Visual Cues Contribute Differentially to Audiovisual Perception of Consonants and Vowels in Improving Recognition and Reducing Cognitive Demands in Listeners With Hearing Impairment Using Hearing Aids

Shahram Moradi, a Björn Lidestam, b Henrik Danielsson, a Elaine Hoi Ning Ng, a and Jerker Rönnberg a

Purpose: We sought to examine the contribution of visual cues in audiovisual identification of consonants and vowels—in terms of isolation points (the shortest time required for correct identification of a speech stimulus), accuracy, and cognitive demands—in listeners with hearing impairment using hearing aids.

Method: The study comprised 199 participants with hearing impairment (mean age = 61.1 years) with bilateral, symmetrical, mild-to-severe sensorineural hearing loss. Gated Swedish consonants and vowels were presented aurally and audiovisually to participants. Linear amplification was adjusted for each participant to assure audibility. The reading span test was used to measure participants' working memory capacity.

Results: Audiovisual presentation resulted in shortened isolation points and improved accuracy for consonants and vowels relative to auditory-only presentation. This benefit was more evident for consonants than vowels. In addition, correlations and subsequent analyses revealed that listeners with higher scores on the reading span test identified both consonants and vowels earlier in auditory-only presentation, but only vowels (not consonants) in audiovisual presentation.

Conclusion: Consonants and vowels differed in terms of the benefits afforded from their associative visual cues, as indicated by the degree of audiovisual benefit and reduction in cognitive demands linked to the identification of consonants and vowels presented audiovisually.
memory capacity (WMC) plays a critical role in successful phoneme identification. Independent studies have shown that working memory for inference making, which in this context involves the perceptual completion of ambiguous sounds, is critical for accurate phoneme identification, especially when the speech signal is not clear or when it is masked by external noise or hearing loss.

Hearing loss has been shown to adversely affect the auditory identification of both consonants (Walden & Montgomery, 1975) and vowels (Arehart, Rossi-Katz, & Swenson-Prutsman, 2005; Nábělek, Czyzewski, & Krishnan, 1992). Studies on consonant identification have shown that although the amplification of sounds with hearing aids improves the identification of consonants compared with unaided conditions (Walden, Grant, & Cord, 2001; Woods et al., 2015), hearing-aid users still show inferior performance compared with their counterparts with typical hearing in the auditory identification of consonants and vowels (Moradi, Lidestam, Hälgren, & Rönnberg, 2014; Walden et al., 2001).

To the best of our knowledge, no study has compared vowel identification under aided and unaided conditions in the same study. However, Bor, Souza, and Wright (2008) reported that providing audibility by nonlinear amplification (multichannel compression) was not a significant factor affecting vowel identification in listeners with hearing impairment, and other factors (i.e., cognitive functions) might contribute to vowel recognition in people with hearing loss. Later, Souza, Wright, and Bor (2012) compared linear amplification versus multichannel compression in vowel recognition, and showed that linear amplification was better; nevertheless, neither of these two amplification settings was able to fully compensate for difficulties in vowel recognition up to the same level as a control group with typical hearing.

To compensate for this lack of clarity of the speech signal, listeners with hearing impairment using hearing aids need to use explicit cognitive resources (e.g., working memory) to disambiguate ambiguous sounds into meaningful phonemes (Davies-Venn & Souza, 2014; Moradi, Lidestam, Saremee, & Rönnberg, 2014). According to the Ease of Language Understanding model (the ELU model; Rönnberg et al., 2013; Rönnberg, Rudner, Foo, & Lunner, 2008), there is collaboration between auditory and cognitive systems in language understanding, and working memory acts as a gateway for speech signals on their way to phonological representation in long-term memory. When speech stimuli are presented in optimum listening conditions (e.g., to young listeners with typical hearing, and with a favorable speech presentation level or signal-to-noise ratio [SNR]), mapping an incoming speech signal with its corresponding phonological representations places less demand on working memory to process the clearly audible speech signal. In such cases, the processing of a speech signal is presumably rapid, automatic, and without cognitive demand. However, the receipt of audible speech signals that are less clear, due to noise or hearing loss, places higher demand on working memory to discriminate phonologically similar phonemes from each other. In such cases, perceiving speech stimuli becomes cognitively demanding, and listeners rely on working memory for inference making, which in this context refers to the perceptual completion of ambiguous sounds as phonemes. Independent studies have shown that working memory capacity (WMC) plays a critical role in successful listening comprehension under degraded listening conditions, especially for listeners with hearing impairment (Foo, Rudner, Rönnberg, & Lunner, 2007; Gordon-Salant & Cole, 2016; Lunner, 2003; Souza & Arehart, 2015), supporting the theoretical framework of the ELU model.

Moradi, Lidestam, Hälgren, and Rönnberg (2014) showed that despite using advanced digital hearing aids, older adults using hearing aids had inferior performance compared with older adults with typical hearing in auditory identification of consonants. The hearing-aid users needed longer isolation points (IPs, the shortest time from the onset of a speech signal required for correct identification) and had lower accuracy (in terms of correct identification) than their age-matched counterparts with typical hearing in the identification of Swedish consonants. In addition, the researchers also showed that hearing-aid users with greater WMC had quicker and more accurate consonant identification than those with fewer explicit cognitive resources. Davies-Venn and Souza (2014) similarly reported that working memory can modulate the adverse consequences of distortion caused by compression amplification in consonant identification for people with moderate to severe hearing loss. Using high-frequency amplification (to increase audibility), Ahlstrom, Horwitz, and Dubno (2014) attempted to improve consonant identification in people with hearing loss. However, they reported that the benefits provided by high-frequency amplification in the identification of consonants were relatively limited and varied among listeners with hearing impairment. They suggested that other factors beyond simple audibility, such as individual differences in cognitive capacity, may have partly influenced consonant identification in their participants.

Face-to-face conversation, which typically occurs in an audiovisual modality, enables the listener to view the talker’s accompanying facial gestures. These facial gestures provide supplementary information about the identity of the speech signal that is not available in an auditory-only modality, such as temporal features (e.g., amplitude envelope, onset, and offset) and content (e.g., manner and place of articulation, which limit the number of lexical neighborhoods and resolve syllabic and lexical ambiguity; for a review, see Peelle & Sommers, 2015). In addition, visual cues direct the attention of a listener to the target talker in a “cocktail party” condition, facilitating auditory-stream segregation, and can increase the certainty of a listener’s prediction about the identity of a forthcoming speech signal. As a consequence, these supportive visual cues facilitate the perception of speech stimuli in terms of accuracy and IP, particularly in degraded listening conditions caused by external noise or hearing loss (Moradi, Lidestam, & Rönnberg, 2013, 2016). Audiovisual presentation of speech stimuli is particularly important for people with hearing difficulties (Desai, Stickney, & Zeng, 2008; Walden, Montgomery, Prosek, & Hawkins, 1990), because they rely more on visual speech cues than do listeners with typical hearing when both auditory and visual speech cues are available to disambiguate the identity of a target speech signal. In a recent study, Moradi et al. (2016) showed that...
the degree of audiovisual benefit provided by the association of visual cues with auditory speech stimuli was greater in older adults using hearing aids than in age-matched counterparts with typical hearing in the audiovisual identification of speech stimuli. In that study, both the hearing-aid users and their counterparts with typical hearing reached almost ceiling level in terms of accuracy in the audiovisual identification of consonants. However, in terms of IPs, the hearing-aid users’ performance was inferior (IPs were longer) when compared with individuals with typical hearing.

