IQR, mean is indicated by the cross.

N1-P2.

N1-P2 peak-to-peak amplitudes did not differ neither across age groups, nor across the different conditions of visual information, however, both factors interacted significantly (figure 4b; table 1). Looking at post-hoc comparisons, only amplitudes in older adults for visually unspecific stimuli differed from still faces, with the former eliciting larger peak-to-peak amplitudes, $t(19) = 3.15$, $p_{adj} = .016$, $g = 0.68$ (all other $p > .189$).

P2-N2.

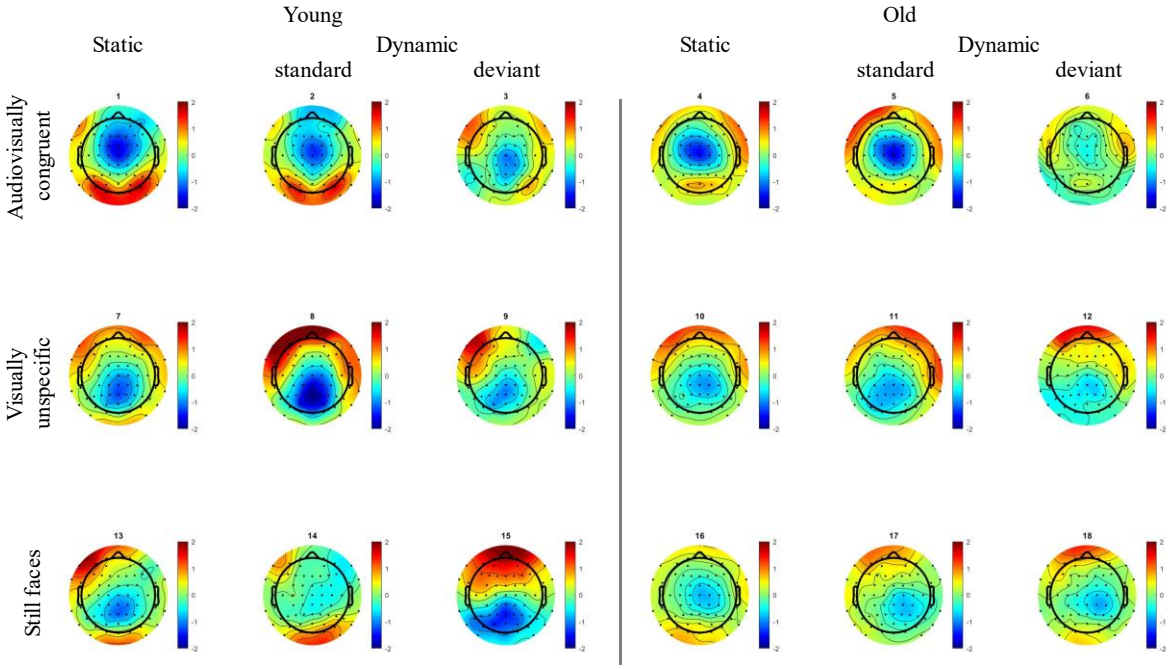
For P2-N2, we found overall smaller peak-to-peak amplitudes in the older compared to the younger group ($M_{old} = 1.68$ µV, $SE_{old} = 0.18$; $M_{young} = 3.30$ µV, $SE_{young} = 0.19$; figure 4c; table 1). We also found a main effect of visual information that was modulated by an interaction with the age group. As revealed by post-hoc analyses, smaller peak-to-peak amplitudes were elicited in the older group for audiovisually congruent, $t(19) = -3.27$, $p_{adj} = .008$, $g = -0.70$, and still faces, $t(19) = -2.36$, $p_{adj} = .044$, $g = -0.51$, compared to visually unspecific stimuli, and no differences were found between audiovisually congruent and still face stimuli, $V = 64$, $p_{adj} = .160$, $r = .34$. In the younger group, however, peak-to-peak amplitudes were smallest when presenting audiovisually congruent information compared to visually unspecific information, $t(21) = 4.85$, $p_{adj} < .001$, $g = -1.00$ and still faces, $t(21) = -4.91$, $p_{adj} < .001$, $g = -1.01$, with no difference between the last two, $t(21) = 0.02$, $p_{adj} = .985$, $g = 0.004$.

CNV.

The CNV, preceding the onset of auditory speech, had a frontocentral topography that was especially pronounced when audiovisually congruent information was presented (figure 5). The ANOVA revealed that the CNV mean amplitude was not modulated by the age group, but by visual information (figure 4d, table 1). That is, larger amplitudes were elicited for audiovisually congruent stimuli (M = -1.31 µV, SE = 0.21) compared to both still faces (M = -0.38 µV, SE = 0.14; V = 104, $p_{adj} < .001$, r = .67), and visually unspecific stimuli (M = -0.35 µV, SE = 0.14; V = 111, $p_{adj} < .001$, r = .66), while the last two did not differ, V = 451, $p_{adj} = 1.00$, $r < .001$.

Figure 5

Topographies of the CNV at 100 to 0 ms



Note. Topographies of the CNV at -100 to 0 ms prior to auditory onset shown separately for the younger and older group, for audiovisually congruent, unspecific, and still visual information, and for the static setting und the standard and deviant trials in the dynamic setting.

3.1.3 Summary: static setting

Overall, the behavioral data of the static setting indicated that younger adults showed a better performance on the task than older adults. Additionally, performance was better with audiovisually congruent information relative to unspecific information and still faces, especially in the older group, which suggests a facilitation of the task when the seen lip movement was congruent to the auditory speech input. Audio-visual information reduced P1-N1 amplitudes relative to unspecific information and still faces in the older group, while this effect was less clear in the younger group. In contrast, reduced P2-N2 amplitudes with audio-visual information relative to unspecific information and still faces were only observed in the younger, but not older group. Finally, in both age groups, the CNV was more pronounced with audio-visual information than with unspecific information and still faces.

3.2 Dynamic setting: Standard vs. deviant stimuli

3.2.1 Behavioral data

Accuracy.

Within the dynamic setting, the rate of correct responses was lower in older, than younger participants (M_{old} = 88.92 %, SE_{old} = 0.90; M_{young} = 92.98 %, SE_{young} = 0.72; figure 2B; table 2). Accuracy was also higher in standard, compared to deviant trials (M_{std} = 95.72 %, SE_{std} = 0.58; M_{dev} = 86.38 %, SE_{dev} = 0.82). In addition, we found an interaction of visual information and target location. Post-hoc comparisons revealed that within standard trials, accuracy was higher for audiovisually congruent (M = 97.42 %, SE = 0.55) than for visually unspecific stimuli (M = 94.68 %, SE = 0.93; V = 438.5, p_{adj} = .010, r = .40) and still faces (M = 95.07 %, SE = 1.34; V = 352.5, p_{adj} = .009, r = .43), with no differences between last two, V = 334, p_{adj} = .994, r = .002. In contrast, within deviant trials, accuracy was lower in audiovisually congruent (M = 82.73 %, SE = 1.66) compared to visually unspecific stimuli (M

= 86.64 %, SE = 1.26; $t(41) = -2.51$, $p_{adj} = .024$, $g = -0.38$) and still faces (M = 89.75 %, SE = 1.13;$t(41) = -5.78$, $p_{adj} =< .001$, $g = -0.88$). Also, accuracy for visually unspecific stimuli was significantly lower than for still faces, $V = 344.5$, $p_{adj} = .028$, $r = .33$.

Response Times.

Response times were overall larger in older than younger participants (M_{old} = 1948.25ms, SE_{old} = 27.40; M_{young} = 1800.26 ms, SE_{young} = 19.12), and in deviant than standard trials (M_{dev} = 2016.63 ms, SE_{dev} = 22.14; M_{std} = 1724.83 ms, SE_{std} = 18.40; figure 2B, table 2). There was also an interaction of visual information and target location. In standard trials, response times in audiovisually congruent stimuli were lower (M = 1630.71 ms, SE = 25.21) compared to visually unspecific (M = 1770.91 ms, SE = 34.61, V = 11, $p_{adj} < .001$, $r = .85$) and still face stimuli (M = 1772.87 ms, SE = 30.67, $t(41) = -9.86$, $p_{adj} < .001$, $g = -1.49$), with no difference between the two latter, $V = 535$, $p_{adj} = .363$, $r = .16$. In contrast, response times in deviant trials were actually higher with audiovisually congruent (M = 2097.45 ms, SE = 37.73), compared to both visually unspecific (M = 1976.68 ms, SE = 41.15, $t(41) = 5.67$, $p_{adj} < .001$, $g = 0.86$) and still face stimuli (M = 1975.75ms, SE = 33.59, $t(41) = 5.74$, $p_{adj} < .001$, $g = 0.87$). Again, the two latter did not differ, $V = 494$, $p_{adj} = .603$, $r = .08$.

Finally, there was a three-way interaction of age, visual information, and target location. Separate analyses for both age groups showed a similar pattern, with a significant interaction of visual information and target location, old: $F(38,2) = 56.03$, $p_{adj} < .001$, η_p^2 = .75, young: $F(42,2) = 59.48$, $p_{adj} < .001$, $\eta_p^2 = .74$. Post-hoc comparisons in the older adults showed smaller response times in standard trials, when comparing audiovisually congruent with visually unspecific information, $t(19) = -4.99$, $p_{adj} < .001$, $g = -1.07$, and still faces, $t(19) = -7.53$, $p_{adj} < .001$, $g = -1.62$. In deviant trials, larger response times followed audiovisually congruent compared to visually unspecific, $t(19) = 3.61$, $p_{adj} = .003$, $g = 0.78$, and still face stimuli, $t(19) = 4.91$, $p_{adj} < .001$, $g = 1.05$. Still face stimuli did not differ from visually

unspecific stimuli in terms of response times in neither standard, $t(19) = -0.13$, $p_{adj} = .900$, $g = -0.03$, nor deviant trials, $t(19) = -1.22$, $p_{adj} = .277$, $g = -0.26$. In the younger adults, a similar pattern could be observed with response times being smaller in audiovisually congruent compared to visually unspecific, $t(21) = -5.54$, $p_{adj} < .001$, $g = -1.14$, and still face stimuli, $t(21) = -7.12$, $p_{adj} < .001$, $g = -1.46$, when shown in standard trials. The last two conditions did not differ, $t(21) = 0.54$, $p_{adj} = .641$, $g = 0.11$. When deviant trials were presented, response times were larger in audiovisually congruent compared to visually unspecific, $t(21) = 4.31$, $p_{adj} < .001$, $g = 0.89$, and still face stimuli, $t(21) = 3.31$, $p_{adj} = .004$, $g = 0.68$. Here, also smaller response times were found in visually unspecific stimuli in comparion to still faces, $t(21) = -2.32$, $p_{adj} = .038$, $g = -0.48$.

3.2.2 ERP data

Looking at the grand-averaged waveforms at frontocentral scalp locations (figure 6), we can observe a similar pattern of activation as found for the static setting (figure 3). Again, prior to the onset of auditory information (0 ms) and thus prior to our components of interest, P1, N1, P2 and N2, we observed a negative, CNV-like deflection, starting about 300 ms, and peaking at approximately 100 ms prior to sound onset.

Table 2

2×3×2 ANOVA results for standard and deviant trials in the dynamic setting

	Accuracy[1]	Response time[2]		P1-N1[1]	N1-P2[1]	P2-N2[1]	CNV[1]
conditions	*F*	*F*	η_p^2	*F*	*F*	*F*	*F*
age	**8.78****	**6.62***	.14	0.36[T]	2.82	**10.21****	.10
visual information	2.40[T]	0.25	.01	**7.43*****	**6.88****	**17.75*****	**12.00*****
target location	**173.54*****	**199.72*****	.83	0.20	0.17	**10.89*****	**17.73*****
age × visual information	0.80	1.40	.03	**3.20***	0.97	**7.46*****	.61
target location × visual information	**17.15*****	**111.40*****	.74	**6.02****	0.90	1.21	2.11
age × target location	0.21	0.04	< .01	2.21	0.25	0.12	.15
age × visual information × target location	1.96	**4.84***	.11	1.22	2.19	0.39	**5.37****