In addition, audiovisual presentation (relative to auditory-only) reduces the cognitive demand required for the processing of an impoverished speech signal (Mishra, Lunner, Stenfelt, Rönberg, & Rudner, 2013; Mishra, Stenfelt, Lunner, Rönberg, & Rudner, 2014; Moradi et al., 2013). By providing supplementary information about the identity of a degraded speech signal, visual cues reduce the signal uncertainty and decrease the computational demand and demands on the inference-making process needed to map degraded speech signals onto their corresponding phonological or lexical representations (see the ELU model, Rönberg et al., 2013). In a gating paradigm study, Moradi et al. (2013) showed that audiovisual presentation (relative to auditory-only) at equivalent SNRs not only expedited the identification of speech stimuli but also greatly reduced the cognitive demand required. Using a dual-task paradigm, Fraser, Gagné, Alepins, and Dubois (2010) similarly showed that audiovisual presentation (relative to auditory-only) at an equivalent SNR reduced the listening effort (i.e., the attention requirements; Hicks & Tharpe, 2002) required for understanding speech in background noise.

For consonants, studies have shown a robust audiovisual benefit (the benefit provided by the addition of visual cues to an auditory speech signal), with audiovisual presentation resulting in earlier and more accurate identification than auditory-only presentation in individuals with hearing impairment (Moradi et al., 2016; Walden et al., 2001; Walden & Montgomery, 1975). This audiovisual benefit is more evident under degraded listening conditions, such as those incurred by hearing loss (i.e., Moradi et al., 2016; Sheffield, Schuchman, & Bernstein, 2015) or in background noise (i.e., Moradi et al., 2013), where access to the critical acoustic cues of consonants is reduced. In such cases, visual cues are complementary rather than redundant (see Moradi et al., 2016), enabling disambiguation of the identity of the target consonant by providing cues about the place of articulation and when and where to expect the onset and offset of a specific consonant (Best, Ozmeral, & Shinn-Cunningham, 2007).

No study has yet investigated the audiovisual benefit in the identification of vowels in adult listeners with hearing impairment. Current research on the audiovisual benefit for vowels in listeners with typical hearing is inconclusive. Some studies have shown that audiovisual presentation (relative to auditory-only) improves the identification of vowels in participants with typical hearing (Blamey, Cowan, Alcantara, Whitford, & Clark, 1989; Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998). For instance, Breeuwer and Plomp (1986) reported that the combination of visual cues and acoustic cues (an audiovisual modality) improved identification of vowels compared with an auditory-only modality (83% in audiovisual vs. 63% in auditory-only). In contrast, other studies have shown very little or no audiovisual benefit at all in the identification of vowels (Kim, Davis, & Groot, 2009; Ortega-Llebaria, Faulkner, & Hazan, 2001).

In addition, some studies have shown that the degree of audiovisual benefit for vowels is less than that for consonants (Borrie, 2015; Kang, Johnson, & Finley, 2016). For instance, Kim et al. (2009) studied the extent to which the addition of visual cues to degraded auditory presentations of consonants and vowels affected identification compared with auditory-only presentations. The auditory presentations of consonants and vowels were filtered using either amplitude modulation (AM condition), in order to have only amplitude envelope cues in the speech signal, or a combination of frequency modulation (FM) and AM (AM + FM condition), to have both envelope and spectral cues in the speech signal. The authors reported evident audiovisual benefit for consonant identification in both the AM and AM + FM conditions. For vowels, there was a small benefit from the addition of visual cues in the AM condition and no benefit at all in the AM + FM condition, such that the mean percent correct scores of vowels in the AM + FM condition were the same in the auditory and audiovisual modalities. The authors suggested this was due to lower visual saliency for vowels than consonants, which resulted in little or no audiovisual benefit in the identification of vowels. Further, Valkenier, Duyne, Andringa, and Başkent (2012) have reported that the amount of audiovisual benefit provided is dependent on SNRs. An improvement in Dutch vowel recognition was observed only under highly taxing noise conditions (at SNRs of −6, −12, and −18 dB), and there was no difference between audiovisual and auditory vowel recognition at SNRs of 30 and 0 dB.

The present study aimed to use a gating paradigm (Grosjean, 1980) to investigate the extent to which the combination of visual cues and an amplified auditory speech signal affects the identification of Swedish consonants and vowels, in terms of IP and accuracy, in listeners with hearing impairment using hearing aids. In the gating paradigm, successive fragments of a given speech token (e.g., a consonant) are presented to participants, whose task is to guess the identity of that speech token as more fragments of the signal are presented. The major aim of the gating paradigm is to measure the IP, which, as noted earlier, is the shortest time from the onset of a speech stimulus that is needed for correct identification of that speech token. In contrast to accuracy, which has a discrete scale (i.e., correct or incorrect), the IP enables a wide range of responses for the identification of speech stimuli, even in silent listening conditions when accuracies will be at ceiling level (Moradi et al., 2013, 2016). In addition, from a cognitive hearing-science perspective (Arlinger, Lunner, Lyxell, & Pichora-Fuller, 2009), the present study investigated the cognitive demands of identifying
consonants and vowels presented in an audiovisual or an auditory-only modality (by examining relationships between participants’ IPs for consonants and vowels in each modality and their WMC).

On the basis of our prior study (Moradi, Lidgett, Häggren, & Rönnberg, 2014), and given the deficit in auditory coding of speech signals in listeners with hearing impairment, even under aided conditions (see Ahlstrom et al., 2014; Bor et al., 2008; Davies-Venn & Souza, 2014), we expected that identification of consonants and vowels presented in an auditory-only modality would be more cognitively demanding. Furthermore, we anticipated that listeners with hearing impairment who had greater WMC would identify consonants and vowels earlier than those who had lower WMC. In the case of evident audiovisual benefit for consonants and vowels, we hypothesized that audiovisual presentation would reduce the cognitive demands of identifying consonants and vowels, and make their identification not cognitively demanding in listeners with hearing impairment using hearing aids (similar to our prior study on listeners with typical hearing; Moradi et al., 2013). However, in the case of little or no audiovisual benefit, we hypothesized that identification of consonants and vowels presented audiovisually would remain cognitively demanding, similar to identification in an auditory-only modality.

Method

Participants

The study comprised 199 listeners with hearing impairment (113 men and 86 women) with bilateral, symmetrical, mild-to-severe sensorineural hearing loss who completed the gated and cognitive tasks in the n200 project (for more details, see Rönnberg et al., 2016). In brief, the n200 project is an ongoing longitudinal study on the interaction of speech signal and cognition in listeners with hearing impairment. The participants were randomly selected from an audiology-clinic patient list at Linköping University Hospital, Sweden. The age range of participants was 33–80 years; the mean age was 61.1 years (SD = 8.2). The participants were experienced hearing-aid users who had used their hearing aids for more than 1 year at the time of testing.

Figure 1 shows the mean hearing thresholds over eight frequencies (250, 500, 1000, 2000, 3000, 4000, 6000, 8000 Hz) for the participants in the present study. The mean hearing thresholds across eight frequencies were 44.36 dB HL (SD = 10.13) for the right ear and 44.30 dB HL (SD = 9.76) for the left ear.

All participants were native Swedish speakers who reported themselves to be in good health, with no history of Parkinson’s disease, stroke, or other neurological disorders that might affect their ability to perform the speech and cognitive tasks. All participants had normal or corrected-to-normal vision with glasses.

The Linköping regional ethical review board approved the study (Dnr: 55-09 T122-09). All participants were fully informed about the study and gave written consent for their participation.

Linear Amplification

In order to assure audibility, linear amplification was adjusted according to each participant’s hearing thresholds. The linear amplification was based on voice-aligned compression (VAC) rationale (Buus & Florentine, 2002; for more technical details, see Ng, Rudner, Lunner, Pedersen, & Rönnberg, 2013; Rönnberg et al., 2016). VAC is an Oticon processing strategy that provides a linear-gain 1:1 compression ratio corresponding to pure-tone input levels ranging from 30 to 90 dB SPL. VAC aims to provide less compression at high input levels and more compression at low input levels via a lower compression knee-point (i.e., increased gain for weaker inputs). In fact, the target objective of VAC is to improve subjective sound quality, so it is heard as natural with no loss of speech intelligibility.

Stimuli

A male native Swedish talker with a general Swedish dialect read the Swedish consonants and vowels at a natural articulation rate, in a quiet studio while looking straight into the camera. A Sony DV Cam DSR-200P was used for the video recordings of speech stimuli. The frame rate of video recordings was 25 fps, with a resolution of 720 × 576 pixels. The talker maintained a neutral facial expression, avoided blinking, and closed his mouth before and after articulation. The hair, face, and top part of his shoulders were visible. The auditory speech stimuli were recorded with an electret condenser microphone attached to the camera.
The sampling rate of the recording was 48 kHz, and the bit depth was 16 bits. Each target speech item was recorded several times, and the best of the recorded items (on the basis of the quality of the audio and video items) were selected. Speech stimuli were saved as .mov files. Each speech item was then edited into separate short clips (gates) to be presented in the gating paradigm. For instance, consonant /l/ consisted of 15 clips, wherein Clip 1 contained the first 40 ms of /l/ (the gate size in the present study was 40 ms for both consonants and vowels; see later), Clip 2 contained the first 80 ms of /l/, and so on, until Clip 15, which contained the complete presentation of /l/. The quality of short clips of each speech item was rechecked to eliminate sound clicks and incongruence between audio and video speech signals.

**Gated Speech Tasks**

**Consonants**

Five Swedish consonants, structured in a vowel–consonant–vowel syllable format (/af/, /aːl/, /am/, /aːs/, and /aːta/), were used in both auditory and audiovisual modalities. The first vowel (/a/) was presented and the gating started immediately at the onset of the consonant. As noted earlier, the gate size was 40 ms; the first gate included the vowel /a/ plus the initial 40 ms of the consonant. The second gate added a further 40 ms of the consonant (a total of 80 ms of the consonant), and so on. The consonant gating task took approximately 7 min to complete.

**Vowels**

Five Swedish vowels, structured in a consonant–vowel format (/pl, /mːl, /mv/, /vːl, /m/), were used in both the auditory and audiovisual modalities. This consonant–vowel format was used because earlier studies have shown that when vowels are presented in a consonant–vowel–consonant format, the critical acoustic and articulatory features of target vowels are not always distinguishable (Lindblom, 1963; Stevens & House, 1963). To deliver better acoustic cues and clear articulation of vowels to the listeners with hearing impairment, we chose the consonantal context for each vowel that met those criteria. Initial consonants were presented, and the gating started from the beginning of the vowel onset. The gate size was 40 ms, similar to the consonant gating task. The vowel gating task took around 7 min to complete.

Participants in the n200 project attended three separate sessions to provide auditory data (e.g., temporal fine-structure assessment, distortion-product otoacoustic emissions testing), speech data (e.g., Hearing In Noise Test, speech gated tasks), and cognitive data (e.g., visuospatial working memory test, reading span test [RST]; for a detailed description of the tasks used in the n200 project, see Rönberg et al., 2016). Each session took 2–3 hr to complete. Collecting the gating data for all 26 Swedish consonants and 23 vowels in auditory and audiovisual modalities would have required at least two separate sessions for these data alone, which was beyond the available time in the n200 project and was likely to have increased the dropout rate of participants. Because of these limitations, the second author of the present study chose five consonants and five vowels that varied in terms of acoustic features. For instance, regarding manner of articulation, the selected consonants comprise a plosive (/t/), fricatives (/l/, /s/), a nasal (/m/), and a lateral (/l/). Regarding place of articulation, they consist of a bilabial (/m/), a labiodental (/f/), and alveolars/dentals (/l/, /s/, /t/). The selected vowels varied in terms of duration (/aː/, /iː/ as long vowels and /a/, /aː/ as short vowels) and mouth shape (/iː/, /iː/ and /aː/). The gated consonants and vowels were presented to participants in the second session of the n200 project. In the current study, we report only the results of the gated consonants and vowels presented in the auditory and audiovisual modalities, and consider their associations with RST scores in order to examine the cognitive demand associated with their identification in each modality.

**Cognitive Test**

The RST (Daneman & Carpenter, 1980; Rönberg, Arlinger, Lyxell, & Kinnefors, 1989) was used to measure participants’ WMC. The RST involves the retention and recall of words embedded within blocks of two to five sentences. Half of the sentences were sensible (semantically correct), such as “Pappan kramade dottern” (“The father hugged his daughter”), and the other half were absurd (semantically incorrect), such as “Räven skrev poesi” (“The fox wrote poetry”). Sentences were presented visually, word by word, in the middle of a computer screen, at a rate of one word per 800 ms. The RST required two parallel actions: comprehension and retention. The participants’ task was to respond “no” to an absurd sentence and “yes” to a sensible sentence. The RST started with two-sentence sets, followed by three-sentence sets and so forth, up to five-sentence sets. After each set of (two, three, four, or five) sentences, the participants were asked to recall either the first or the last words of each sentence in the current set in their correct serial order. Participants’ RST scores were determined on the basis of the total number of correctly recalled words across all sentences. The maximum RST score was 28.