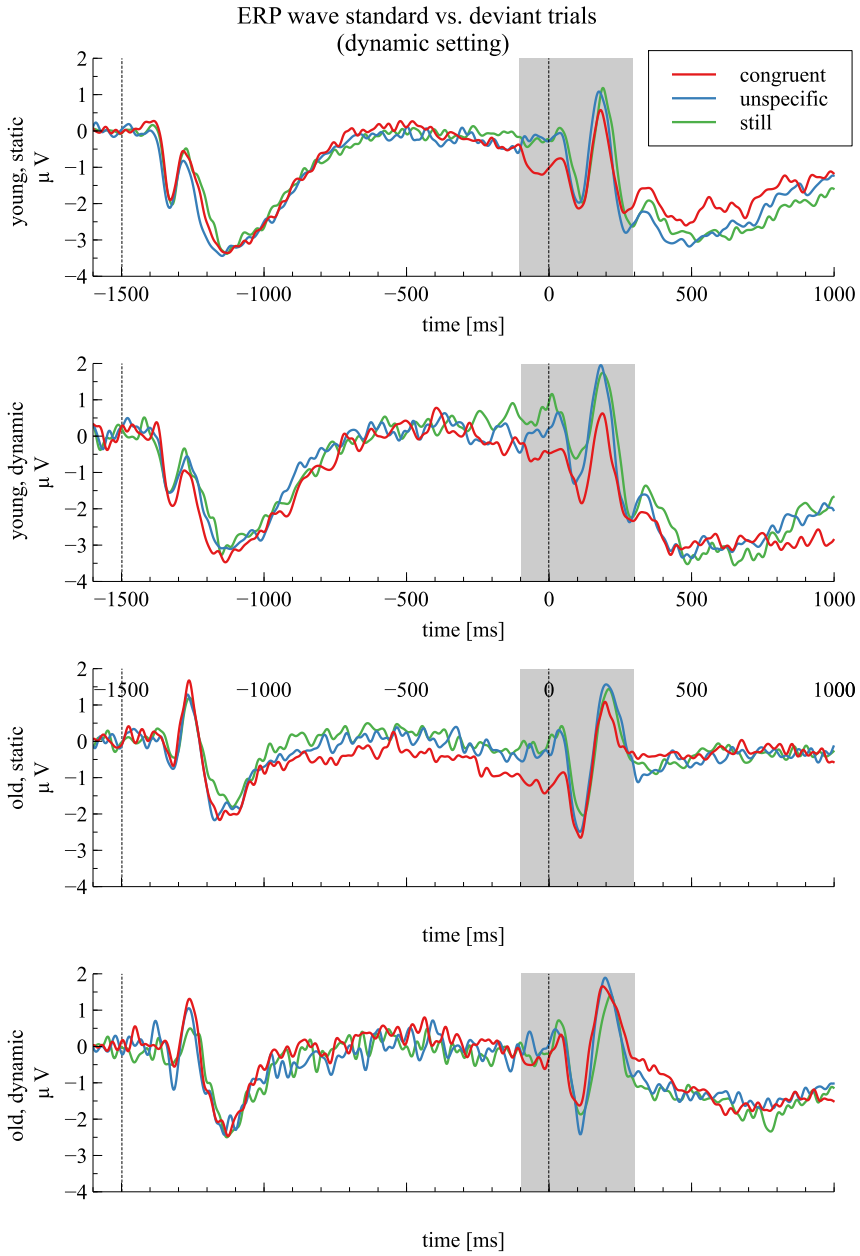
Notes. [1] ANOVA-type statistic, except for whole-plots factors, denominator degrees of freedom are infinite, thus reporting only nominator degrees of freedom. Degrees of freedom: age = 1, 39.23 (accuracy), 1, 38.75 (P1-N1), 1, 39.53 (N1-P2), 1, 38.17 (P2-N2), 1, 40 (CNV); visual information, age × visual information = 1.74 (accuracy), 1.89 (P1-N1), 1.97 (N1-P2), 2.00 (P2-N2), 2.00 (CNV); target location, group × target location = 1; target location × visual information, age × target location × visual information = 1.96 (accuracy), 1.92 (P1-N1), 1.81 (N1-P2), 1.97 (P2-N2), 2.0 (CNV)

[2] Degrees of freedom: Age, target location, age × target location = 1, 40, visual information, age × visual information, visual information × target location, age × visual information × target location = 2, 80.

[T] p <.10, * p < .05, ** p < .01, *** p < .001.

Figure 6

ERP waveforms at electrode cluster Cz, FCz, Fz, FC1 and FC2 in dynamic setting



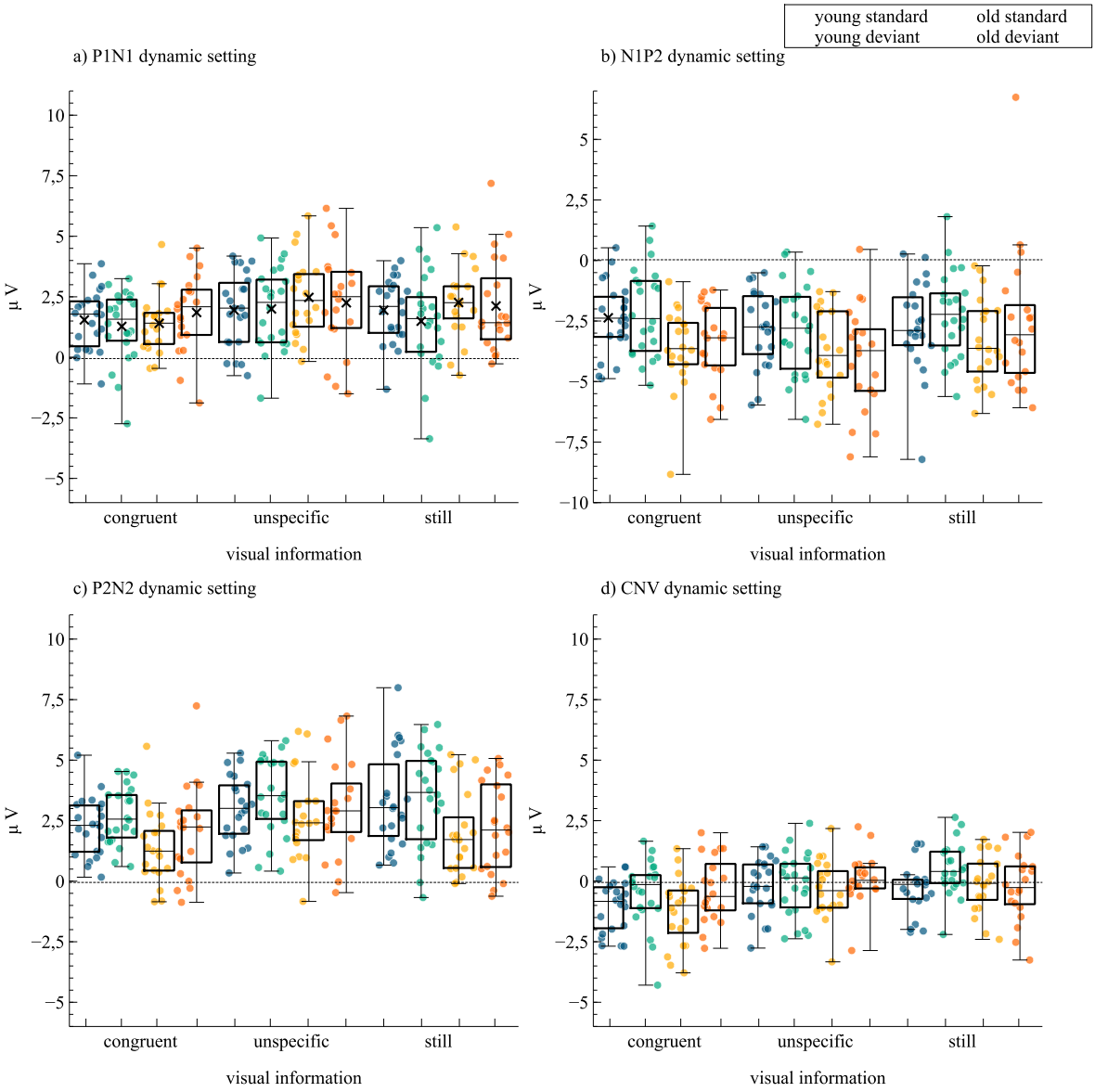
Note. Waveforms are shown for separately for the younger and older group, for audiovisually congruent, unspecific, and still visual information, and for standard and deviant trials. The time interval shows length of one epoch, baseline interval was -1600 to -1500, i.e. prior to stimulus onset, sound onset was at 0 ms. The grey area marks the analyzed part of the ERP wave.