**Procedure**

The gated consonants and vowels were presented in quiet to participants seated in a sound booth. An Apple MacBook Pro equipped with Tc/TK and Quick TimeTel software was used to present the gated speech stimuli, monitor participants’ progress, and collect responses. The MacBook Pro was outside the sound booth and was configured for dual-screen presentation. This was used to display the face, hair, and top part of the talker’s shoulders against a gray background on a 17-in. Flatron monitor (LG L1730SF) inside the sound booth, viewed from a distance of about 50 cm. The monitor was turned off during the auditory-only presentation.

All participants received linear amplification (VAC, see earlier) on the basis of their audiograms.
This type of linear amplification has been used by Ng et al. (2013) to investigate the effect of noise and WMC on memory processing of speech stimuli in hearing-aid users. To deliver a linearly amplified audio speech signal to each participant, the MacBook Pro was routed to the input of an experimental hearing aid (Oticon Epq XW behind-the-ear) located in an anechoic box (Brüel & Kjær, Type 4232), the output of which was coupled with an IEC-711 ear simulator (Brüel & Kjær, Type 4157). The auditory speech signal was then transferred via an equalizer (Behringer Ultra-Curve Pro, Model DEQ2496) and another measuring amplifier (Brüel & Kjær, Type 2636) into a pair of ER3A insert earphones inside the sound booth, where the participants sat.

A microphone (in the sound booth, routed into an audiometry device) delivered the participants’ verbal responses to the experimenter through a headphone connected to the audiometry device. Participants gave their responses orally and the experimenter wrote these down.

All participants began with the consonant-identification task, followed by the vowel-identification task. The modality of presentation (audiovisual vs. auditory) within each gated task (consonants and vowels) was counterbalanced across participants, such that half of the participants started with the audiovisual modality (for both consonants and vowels) and the other half started with the auditory modality (for both consonants and vowels). The participants received written instruction about how to perform the gated tasks. They were asked to attempt identification after each gated phoneme had been presented, regardless of how uncertain they were about their identification of that phoneme, but to avoid random guessing. There was no feedback from the experimenter during the presentation of gated stimuli with regard to the correctness of answers. In order to avoid random guessing, the presentation of gates continued until three consecutive correct answers had been given. If the participants correctly repeated their response for three consecutive gates, it was considered a correct response. The IP in this case was the first gate for which the participant gave the correct response. After three correct answers, the presentation of gates for that item was stopped and the gating for a new item was started.

When a target phoneme was not correctly identified, the IP for that phoneme was scored as its total duration plus one gate size (this scoring method matches with our prior studies and other studies that have utilized the gating paradigm; Elliott, Hammer, & Evan, 1987; Hardison, 2005; Lidestam, Moradi, Pettersson, & Ricklefs, 2014; Metsala, 1997; Moradi et al., 2013; Moradi et al., 2014; Moradi, Lidestam, Sarem, & Rönnberg, 2014).

Results

Figure 2 displays the mean IPs and accuracies for the gated speech tasks. A 2 (modality: audiovisual, auditory) × 2 (phoneme: consonants, vowels) repeated-measure analysis of variance (ANOVA) was conducted to examine the effect of modality on the mean IPs and accuracies of the gated speech tasks. In terms of IPs, the results showed a main effect of modality, $F(1, 198) = 133.26, p < .001, \eta^2_p = .40$, a main effect of phoneme class, $F(1, 198) = 20.25, p < .001, \eta^2_p = .09$, and a Modality × Phoneme class interaction, $F(1, 198) = 98.20, p < .001, \eta^2_p = .33$. Planned comparisons showed that audiovisual presentation (relative to auditory-only) significantly shortened IPs for both consonants, $(198) = 13.64, d = 1.06$, and vowels, $(198) = 2.61, p = .010, d = 0.19$.

In terms of accuracy, the results showed a main effect of modality, $F(1,198) = 84.65, p < .001, \eta^2_p = .30$, a main effect of phoneme class, $F(1, 198) = 101.43, p < .001, \eta^2_p = .34$, and a Modality × Phoneme class interaction, $F(1, 198) = 13.62, p < .001, \eta^2_p = .06$. Planned comparisons using McNemar’s test for paired data showed that audiovisual presentation (relative to auditory-only) significantly improved accuracy for both consonants ($p < .001$) and vowels ($p < .001$).

In addition, statistical results are reported separately for consonants and vowels in order to examine the effect of modality within consonants and vowels. The mean RST score of participants was $16.05 (SD = 3.84)$. The Appendix presents the confusion matrices and the $d’$ scores for the consonants and vowels in auditory and audiovisual modalities; data were extracted from the correct and incorrect responses across all gates in the gating tasks.

Consonants

IPs

Figure 3 displays the mean IPs and accuracies for each of the five gated consonants in the auditory and audiovisual modalities. A 2 (modality: audiovisual, auditory) × 5 (consonant: /l/, /s/, /m/, /t/, /f/) repeated-measure ANOVA was computed to examine the effects of modality on the mean IPs of consonants. The results showed a main effect of modality, $F(1, 198) = 186.17, p < .001, \eta^2_p = .49$, a main effect of consonant, $F(4, 792) = 34.60, \eta^2_p = .34$, and a Modality × Consonant interaction, $F(4, 792) = 39.46, p < .001, \eta^2_p = .49$. Planned comparisons with Bonferroni adjustments showed that audiovisual presentation (relative to auditory-only) significantly shortened the IPs for all consonants.

Accuracy

A 2 (modality: audiovisual, auditory) × 5 (consonant: /l/, /s/, /m/, /t/, /f/) repeated-measure ANOVA was conducted to examine the effects of modality on the mean accuracy of consonant identification. The results showed a main effect of modality, $F(1, 198) = 92.59, p < .001, \eta^2_p = .32$, a main effect of consonant, $F(4, 792) = 3.437, \eta^2_p = .09$, and Modality × Consonant interaction, $F(4, 792) = 14.81, p < .001, \eta^2_p = .07$. Planned comparisons using McNemar’s test for paired data with Bonferroni correction showed that audiovisual
presentation (relative to auditory-only) improved accuracy for /s/, /t/, and /f/.

**Vowels**

**IPs**

Figure 3 displays the mean IPs and accuracies for each of the five gated vowels in the auditory and audiovisual modalities. A 2 (modality: audiovisual, auditory) × 5 (vowel: /ɪ/, /ə:/, /ʏ/, /iː/, /a/) repeated-measure ANOVA was computed to examine the effects of modality on the mean IPs of vowels. The results showed a main effect of modality, $F(1, 198) = 6.78, p = .010$, $\eta_p^2 = .03$, a main effect of vowel, $F$(Greenhouse–Geisser corrected: 3.107, 615.192) = 15.32, $p < .001$, $\eta_p^2 = .07$, and a Modality × Vowel interaction, $F$(Greenhouse–Geisser corrected: 3.314, 656.235) = 8.84, $p = .04$, $\eta_p^2 = .04$. Planned comparisons with Bonferroni adjustments showed that the audiovisual presentation (relative to auditory-only) shortened IPs only for /ʏ/.