P1-N1.

For the P1-N1 peak-to-peak amplitude, we found a main effect of visual information, as well as interactions of visual information with the age group and the target location (see figure 7a, table 2). Post-hoc analyses revealed that amplitudes with regard to the different conditions of visual information only differed in the older, but not in the younger group (all $p_{adj} > .287$). In the older group, smaller peak-to-peak amplitudes were elicited in stimuli with audiovisually congruent ($M = 1.86$ µV, $SE = 0.19$) compared to visually unspecific ($M = 2.82$ µV, $SE = 0.23$; $t(39) = -6.29$, $p_{adj} < .001$, $g = -0.98$) and still face stimuli ($M = 2.43$ µV, $SE = 0.25$, $V = 193$, $p_{adj} = .009$, $r = .46$), with no difference between the last two conditions, $V = 270$, $p_{adj} = .120$, $r = .30$. Within standard trials, audiovisually congruent stimuli elicited significantly smaller amplitudes ($M = 1.57$ µV, $SE = 0.17$) compared to visually unspecific ($M = 2.27$ µV, $SE = 0.23$; $t(41) = -3.95$, $p_{adj} < .001$, $g = -0.60$) and still face stimuli ($M = 2.22$ µV, $SE = 0.19$; $t(41) = -4.28$, $p_{adj} < .001$, $g = -0.65$), whereas the last two conditions did not differ, $t(41) = 0.31$, $p_{adj} = .760$, $g = 0.05$. In contrast, within deviant trials, amplitudes for audiovisually congruent stimuli ($M = 1.90$ µV, $SE = 0.18$) were smaller than for visually unspecific stimuli ($M = 2.47$ µV, $SE = 0.23$; $t(41) = -2.76$, $p_{adj} = .018$, $g = -0.42$), whereas there were no differences between both audiovisually congruent, $t(41) = -0.397$, $p_{adj} = .760$, $g = -0.06$, and visually unspecific, $V = 306$, $p_{adj} = .105$, $r = .28$, compared to still face stimuli ($M = 2.00$ µV, $SE = 0.26$).

Figure 7

ERP amplitudes in dynamic setting



Note. Mean amplitudes of a) P1-N1, b) N1-P2 and c) P2-N2 absolute difference values as well as d) CNV shown separately for the younger and older group, for audiovisually congruent, unspecific, and still visual information, and for standard and deviant trials. Box plot whiskers mark 1.5 IQR, mean is indicated by the cross.

N1-P2.

The N1-P2 peak-to-peak amplitudes only depended on visual information. That is, audiovisually congruent stimuli (M = 2.88 µV, SE = 0.19) elicited smaller amplitudes in comparison to visually unspecific stimuli (M = 3.39 µV, SE = 0.21; V = 1111, p_{adj} = .009, r = .33), but not to still face stimuli (M = 2.90 µV, SE = 0.19; $t(83)$ = -0.11, p_{adj} = .913, g = -0.01), while amplitudes were larger for stimuli with visually unspecific information compared to still faces, $t(83)$ = 2.82, p_{adj} = 0.009, g = 0.31 (figure 7b, table 2).

P2-N2.

For the P2-N2 peak-to-peak measure, smaller amplitudes were observed in the older than in the younger group (M_{old} = 2.09 µV, SE_{old} = 0.15; M_{young} = 3.32 µV, SE_{young} = 0.14) as well as in standard than in deviant trials (M_{std} = 2.45 µV, SE_{std} = 0.15; M_{dev} = 2.98 µV, SE_{dev} = 0.16; figure 7c, table 2). Moreover, visual information had an influence on the P2-N2 amplitudes, which interacted with the age group. As revealed by post-hoc comparisons, peak-to-peak amplitudes in the older group elicited for visually unspecific stimuli (M = 2.58 µV, SE = 0.26) were larger compared to audiovisually congruent (M = 1.84 µV, SE = 0.23; $t(39)$ = -3.61, p_{adj} = .001, g = -0.56) and still face stimuli (M = 1.84 µV, SE = 0.26; $t(39)$ = -4.37, p_{adj} < .001, g = -0.68), while the two latter did not differ, V = 399, p_{adj} = .889, r = .02. In the younger group, however, we found smaller amplitudes in audiovisually congruent stimuli (M = 2.60 µV, SE = 0.22) compared to still face (M = 3.76 µV, SE = 0.27; $t(43)$ = -5.01, p_{adj} < .001, g = -0.74) and visually unspecific stimuli (M = 3.60 µV, SE = 0.23 $t(43)$ = -4.84, p_{adj} < .001, g = -0.72), while the last two did not differ, $t(43)$ = 0.725, p_{adj} = .472, g = 0.11 (figure 7c; table 2).

CNV.

The frontocentral CNV was most pronounced when audiovisually congruent stimuli were presented at the standard target location, but faded away with visually unspecific stimuli und still faces, and when targets were presented at a deviant location (figure 5). This pattern was more evident in the older group. Accordingly, the ANOVA indicated larger CNV amplitudes in standard than in deviant trials (M_{std} = -0.55 µV, SE_{std} = 0.11; M_{dev} = -0.09 µV, SE_{dev} = 0.11; figures 7d, table 2). Also, there was a main effect of visual information as well as a three-way interaction of age group, visual information, and target location. To further investigate this three-way-interaction, we calculated deviant minus standard difference values (i.e., the decrease in CNV amplitudes in deviant relative to standard trials) and tested effects of visual information on these difference values. With the interaction term of group and visual information still being significant, $F(2, 80) = 5.37$, $p = .007$, $\eta_p^2 = .12$, we conducted post-hoc pairwise comparisons. For the older group, results indicated a larger decrease in CNV amplitude with audiovisually congruent stimuli (M = +0.87 µV, SE = 0.23) compared to still face stimuli (M = -0.08 µV, SE = 0.15; $V = 175$, $p_{adj} = .042$, $r = 0.58$), but not to visually unspecific stimuli (M = +0.44 µV, SE = 0.18; $t(19) = 1.25$, $p_{adj} = .270$, $g = 0.27$). Also, the decrease in CNV did not differ between visually unspecific and still face stimuli, $t(19) = 2.22$, $p_{adj} = .078$, $g = 0.48$. For the younger group no significant differences between visual conditions (all $t < 2.35$; $p > .085$) were found.