**Accuracy**

A 2 (modality: audiovisual, auditory) × 5 (vowel: /ɪ/, /ə:/, /ʏ/, /iː/, /a/) repeated-measure ANOVA was conducted to examine the effects of modality on the mean accuracy of vowel identification. The results showed a main effect of modality, $F(1, 198) = 16.70, p < .001$, $\eta_p^2 = .08$, a main effect of vowel, $F$(Greenhouse–Geisser corrected: 3.542, 701.233) = 34.45, $p < .001$, $\eta_p^2 = .15$, and a Modality × Vowel interaction, $F$(Greenhouse–Geisser corrected: 3.701, 732.891) = 11.77, $p < .001$, $\eta_p^2 = .06$. Planned comparisons using McNemar’s test for paired data with Bonferroni correction showed that the audiovisual presentation (relative to auditory-only) improved the accuracy for only /ʏ/.

**Cognitive Demands of Consonant and Vowel Identification**

A correlation matrix was generated to assess the relationships between participant age, hearing-threshold variables, audiovisual and auditory IPs for consonants and vowels, and RST scores in listeners with hearing impairment using hearing aids (Table 1). Hearing-threshold variables consisted of two separate variables. The first was the mean pure-tone frequencies of 500, 1000, 2000, and 4000 Hz (or PTF4). Nábělek (1988) showed that PTF4 had the highest correlation coefficients with vowel identification; hence, PTF4 was included in the correlation matrix to examine its correlation with vowels and other variables in the present study. The second variable was hearing-threshold average (HTA) for all seven frequencies from 250 to 8000 Hz.

The results showed that age was significantly correlated with all other measures: Increasing age was associated with poorer PTF4 and HTA, lower WMC, and longer audiovisual and auditory IPs for consonants and vowels. In the auditory modality, HTA had the greatest correlations with consonants and vowels; PTF4 was correlated with consonants but not vowels. In the audiovisual modality, only HTA was correlated with consonants; neither of the
hearing-threshold variables was correlated with vowels in the audiovisual modality. Of particular interest to the present study are the correlations between audiovisual and auditory IPs for consonants and vowels and RST scores. Figure 4 shows the correlation plots between RST scores and audiovisual and auditory IPs for consonants and vowels. The results showed that better performance in the RST was associated with earlier identification of consonants in the auditory modality (but not in the audiovisual modality) and earlier identification of vowels in both auditory and audiovisual modalities.

In order to further explore the contribution of WMC to the audiovisual and auditory IPs for consonants and vowels, we created two groups of participants: high- and low-WMC.

Table 1. Correlation matrix for age, hearing-threshold average (HTA), pure-tone frequencies (PTF4), audiovisual and auditory isolation points (IPs) for consonants and vowels, and the reading span test (RST) scores in aided listeners with hearing impairment.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Age</td>
<td>0.30**</td>
<td>0.16*</td>
<td>0.28**</td>
<td>0.25**</td>
<td>0.24**</td>
<td>0.18</td>
<td>−0.35**</td>
<td></td>
</tr>
<tr>
<td>2. HTA</td>
<td>0.89**</td>
<td>0.25**</td>
<td>0.23**</td>
<td>0.17</td>
<td>0.11</td>
<td>0.06</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>3. PTF4</td>
<td>0.17</td>
<td>0.18</td>
<td>0.39**</td>
<td>0.12</td>
<td>0.19**</td>
<td>0.18</td>
<td>0.24**</td>
<td>−0.05</td>
</tr>
<tr>
<td>4. Consonants (auditory)</td>
<td>0.27**</td>
<td>0.27**</td>
<td>0.17</td>
<td>0.24**</td>
<td>0.18</td>
<td>0.60**</td>
<td>−0.21**</td>
<td></td>
</tr>
<tr>
<td>5. Consonants (audiovisual)</td>
<td>0.17</td>
<td>0.24**</td>
<td>−0.05</td>
<td>−0.18</td>
<td>−0.05</td>
<td>−0.22**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Vowels (auditory)</td>
<td>0.17</td>
<td>0.24**</td>
<td>0.12</td>
<td>0.19**</td>
<td>0.24**</td>
<td>0.18</td>
<td>0.60**</td>
<td>−0.21**</td>
</tr>
<tr>
<td>7. Vowels (audiovisual)</td>
<td>0.17</td>
<td>0.24**</td>
<td>0.12</td>
<td>0.19**</td>
<td>0.24**</td>
<td>0.18</td>
<td>0.60**</td>
<td>−0.21**</td>
</tr>
<tr>
<td>8. RST</td>
<td>0.17</td>
<td>0.24**</td>
<td>0.12</td>
<td>0.19**</td>
<td>0.24**</td>
<td>0.18</td>
<td>0.60**</td>
<td>−0.21**</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01.
To do this, we classified participants as having high or low WMC depending on whether their scores fell within the upper or lower quartiles of the RST score distribution, respectively. Fifty-seven participants (29 men and 28 women, mean age = 57.96 years, SD = 9.34) were categorized as having high WMC (mean RST score = 20.60, SD = 1.79), and 74 (49 men and 25 women, mean age = 64.53 years, SD = 6.24) were categorized as having low WMC (mean RST score = 12.19, SD = 2.16).

Figure 5 shows the mean IPs for consonants and vowels presented via auditory-only and audiovisual modalities in the high- and low-WMC groups. In the auditory-only modality, the t-test results for independent groups showed that the mean IPs for the high-WMC group were significantly shorter for both consonants (260 vs. 307 ms), t(129) = 2.70, p = .008, and vowels (201 vs. 244 ms), t(129) = 2.87, p = .005. In the audiovisual modality, there was no significant difference between high- and low-WMC groups in terms of mean IPs for consonants (193 vs. 200 ms), t(129) = 0.69, p = .495. However, the high-WMC group had significantly shorter IPs for vowels compared with the low-WMC group (183 vs. 228 ms), t(129) = 3.38, p < .001. Together, these findings are in agreement with the correlation coefficients (see Figure 4); they indicate that individuals with greater WMC were able to identify consonants and vowels in the auditory-only modality earlier than those with lower WMC. In the audiovisual modality, however, individuals with greater WMC were only able to identify vowels (not consonants) earlier than those with lower WMC.

We also conducted a multiple-regression analysis to investigate the predictive effect of WMC on the audiovisual and auditory IPs for consonants and vowels. Because participant age was correlated with both WMC and hearing-threshold variables (HTA and PTF4), and given that there were also high correlation coefficients within hearing-threshold variables (see Table 1), only HTA and WMC were included in the analyses as predictor variables to avoid the possibilities of a suppressor effect and multicollinearity. The multiple-regression analyses indicated that WMC is a significant predictor of IPs for consonants and vowels in the auditory modality and for vowels (but not consonants) in the audiovisual modality (see Table 2).

**Discussion**

Overall, the present study shows that although audiovisual presentation (relative to auditory-only) facilitated identification of both consonants and vowels in listeners with hearing impairment using hearing aids, this audiovisual benefit was more evident for consonants than vowels. Listeners with hearing impairment using hearing aids who had...
greater WMC identified consonants and vowels earlier in the auditory-only modality, implying cognitively demanding auditory identification of consonants and vowels. Audiovisual presentation reduced the cognitive demand required for the identification of consonants but not vowels.

The Audiovisual Benefit for Consonants and Vowels in Listeners With Hearing Impairment Using Hearing Aids

Consonants

Audiovisual presentation of consonants resulted in earlier IPs (196 vs. 288 ms) and more accurate identification (96% vs. 81%) than auditory-only presentation (see Figure 2). In terms of IPs, this audiovisual benefit was observed for all consonants used in the present study (/f/, /l/, /m/, /s/, and /t/). In terms of accuracy, this benefit was only observed for /s/, /t/, and /f/. The accuracy for /l/ was at ceiling level in auditory-only (97%) and audiovisual modalities (100%), which may explain the lack of audiovisual benefit for this consonant. For /m/ (a bilabial consonant), confusion matrices showed that audiovisual presentation of this consonant was more difficult for some listeners with hearing impairment using hearing aids, because they perceived it onto other bilabials in Swedish such as /b/ and /p/. Hence, it seems that the associative visual cue of /m/ with its amplified auditory signal was not sufficiently helpful to resolve confusions for visemes of the same class.

Together, these findings corroborate those of our recent study (Moradi et al., 2016), in which it was reported that audiovisual presentation generally shortened IPs and improved accuracy for the identification of consonants

Table 2. Summary of multiple regression analyses for variables predicting audiovisual and auditory isolation points for consonants and vowels (n = 199). HTA = hearing-threshold average; WMC = working memory capacity.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Consonants</th>
<th></th>
<th></th>
<th>Vowels</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Auditory</td>
<td>Audiovisual</td>
<td></td>
<td>Auditory</td>
<td>Audiovisual</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>SE B</td>
<td>β</td>
<td>B</td>
<td>SE B</td>
<td>β</td>
<td>B</td>
</tr>
<tr>
<td>HTA</td>
<td>2.64</td>
<td>0.74</td>
<td>0.24***</td>
<td>1.34</td>
<td>0.40</td>
<td>0.23***</td>
</tr>
<tr>
<td>WMC</td>
<td>-4.44</td>
<td>1.81</td>
<td>-0.17*</td>
<td>-0.55</td>
<td>0.99</td>
<td>-0.04</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001.
(18 Swedish consonants) in older adults using hearing aids who wore their own hearing aids during the experiment.

A more detailed comparison of the findings of the present study and those of the Moradi et al. (2016) study revealed that audiovisual presentation shortened the IPs for /l/, /ml/, /fl/, and /s/ in both studies. Although audiovisual presentation shortened the IPs for /l/ in the present study, it did not in the previous study. The findings of that previous study are consistent with those of Walden et al. (2001), who reported that visual cues provided the least benefit for /l/ in listeners with hearing impairment who wore hearing aids with nonlinear settings for 10 weeks (and had acclimatized to their hearing aids). One explanation might be that participants in the Moradi et al. (2016) study had worn their own digital hearing aids with nonlinear amplification settings for at least 1 year (and had acclimatized to their hearing aids), whereas participants in the present study received linear amplification during the experiment. Hence, differences in the type of amplification (linear vs. nonlinear) and/or acclimatization to hearing aids may, to some extent, affect the benefit for identifying a given consonant associated with the addition of visual cues.

Related to this, independent studies have shown that although there is generally no difference in the identification of consonants when different amplification settings are used, each amplification setting can have quite specific effects on given consonants (Strelec, Li, Rodriguez, Kalluri, & Edwards, 2013). For instance, Souza and Gallun (2010) showed that wide dynamic range compression (a nonlinear amplification setting commonly used in current digital hearing aids) was better at reducing the similarity of /l/ to other consonants than compressed limiting amplification. Because nonlinear amplification provides better audibility of /l/ than linear amplification, we hypothesize that the association of visual cues with the nonlinear amplification of /l/ might have resulted in a redundancy effect, whereas the linear amplification of /l/ might have resulted in a complementary effect as illustrated in the present study. Overall, the results of the present study are in agreement with those of prior research, showing the superiority of audiovisual presentation (relative to auditory-only) in improving consonant identification in listeners with hearing impairment using hearing aids (Moradi et al., 2016; Walden et al., 2001).

Vowels

On closer inspection, an audiovisual benefit was observed only for /ʊ/—in terms of both IP and accuracy—which globally resulted in a small but significant overall audiovisual benefit for the vowels used in the present study in terms of shortened IPs (208 vs. 222 ms) and improved accuracy (77% vs. 71%). However, the observed overall audiovisual benefit is in agreement with studies that have shown the facilitative effect of audiovisual over auditory presentation in the identification of vowels (Blamey et al., 1989; Breeuwer & Plomp, 1986; Robert-Ribes et al., 1998).

The only explanation we can offer for observing audiovisual benefit for only /ʊ/ is based on the confusion matrices, which suggest that the associative visual cue of /ʊ/ with its amplified auditory presentation substantially helped the listeners with hearing impairment using hearing aids to discard other phonologically similar vowels (e.g., /u/, /ʊ/, /l/, /u/) in the process of audiovisual identification of /ʊ/, compared with auditory-only identification. The number of incorrect responses for /ʊ/, interestingly, was increased in the audiovisual relative to the auditory-only modality (18 in audiovisual vs. 6 in auditory-only; see the confusion matrices). This suggests that the listeners with hearing impairment using hearing aids were struggling to differentiate /ʊ/ from the closest viseme neighbor in the audiovisual modality (/ɛ/), hindering correct identification.

When comparing the extent of audiovisual benefit for consonants and vowels, our findings demonstrated a greater benefit for identification of consonants than for vowels. First, the relative effect size of audiovisual benefit in IPs was large for consonants and small for vowels ($d = 1.06$ vs. $d = 0.19$). Second, as noted earlier, audiovisual presentation shortened IPs in the identification of all of five consonants used in the present study and improved accuracy for three consonants, whereas it shortened the IPs and improved the accuracy for only one vowel. Third, in terms of accuracy, the identification of consonants presented audiovisually almost reached ceiling level (96%), but this was not the case for vowels (77%).