3.2.3 Summary: dynamic setting

Accuracy declined and response times increased when the target stimulus was presented from a deviant position, and in both age groups these effects were stronger with audiovisually congruent stimuli than visually unspecific stimuli and still faces. A reduction of P1-N1 amplitudes with audiovisually congruent relative to still faces was only observed in

standard, but not in deviant trials. P2-N2 amplitudes were overall larger in deviant, than standard trials. Also, the CNV was overall smaller in deviant trials, and this reduction in CNV was mainly pronounced in the older group with audio-visually congruent stimuli relative to visually unspecific stimuli and still faces. Irrespective of the target location, the audio-visually congruent information reduced P1-N1 amplitudes (mainly in the older group) and P2-N2 amplitudes (mainly in the younger group) relative to unspecific information and still faces. The N1-P2 amplitudes, however, were increased with visually unspecific stimuli, relative to audio-visually congruent information and still faces.

4. Discussion

The present study investigated the processing of simultaneously presented audiovisual speech in a multi-talker scenario, comparing the neuro-cognitive mechanisms of audio-visual speech integration in younger and older adults. We varied the informational content of the visual input by presenting audiovisually congruent, visually unspecific, and still face speech stimuli. Moreover, we tested, whether expected benefits of audiovisually congruent speech information (relative to incongruent and still face, that is auditory-only information) also occur under highly dynamic listening conditions. We therefore used either a static setting in which the talker of the target information was kept constant, or a dynamic setting in which the target talker occasionally switched to a deviant (lateral) position.

In both age groups, audiovisually congruent information proved to be beneficial for a faster and more accurate performance, relative to audio-only information (i.e., the still – face condition), while visually unspecific information was not particularly helpful. Overall, performance was declined in older adults, with lower accuracy rates and higher response times. Moreover, the static setting was superior to the dynamic one, in which the audiovisual

benefit vanished when the target information was not presented at the expected standard location.

These behavioral results are reflected in the examined ERPs as well. Audiovisually congruent speech information reduced amplitudes in early to late components, while also here, we did not observe a clear benefit of visually unspecific information. Age-related amplitude modulations were found, with enhanced amplitudes in older adults in early processing stages and reduced amplitudes in later stages. The occurrence of a deviant target location seemed to be mostly relevant in later components, with amplitudes being enhanced in deviant trials. In the following, specific results regarding audiovisual speech processing, dynamic switches in talker location and age-related changes are discussed in detail.

4.1 Audiovisual speech comprehension in static and dynamic settings

Looking at the static setting, in which no switches in target location occurred, we observed a benefit of audiovisual information, which was being reflected both behaviorally and electrophysiologically. There were faster and more accurate responses with audiovisually congruent stimuli compared to visually unspecific or still face stimuli. In terms of accuracy, the benefit of congruent in contrast to unspecific information was especially visible in older, but not younger adults. On the one hand, this could be due to ceiling effects in the younger group, with error rates of less than 3 %. On the other hand, it has been shown that younger adults are less biased by ambiguous visual information (Cienkowski & Carney, 2002), while older adults need high quality visual input in order to benefit from it (Tye-Murray et al., 2011). Generally, our study is in line with the notion that audiovisually congruent information supports effective speech comprehension by integrating coherent and redundant information from both modalities (Besle et al., 2004; Bronkhorst, 2015; Campbell, 2008; Lindström, 2012; van Wassenhove et al., 2005). In addition, Stekelenburg and Vroomen (2007) argued that enhancement due to successful integration does not primarily depend on the congruency

of both modalities, but on the predictive value of the visual input. That is, additional visual input only leads to a decreased auditory threshold (Grant & Seitz, 2000) and enhanced intelligibility (Schwartz et al., 2004) when it allows for a valid prediction of the subsequent auditory information. Consistently, our visually unspecific condition demonstrates that preceding lip movements which might serve as a cue to the onset of the following speech sound, but which are unrelated to the auditory information, do not seem to be sufficient to result in behavioral benefits. In conclusion, not only the timing, but also the content is important for a facilitation in speech comprehension.

When presented within the dynamic setting, a reversion of the benefit of audiovisually congruent information became visible as soon as an unexpected change in target talker location occurred. Previous research in the auditory modality has already demonstrated that switches in target talker location deteriorate speech comprehension in cocktail-party situations (Best et al., 2008). This has been attributed to the re-analysis of the auditory scenary and the re-focussing of auditory attention after a switch has occurred (Getzmann et al., 2015; Koch et al., 2011; Lin & Carlile, 2015). Additionally, when presented with the audiovisually congruent stimuli in the present study, the fixated faces at the (central) standard location became temporarily incongruent in case of a location switch: Instead of the expected visual speech information (the target words “yes” or “no”), a deviant visual input occurred not matching the expected subsequent auditory input. Incongruent audiovisual input has been shown to result in increased response times compared to congruent stimuli (e.g., Altieri, Lentz, Townsend, & Wenger, 2016). This obvious disadvantage of audiovisual speech in dynamic multi-talker scenarios corresponds to results of a previous study, in which speech recognition in a multi-talker context with changing talkers was slower in an audio-visual condition compared to an audio-only condition (Heald & Nusbaum, 2014). The authors attributed this effect to an enhanced talker variability when presented with audiovisual stimuli.

The deviance-related increase in response times with audiovisual information compared to still faces or unspecific lip movement was more pronounced in older than in younger adults, suggesting that they suffered more from the sudden switch in target talker location than the younger group. In line with this, previous research has shown that older adults experience higher conflict costs in multi-talker scenarios, resulting in less accurate responses, when the attentional load is high (Passow et al., 2012), and higher response times following distraction due to unpredictable sound modulations (Volosin et al., 2017) or deviant target positions (Getzmann et al., 2015).

4.2 ERPs to audiovisual speech onset

The above-mentioned behavioral patterns are reflected in the ERP measures. In the static setting, the benefit due to audiovisually congruent information in comparison to visually unspecific information and still faces was associated with a reduction in the early P1-N1 complex. This observation is in line with past research and has been interpreted as an indicator of facilitated speech processing at an early perceptual level, where both modalities of speech information are integrated (Baart, 2016; Ganesh et al., 2014; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005). Moreover, this early integration of both modalities could enable a more effective pre-attentive filtering of sensory information (e.g., Lebib et al., 2003).

Interestingly, reduced P1-N1 amplitudes to audiovisually congruent stimuli relative to visually unspecific stimuli and still faces were only observed in the older group, but not in the younger group. This is in line with a previous EEG study in which larger reductions in P1 amplitude in older compared to younger participants were found when being presented with audiovisual versus auditory-only speech (Winneke & Phillips, 2011). In contrast to the static setting, P1-N1 amplitudes to audiovisually congruent stimuli were not reduced in comparison to still faces when a switch in target location, that is a deviant, occurred in the dynamic

setting. Here, we observed only a reduced amplitude for audiovisually congruent in comparison to visually unspecific stimuli, with no differences between still faces and both audiovisual conditions. Yet, the typical P1-N1 reduction associated with audiovisually congruent stimuli (compared to still faces and visually unspecific information) was still present in standard-dynamic trials (i.e., when no switch occurred in the dynamic setting).Thus, this suggests that although a benefit due to audiovisual information was not present in switch trials, our participants still seemed to process both audiovisual stimuli differently. A study by Lebib and colleagues (2003) revealed larger P1 amplitudes for incongruent audiovisual stimuli when the input from both modalities was harder to distinguish. This was interpreted as an indicator for early detection of non-redundant audiovisual information, with the lip movement being processed even before vowel sound onset. According to the “analysis by synthesis” approach by Wassenhove and colleagues (2005) audiovisual integration is suggested to take place as soon as the N1, that is, on a perceptual level. Several studies could further show that binding processes can be observed at the time interval of the N1 peak, mainly being driven by the anticipation due to the preceding visual speech input (Ganesh et al., 2014; Stekelenburg & Vroomen, 2007).