This latter point is important, because the association of visual cues with the amplified speech signal of consonants was predominantly complementary, helping the listeners with hearing impairment using hearing aids to finally identify the consonants. In contrast, the association of visual cues with the amplified speech signal of vowels was considerably redundant, such that the listeners could not effectively resolve the confusion between neighboring vowels in the audiovisual modality to correctly identify the vowels (see the confusion matrices). This is most likely because the visual cues for vowels are not sufficiently distinguishing. For instance, the visual cues of the neighboring vowels /o/, /u/, /l/, /l/, /l/, /l/, and /r/; and /l/, /l/, /l/, and /l/ are almost the same, and listeners need to hear auditory cues to distinguish them from each other. Although visual cues can enable one to distinguish long vowels from short vowels (e.g., /a/ vs. /a/ or /i/ vs. /i/), a study by Lidestam (2009) with young Swedish listeners with typical hearing showed no effect of adding visual cues in the discrimination of Swedish vowel duration. Further, although lip rounding seems to provide reliable visual cues in the discrimination of vowels, Kang et al. (2016) found that differences in vowel lip rounding had no effect on their audiovisual identification when visual cues were added to the auditory presentation of speech stimuli.

Together, our findings are in agreement with studies that show less audiovisual benefit for vowels than consonants (Borrie, 2015; Kim et al., 2009). Overall, our findings suggest that the association of visual cues with the auditory speech signal was superadditive to consonants and additive to vowels in aided listeners with hearing impairment. This is probably a result of the lower visual saliency (decreased
visibility of the speech signal) for vowels than consonants. The degree of visual saliency has been shown to be a key factor in audiovisual benefit (Arnal, Morillon, Kell, & Giraud, 2009; Hazan et al., 2006; van Wassenhove, Grant, & Poeppel, 2005), because highly visible phonemes are processed more rapidly than less visible phonemes (van Wassenhove et al., 2005).

**Cognitive Demands in the Identification of Consonants and Vowels Presented in the Auditory and Audiovisual Modalities**

**Consonants**

The results of the present study are in agreement with studies that show that simply providing audibility by amplification of sounds, either linearly or nonlinearly, does not fully restore consonant intelligibility in people with hearing loss (Ahlstrom et al., 2014; Davies-Venn & Souza, 2014; Moradi, Lijestad, Hallgren, & Rönnberg, 2014). This makes the identification of consonants cognitively demanding for listeners with hearing impairment using hearing aids (Moradi, Lijestad, Hallgren, & Rönnberg, 2014). This finding is in line with the ELU model’s prediction (Rönnberg et al., 2008, 2013), in that explicit cognitive resources such as working memory are needed to infer a phoneme from a given ambiguous sound through a perceptual completion process (see also Moradi, Lijestad, Hallgren, & Rönnberg, 2014).

The combination of visual cues and amplified auditory presentation of consonants reduced the cognitive demands of consonant identification and made it cognitively nondemanding. In an audiovisual modality, the visual articulations of consonants are typically available earlier than auditory cues (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Smeele, 1994). These initial visual cues elicit only predictions (residual errors) that are matched with this initial visual articulation (predictive coding hypothesis; Friston & Kiebel, 2009). For instance, the initial visual articulation of /r/ corresponds to the initial articulation of both /r/ and /l/, and hence listeners need to hear and see a little more of the incoming signal to correctly identify the given phoneme. In an auditory-only modality, the number of residual errors made when hearing the initial parts of given consonants (e.g., /l/) is greatly increased, and this necessitates explicit cognitive resources to perceptually complete ambiguous sounds as phonemes for identification. Having fewer residual errors in an audiovisual relative to an auditory-only modality frees up cognitive resources and subsequently reduces the working memory processing demands of identifying consonants (see Fritsouva, W inneke, & Phillips, 2013; Mishra et al., 2013).

**Vowels**

Similar to consonants, identification of vowels presented in an auditory modality was cognitively demanding, and listeners with hearing impairment using hearing aids who had greater WMC identified vowels earlier than those with lower WMC. This is in line with findings by Molis and Leek (2011), who suggested that perceptual uncertainty in the identification of vowels in listeners with hearing impairment would increase the cognitive effort required in the process of vowel identification. In older adults with typical hearing, Gilbertson and Lutfi (2014) reported the contribution of cognitive function (inhibitory control) on masked vowel recognition. Note that they also used the Wechsler Memory Scale–Revised Digit Span test (Wechsler, 1981) as a measure of WMC; however, they found no relationship between those results and masked vowel recognition performance. The discrepancy in findings between their study and ours might be a result of the different means of measuring WMC. The digit-span test is mainly a short-term memory test in which performance is dependent upon storage capacity (maintaining a sequence of digits and then repeating it), whereas performance in the RST is dependent upon both storage and processing, and the RST is more cognitively taxing than the digit-span test. In a review, Akeroyd (2008) argued that only those cognitive tasks which are sufficiently taxing, such as the RST, are correlated with measures of speech recognition in degraded listening conditions.

In the present study, there was no significant correlation between PTF4 and auditory IPs for vowels. This is at odds with the findings of Nábělek (1988), who reported that PTF4 had the highest correlation with vowel recognition, particularly in noise and reverberation conditions. This discrepancy might be due to differences in the types of participants included in the studies. The participants in the Nábělek study were extremely heterogeneous in terms of age and hearing loss, belonging to four separate groups: young listeners with typical hearing, older adults with typical hearing, listeners with hearing impairment with mild hearing loss, and listeners with hearing impairment with moderate hearing loss. In addition, vowels were presented monaurally to the preferred ear at a comfortable presentation level. In contrast, the participants in the present study had hearing impairment, and the presentation of speech stimuli was individually amplified in a linear manner. The differences in the types of participants and presentation of speech stimuli may explain why PTF4 was not correlated with IPs for vowel recognition in the present study.

In contrast to the findings for consonants, audiovisual identification of vowels in listeners with hearing impairment using hearing aids was still surprisingly cognitively demanding, most likely a result of the lower visual saliency provided by vowels. In line with the predictive coding hypothesis, we argue that in contrast to consonants, the number of residual errors made in identifying a given vowel in an audiovisual modality is considerably high (similar to the visual articulation of vowels within a visemic class, as described already). Hence, listeners with hearing impairment using hearing aids require explicit cognitive resources to discriminate audiovisually similar vowels from each other and correctly perceive an ambiguous audiovisual signal as a given vowel.

This cognitively demanding audiovisual identification of vowels challenges the notion of cognitive spare
capacity (Mishra et al., 2013, 2014), which suggests that under degraded listening conditions, solely adding visual cues to auditory speech stimuli will reduce the cognitive demands of speech-stimuli processing. On the basis of our findings, we argue that the degree of visual saliency is the key factor in reducing the cognitive demand in audiovisual identification of speech stimuli under degraded listening conditions. That is, higher visual saliency (e.g., in terms of complementary consonants) greatly reduces the cognitive demand in the process of audiovisual identification of consonants. However, lower visual saliency (e.g., in terms of redundancy vowels) has little or no impact in reducing the cognitive demand.