Later in the processing cascade, the N1-P2 amplitudes were larger with visually unspecific compared to still face stimuli, in the static setting only for the older group, in the dynamic setting for both groups. This may be due to a higher demand in attention. Since the visually unspecific speech input did not contain any relevant information, an active suppression of the visual content may have led to this result. No differences were found between audiovisually congruent stimuli and still faces, however. This is in contrast to previous studies, showing amplitude reductions in both the N1 and P2 amplitude when presenting an audiovisual stimulus, making the components an indicator for successful audiovisual integration (e.g., Baart, 2016; Baart et al., 2014; Ganesh et al., 2014; Klucharev et al., 2003), although it has been shown that a higher attentional load may disrupt effective

audiovisual integration (Alsius et al., 2014). Taken together, the results from both P1-N1 and N1-P2 clearly demonstrated the difference between congruent, informationally redundant and unspecific, informationally irrelevant visual input in the context of audiovisual speech perception. Despite its temporal precedence relative to the auditory signal, unspecific visual information did not provide valid or helpful information, but rather seemed to demand more cognitive resources to process.

Finally, in the static and dynamic setting, a reduced P2-N2 amplitude was observed with audiovisually congruent relative to unspecific stimuli (and in the younger group also relative to still faces). One possible explanation may be that audiovisually congruent information results in a reduction of the required cognitive control for the processing of audio-visual language in a mixture of speech stimuli. However, independent of the presented visual input, we found an increase in P2-N2 amplitude in the dynamic setting when the target stimulus came from the deviant position. This suggests that a higher amount of cognitive control was needed when the target location changed. Similar results have been shown for non-speech auditory stimuli, where N2 deflections were elicited when presenting preceding visually incongruent information (Lindström, 2012).

4.3 Preparatory ERP activity preceding auditory speech onset

In addition to the above measures, we also explored the pronounced CNV right before sound onset. Interestingly, our results showed stronger CNV amplitudes for audiovisually congruent information than for visually unspecific or still face stimuli. This suggests more pronounced preparatory processes, when the visual information was known to be relevant. In this regard, it should be noted that we chose a blocked design in our experiment. Thus, participants would know that the lip movement contained valid information as opposed to the still faces and simple mouth opening and closing movements in the other two conditions. The CNV is typically elicited, when a preceding stimulus reliably predicts a subsequent stimulus,

containing task-relevant information (Brunia, van Boxtel, & Böcker, 2011; Walter et al., 1964). Anticipatory slow-waves such as the CNV in the context of audiovisual integration have been sparsely studied so far, although their occurrence is known (Teder-Salejarvi et al., 2002). Our findings are in line with the notion of the CNV as an expectancy and anticipation-related component. In particular, the more pronounced CNV amplitude for audiovisually congruent stimuli might reflect a heightened attentional focusing towards the visual speech input and might also be an indication of successful integration as soon as the target sound would be presented.

In the dynamic setting, the CNV was decreased in deviant compared to standard trials, when presented with audiovisually congruent stimuli. This is reasonable, given that only the lip movement in the congruent condition could be a cue to a change in target location because then, the lip movement would not match the expectancy anymore. In contrast, in case of still faces and unspecific lip movement, the visual information would not be an indicator of a target switch. Thus, the deviance-induced reduction in CNV amplitude suggests an anticipation of the preceding switch in target talker location well before the auditory speech onset. Notably, this effect was pronounced only in the older group. This contrasts the predominant finding that preparatory attentional processes in general seem to decline with age, being reflected in diminished CNV amplitudes (for overview, see Wild-Wall, Hohnsbein, & Falkenstein, 2007), especially when a widened focus of attention was needed, that is, in a divided opposed to a selective attention task (Wild-Wall & Falkenstein, 2010). Accordingly, a reduced CNV of older participants has primarily been observed when attention had to be divided between talkers in a cocktail-party situation (Getzmann et al., 2016). Here, in contrast, the well-pronounced CNV to audio-visually congruent speech in the present task as well as the decrease in CNV in deviant trials suggest intact preparatory attention and gating of subsequent task-related auditory speech information in the older group of participants.

4.4 Age effects

We observed general age effects in both behavioral and neuronal measures, demonstrating well-assessed age-related decline in performance for older adults in both cocktail-party scenarios (Getzmann & Wascher, 2016) and (audiovisual) speech processing (Sekiyama et al., 2014; Winneke & Phillips, 2011). In P1-N1, or early stages of processing, we found larger amplitudes for the older group with simultaneously decreased P2-N2 amplitudes. The increase in P1 and decrease in N2 has repeatedly been associated with an age-related deficit of inhibitory processing of irrelevant information in line with the Inhibitory Deficit Hypothesis, suggesting an age-related deficit in inhibitory processing in early stages (Friedman, 2011; Getzmann et al., 2015; Stothart & Kazanina, 2016). Opposed to research showing an age-related modulation of N1 an P2 amplitude (see Friedman, 2011), we could not assess a clear age-effect on N1-P2. Unfortunately, using peak-to-peak measures due to preparatory activity before sound onset (CNV) makes it impossible to show specific modulations on the single ERPs. While P2 amplitudes tend to be enhanced due to age, N1 modulations have been ambiguous throughout studies (for review see Friedman, 2011). In fact, it may be possible that age-related enhancement or decrease in both amplitudes may have been cancelled out in the peak-to-peak analysis.

4.5 Conclusion

To summarize, the ERP results from our study indicate that congruent visual speech information is actively used in preparation for the subsequent auditory input (CNV), enables early audiovisual integration (P1-N1), and a facilitation of cognitive control processes in the further processing of multi-talker speech (P2-N2), leading to a better speech perception as reflected by the enhanced performance (i.e., accuracy and response times). However, these benefits are only visible as long as the multi-talker scenario remains stable. In contrast, unexpected switches in target talker location lead to an audiovisual incongruency, cancelling

the benefits of successfully integrated congruent audiovisual speech. This was observable in amplitude modulations in the P1-N1, P2-N2, and CNV measures and ultimately also in deteriorated performances measures. Although we found age-related changes in the processing of the speech stimuli, probably due to inhibitory deficits and worsened resource allocation, we still found that older adults were able to profit from congruent audiovisual information. The benefits were visible especially in early phases of processing (P1-N1), indicating a more effective pre-attentive filtering of sensory information, when presented with relevant additional visual input.

References

Alain, C., McDonald, K. L., Ostroff, J. M., & Schneider, B. (2004). Aging: A Switch From Automatic to Controlled Processing of Sounds? *Psychology and Aging*, *19*(1), 125–133. https://doi.org/10.1037/0882-7974.19.1.125

Alain, C., & Tremblay, K. (2007). The Role of Event-Related Brain Potentials in Assessing Central Auditory Processing. *Journal of the American Academy of Audiology*, *18*(07), 573–589. https://doi.org/10.3766/jaaa.18.7.5

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), 1–9. https://doi.org/10.3389/fpsyg.2014.00727

Altieri, N., Lentz, J. J., Townsend, J. T., & Wenger, M. J. (2016). The McGurk effect: An investigation of attentional capacity employing response times. *Attention, Perception, & Psychophysics*, *78*(6), 1712–1727. https://doi.org/10.3758/s13414-016-1133-4

Baart, M. (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. *Psychophysiology*, *53*(9), 1295–1306. https://doi.org/10.1111/psyp.12683

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*(1), 115–121. https://doi.org/10.1016/j.neuropsychologia.2013.11.011

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*(8), 2225–2234. https://doi.org/10.1111/j.1460-9568.2004.03670.x

Best, V., Ozmeral, E. J., Kopco, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences*, *105*(35), 13174–13178. https://doi.org/10.1073/pnas.0803718105

Brault, L. M., Gilbert, J. L., Lansing, C. R., McCarley, J. S., & Kramer, A. F. (2010). Bimodal Stimulus Presentation and Expanded Auditory Bandwidth Improve Older Adults' Speech Perception. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(4), 479–491. https://doi.org/10.1177/0018720810380404

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, and Psychophysics*, *77*(5), 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Brunia, C. H. M., van Boxtel, G. J. M., & Böcker, K. B. E. (2011). Negative Slow Waves as Indices of Anticipation: The Bereitschaftspotential, the Contingent Negative Variation, and the Stimulus-Preceding Negativity. In E. S. Kappenman & S. J. Luck (Eds.), *The Oxford Handbook of Event-Related Potential Components* (pp. 1–32). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0108

Burke, D. M., & Shafto, M. A. (2015). Language and Aging. In *The Handbook of Aging and Cognition* (pp. 373–443). Routledge. https://doi.org/10.4324/9780203837665.ch8

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001–1010. https://doi.org/10.1098/rstb.2007.2155

Carlile, S. (2014). Active listening: Speech intelligibility in noisy environments. *Acoustics Australia*,

42(2), 90–96.

Carson, N., Leach, L., & Murphy, K. J. (2018). A re-examination of Montreal Cognitive Assessment (MoCA) cutoff scores. *International Journal of Geriatric Psychiatry*, *33*(2), 379–388. https://doi.org/10.1002/gps.4756

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*.

Cienkowski, K. M., & Carney, A. E. (2002). Auditory-Visual Speech Perception and Aging. *Ear and Hearing*, *23*(5), 439–449. https://doi.org/10.1097/00003446-200210000-00006

Cohen, J. (1973). Eta-Squared and Partial Eta-Squared in Fixed Factor Anova Designs. *Educational and Psychological Measurement*, *33*(1), 107–112. https://doi.org/10.1177/001316447303300111

Correa-Jaraba, K. S., Cid-Fernández, S., Lindín, M., & Díaz, F. (2016). Involuntary Capture and Voluntary Reorienting of Attention Decline in Middle-Aged and Old Participants. *Frontiers in Human Neuroscience*, *10*(129), 1–13. https://doi.org/10.3389/fnhum.2016.00129

Dai, L., Best, V., & Shinn-Cunningham, B. G. (2018). Sensorineural hearing loss degrades behavioral and physiological measures of human spatial selective auditory attention. *Proceedings of the National Academy of Sciences*, *115*(14), E3286–E3295. https://doi.org/10.1073/pnas.1721226115

Davis, S. W., Dennis, N. A., Daselaar, S. M., Fleck, M. S., & Cabeza, R. (2008). Qué PASA? the posterior-anterior shift in aging. *Cerebral Cortex*, *18*(5), 1201–1209. https://doi.org/10.1093/cercor/bhm155

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles: Sage.

Friedman, D. (2011). The Components of Aging. In E. S. Kappenman & S. J. Luck (Eds.), *The Oxford Handbook of Event-Related Potential Components* (Vol. 1, pp. 1–42). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0243

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. https://doi.org/10.1037/a0024338

Ganesh, A. C., Berthommier, F., Vilain, C., Sato, M., & Schwartz, J.-L. (2014). A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. *Frontiers in Psychology*, *5*(1340), 1–13. https://doi.org/10.3389/fpsyg.2014.01340

Getzmann, S., Falkenstein, M., & Wascher, E. (2015). ERP correlates of auditory goal-directed behavior of younger and older adults in a dynamic speech perception task. *Behavioural Brain Research*, *278*, 435–445. https://doi.org/10.1016/j.bbr.2014.10.026

Getzmann, S., Golob, E. J., & Wascher, E. (2016). Focused and divided attention in a simulated cocktail-party situation: ERP evidence from younger and older adults. *Neurobiology of Aging*, *41*, 138–149. https://doi.org/10.1016/j.neurobiolaging.2016.02.018

Getzmann, S., & Wascher, E. (2016). Postdeviance distraction in younger and older adults: Neuro-behavioral evidence from speech perception. *Psychology and Aging*, *31*(8), 943–957. https://doi.org/10.1037/pag0000121

Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197. https://doi.org/10.1121/1.1288668

Heald, S. L. M., & Nusbaum, H. C. (2014). Talker variability in audio-visual speech perception. *Frontiers in Psychology*, *5*(698), 1–9. https://doi.org/10.3389/fpsyg.2014.00698

Hill, K. T., & Miller, L. M. (2010). Auditory Attentional Control and Selection during Cocktail Party Listening. *Cerebral Cortex*, *20*(3), 583–590. https://doi.org/10.1093/cercor/bhp124

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a Cocktail Party. *Psychological Science*, *24*(10), 1995–2004. https://doi.org/10.1177/0956797613482467

Kidd, G., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America*, *118*(6), 3804–3815. https://doi.org/10.1121/1.2109187

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research*, *18*(1), 65–75. https://doi.org/10.1016/j.cogbrainres.2003.09.004

Koch, I., Lawo, V., Fels, J., & Vorländer, M. (2011). Switching in the cocktail party: Exploring intentional control of auditory selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1140–1147. https://doi.org/10.1037/a0022189