Clinical Implications

From a clinical perspective, we suggest that other rehabilitation approaches are needed (in addition to hearing aids) to compensate more fully for the difficulties experienced by people with hearing loss in perceiving phonemes. Auditory training has been shown to improve phoneme recognition in those with hearing loss, both in those who do not use hearing aids (Ferguson, Henshaw, Clark, & Moore, 2014) and in those who do (Stecker et al., 2006; Walden, Erdman, Montgomery, Schwartz, & Prosek, 1981; Woods & Yund, 2007). On the basis of the findings of this study, we suggest that facing communication partners (as face-to-face training) in a well-lit room (and wearing eyeglasses if needed) would be another rehabilitative approach in improving phoneme recognition in people with hearing loss. Audiovisual training can be another rehabilitative approach to improve phonemic recognition in people with hearing loss. To the best of our knowledge, no study has evaluated the efficiency of audiovisual training for improving identification of consonants or vowels in people with hearing loss. However, in listeners with typical hearing, Richie and Kewley-Port (2008) showed that audiovisual vowel training improved vowel identification. Shahin and Miller (2009) have also reported that audiovisual training improved phonemic-restoration ability in listeners with typical hearing (involving a top-down perceptual mechanism in which the individual forms a coherent speech percept from a degraded auditory signal; Warren, 1970). This finding by Shahin and Miller (2009) is critically important, because it indicates that audiovisual training may be used to repair deficits in phonemic restoration caused by hearing loss in listeners with hearing impairment (Başkent, Eiler, & Edwards, 2010).

Limitations and Future Considerations

The findings of the present study were based on five Swedish vowels and five Swedish consonants (out of 26 consonants and 23 vowels in the Swedish language), which may raise concerns about the generalizability of our findings to Swedish consonants and especially vowels as a whole. We suggest that future studies explore the audiovisual benefit and subsequent reduction in cognitive demands using a larger sample of the consonants and vowels that are available in a given language. Nevertheless, our results not only replicate but also extend the findings of prior independent studies, and make theoretical sense.

In the present study, cognitive demand was determined on the basis of the association of RST scores with audiovisual and auditory IPs for consonants and vowels. For future studies, we suggest that listening effort related to the identification of consonants and vowels presented aurally or audiovisually in individuals with hearing impairment be measured using pupillometry (e.g., Zekveld, Kramer, & Festen, 2010, 2011) or the dual-task paradigm (e.g., Sarampalis, Kalluri, Edwards, & Hafter, 2009; Sommers & Phelps, 2016).

Linear amplification of sounds during the performance of the gated task was new to listeners with hearing impairment, who wore their own hearing aids for daily communication. This new amplification setting (for which there was no adequate acclimatization time) may require more working memory processing for perceiving speech stimuli (see Lunner, Rudner, & Rönnberg, 2009). We suggest that future studies investigate the extent to which acclimatization to hearing-aid settings may affect the cognitive demand involved in identifying consonants and vowels presented aurally or audiovisually. In addition, we studied only aided identification of consonants and vowels in auditory and audiovisual modalities. For future research, we suggest comparing aided and unaided identification of consonants and vowels in auditory and audiovisual modalities, to examine the effects of amplification and visual cues on recognition and the cognitive demand posed by identification.

Only one talker was used to produce the gated speech stimuli in the auditory-only and audiovisual modalities. Individual differences in lipreading might influence the audiovisual benefit received by participants when identifying consonants or vowels (see Grant & Seitz, 1998; Tye-Murray, Spehar, Myerson, Hale, & Sommers, 2016). The extent to which lipreading ability influences audiovisual identification of consonants and particularly vowels would be an interesting research question for future studies.

Conclusion

Audiovisual presentation improved accuracy and reduced the phoneme duration necessary for identification of consonants and vowels relative to auditory-only presentation in listeners with hearing impairment using hearing aids. However, this audiovisual benefit was more evident for consonants than vowels. Despite linear amplification, auditory identification of consonants and vowels was cognitively demanding; listeners with hearing impairment who had greater WMC identified consonants and vowels earlier. The combination of visual cues and an amplified speech signal reduced the cognitive demand of identifying consonants but not vowels.
Acknowledgments

This research was supported by a grant from the Swedish Research Council (awarded to Jerker Rönberg) and a program grant from the Swedish Research Council for Health, Working Life, and Welfare (awarded to Jerker Rönberg). We thank Tomas Björm, Helena Torlofson, and Wychiffle Yumba, who assisted in collecting data, and Mathias Hållgren for his technical support.

References


Confusion matrices and $d'$ scores for consonants and vowels presented in an auditory or audiovisual modality

**Consonants-Auditory Modality**

<table>
<thead>
<tr>
<th>consonant</th>
<th>f</th>
<th>l</th>
<th>m</th>
<th>s</th>
<th>t</th>
<th>d</th>
<th>k</th>
<th>n</th>
<th>η</th>
<th>r</th>
<th>j</th>
<th>g</th>
<th>f</th>
<th>h</th>
<th>v</th>
<th>b</th>
<th>p</th>
<th>$d'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>133</td>
<td>29</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>11</td>
<td>2</td>
<td>2.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>l</td>
<td>193</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>11</td>
<td>4.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>13</td>
<td>161</td>
<td>1</td>
<td>14</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>2.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>7</td>
<td>168</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>2.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>1</td>
<td>154</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>21</td>
<td>1</td>
<td>3.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Consonants-Audiovisual Modality**

<table>
<thead>
<tr>
<th>consonant</th>
<th>f</th>
<th>l</th>
<th>m</th>
<th>s</th>
<th>t</th>
<th>d</th>
<th>k</th>
<th>n</th>
<th>η</th>
<th>r</th>
<th>j</th>
<th>g</th>
<th>f</th>
<th>h</th>
<th>v</th>
<th>b</th>
<th>p</th>
<th>$d'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>198</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>5.38</td>
</tr>
<tr>
<td>l</td>
<td>198</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>5.25</td>
</tr>
<tr>
<td>m</td>
<td>3</td>
<td>178</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>12</td>
<td>1</td>
<td>4.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>2</td>
<td>187</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td>1</td>
<td>4.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>2</td>
<td>190</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Vowels-Auditory Modality**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>154</td>
<td>39</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a:</td>
<td>8</td>
<td>186</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>3.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ɪ</td>
<td>136</td>
<td>32</td>
<td>9</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>1.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>28</td>
<td>140</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>9</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td>3</td>
<td>2.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ʏ</td>
<td>45</td>
<td>2</td>
<td>87</td>
<td>6</td>
<td>2</td>
<td>25</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>15</td>
<td>2.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Vowels-Audiovisual Modality**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>157</td>
<td>35</td>
<td></td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>3.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a:</td>
<td>11</td>
<td>184</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>3.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ɪ</td>
<td>132</td>
<td>31</td>
<td>1</td>
<td></td>
<td></td>
<td>14</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>1.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>43</td>
<td>153</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td>2.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ʏ</td>
<td>7</td>
<td>1</td>
<td>143</td>
<td>18</td>
<td>2</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data were extracted from correct and error responses across all gates. Rows represent stimuli present, and columns represent participants’ responses.  
*The number of non-responses for a given consonant or vowel after whole presentation of that speech item.