Kropotov, J., Ponomarev, V., Tereshchenko, E. P., Müller, A., & Jäncke, L. (2016). Effect of Aging on ERP Components of Cognitive Control. *Frontiers in Aging Neuroscience*, *8*(69), 1–15. https://doi.org/10.3389/fnagi.2016.00069

Lawo, V., Fels, J., Oberem, J., & Koch, I. (2014). Intentional attention switching in dichotic listening: Exploring the efficiency of nonspatial and spatial selection. *Quarterly Journal of Experimental Psychology*, *67*(10), 2010–2024. https://doi.org/10.1080/17470218.2014.898079

Lebib, R., Papo, D., de Bode, S., & Baudonnière, P.-M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neuroscience Letters*, *341*(3), 185–188. https://doi.org/10.1016/S0304-3940(03)00131-9

Lin, G., & Carlile, S. (2015). Costs of switching auditory spatial attention in following conversational turn-taking. *Frontiers in Neuroscience*, *9*(124), 1–11. https://doi.org/10.3389/fnins.2015.00124

Lindström, R. (2012). Processing of audiovisual associations in the human brain: dependency on expectations and rule complexity. *Frontiers in Psychology*, *3*(159), 1–6. https://doi.org/10.3389/fpsyg.2012.00159

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*(213), 1–14. https://doi.org/10.3389/fnhum.2014.00213

Mäntyjärvi, M., & Laitinen, T. (2001). Normal values for the Pelli-Robson contrast sensitivity test. *Journal of Cataract & Refractive Surgery*, *27*(2), 261–266. https://doi.org/10.1016/S0886-3350(00)00562-9

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, *48*(2), 229–240. https://doi.org/10.1111/j.1469-8986.2010.01061.x

Nasreddine, Z. S., Phillips, N. A., BÃ©dirian, V., Charbonneau, S., Whitehead, V., Collin, I., … Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD : An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments. *Journal of Statistical*

Software, *50*(12), 1–23. https://doi.org/10.18637/jss.v050.i12

Owsley, C. (2011). Aging and vision. *Vision Research*, *51*(13), 1610–1622. https://doi.org/10.1016/j.visres.2010.10.020

Passow, S., Westerhausen, R., Wartenburger, I., Hugdahl, K., Heekeren, H. R., Lindenberger, U., & Li, S.-C. (2012). Human aging compromises attentional control of auditory perception. *Psychology and Aging*, *27*(1), 99–105. https://doi.org/10.1037/a0025667

Peelle, J. E., & Wingfield, A. (2016). The Neural Consequences of Age-Related Hearing Loss. *Trends in Neurosciences*, *39*(7), 486–497. https://doi.org/10.1016/j.tins.2016.05.001

Pelli, D. G., Robson, J. G., & Wilkins, A. J. (1988). The design of a new letter chart for measuring contrast sensitivity. *Clin Vis Sci*, *2*(3), 187–199. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.321.5971

Ritter, W., Simson, R., Vaughan, H., & Friedman, D. (1979). A brain event related to the making of a sensory discrimination. *Science*, *203*(4387), 1358–1361. https://doi.org/10.1126/science.424760

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69–B78. https://doi.org/10.1016/j.cognition.2004.01.006

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology*, *5*(323), 1–12. https://doi.org/10.3389/fpsyg.2014.00323

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing*, *26*(3), 263–275. https://doi.org/10.1097/00003446-200506000-00003

Stekelenburg, J. J., & Vroomen, J. (2007). Neural Correlates of Multisensory Integration of Ecologically Valid Audiovisual Events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–1973.

Stothart, G., & Kazanina, N. (2016). Auditory perception in the aging brain: the role of inhibition and facilitation in early processing. *Neurobiology of Aging*, *47*, 23–34. https://doi.org/10.1016/j.neurobiolaging.2016.06.022

Teder-Salejarvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research*, *14*, 106–114.

Tewes, U. (1991). *Hamburg-Wechsler-Intelligenztest für Erwachsene. Revision 1991. HAWIE-R*. Bern: Huber.

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging*, *31*(4), 380–389. https://doi.org/10.1037/pag0000094

Tye-Murray, N., Spehar, B., Myerson, J., Sommers, M. S., & Hale, S. (2011). Cross-Modal Enhancement of Speech Detection in Young and Older Adults: Does Signal Content Matter? *Ear and Hearing*, *32*(5), 650–655. https://doi.org/10.1097/AUD.0b013e31821a4578

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Volosin, M., Gaál, Z. A., & Horváth, J. (2017). Task-optimal auditory attention set restored as fast in older as in younger adults after distraction. *Biological Psychology*, *126*, 71–81. https://doi.org/10.1016/j.biopsycho.2017.04.007

Wagener, K., Kühnel, V., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die

deutsche Sprache I: Design des Oldenburger Satztests. *Zeitschrift Für Audiologie*, *38*(1), 1–32.

Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., & Winter, A. L. (1964). Contingent Negative Variation : An Electric Sign of Sensori-Motor Association and Expectancy in the Human Brain. *Nature*, *203*(4943), 380–384. https://doi.org/10.1038/203380a0

Wilcoxon, F. (1945). Individual Comparisons of Grouped Data by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.1093/jee/39.2.269

Wild-Wall, N., & Falkenstein, M. (2010). Age-dependent impairment of auditory processing under spatially focused and divided attention: An electrophysiological study. *Biological Psychology*, *83*(1), 27–36. https://doi.org/10.1016/j.biopsycho.2009.09.011

Wild-Wall, N., Hohnsbein, J., & Falkenstein, M. (2007). Effects of ageing on cognitive task preparation as reflected by event-related potentials. *Clinical Neurophysiology*, *118*(3), 558–569. https://doi.org/10.1016/j.clinph.2006.09.005

Winneke, A. H., & Phillips, N. A. (2011). Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. *Psychology and Aging*, *26*(2), 427–438. https://doi.org/10.1037/a0021683

DECLARATIONS

Contributions

Stephan Getzmann, Alexandra Begau, Edmund Wascher and Daniel Schneider designed the study, Alexandra Begau collected the data and conducted the data analysis supported by the helpful recommendations and discussions by and with Laura Klatt. All authors intensively discussed the results and their interpretations, the manuscript was drafted by Alexandra Begau, and revised by all authors.

Conflicts of interest/Competing interests

All authors disclose no actual or potential conflicts of interest including any financial, personal, or other relationships with other people or organizations that could inappropriately influence (bias) their work.

Funding

This work was supported by a grant from the Deutsche Forschungsgemeinschaft (GE 1920/4-1).

Acknowledgements

The authors are grateful to Ludger Blanke, Christele Motcho and Nina Abich for preparing the software and parts of the electronic equipment and to Stefan Weber, Kristin Koberzin, Vera Birgel and Denise Böhle for their help in running the experiment. Further, they would like to thank Nina Abich, Michaela Djudjaj, Kimberly Freytag and Bianca Zickerick for allowing to use their faces and voices as stimulus material.