MARCH 03 2025

Unimodal speech perception predicts stable individual differences in audiovisual benefit for phonemes, words and sentences[a)]

Jacqueline von Seth ; Máté Aller ; Matthew H. Davis

J. Acoust. Soc. Am. 157, 1554–1576 (2025)
https://doi.org/10.1121/10.0034846

JASA ARTICLE

Unimodal speech perception predicts stable individual differences in audiovisual benefit for phonemes, words and sentences[a)]

Jacqueline von Seth,[b)] Máté Aller, and Matthew H. Davis

Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Road,
Cambridge CB2 7EF, United Kingdom

ABSTRACT:
There are substantial individual differences in the benefit that can be obtained from visual cues during speech perception. Here, 113 normally hearing participants between the ages of 18 and 60 years old completed a three-part experiment investigating the reliability and predictors of individual audiovisual benefit for acoustically degraded speech. Audiovisual benefit was calculated as the relative intelligibility (at the individual-level) of approximately matched (at the group-level) auditory-only and audiovisual speech for materials at three levels of linguistic structure: meaningful sentences, monosyllabic words, and consonants in minimal syllables. This measure of audiovisual benefit was stable across sessions and materials, suggesting that a shared mechanism of audiovisual integration operates across levels of linguistic structure. Information transmission analyses suggested that this may be related to simple phonetic cue extraction: sentence-level audiovisual benefit was reliably predicted by the relative ability to discriminate place of articulation at the consonant-level. Finally, whereas unimodal speech perception was related to cognitive measures (matrix reasoning and vocabulary) and demographics (age and gender), audiovisual benefit was predicted only by unimodal speech perceptual abilities: Better lipreading ability and subclinically poorer hearing (speech reception thresholds) independently predicted enhanced audiovisual benefit. This work has implications for practices in quantifying audiovisual benefit and research identifying strategies to enhance multimodal communication in hearing loss. https://doi.org/10.1121/10.0034846

(Received 14 June 2024; revised 18 November 2024; accepted 17 December 2024; published online 3 March 2025)

[Editor: James F. Lynch] Pages: 1554–1576

I. INTRODUCTION

Speech production is inherently linked to observable motion in the face, which includes the jaw, lips, and tongue of the speaker. Throughout the course of our lives, we acquire substantial experience with these signals during face-to-face conversation. It is well-established that when the acoustic speech signal is degraded, speech cues encoded in facial movements can provide a significant benefit to speech perception (Sumby and Pollack, 1954). Yet, despite the ubiquity of these signals in our everyday perceptual experience, not everyone benefits equally. Previous work has reported substantial individual differences in measures of the audiovisual advantage across a wide range of speech materials: from minimal nonsense syllables to meaningful sentences (Aller *et al.*, 2022; Grant *et al.*, 1998; Grant and Seitz, 1998; Sommers *et al.*, 2005; Tye-Murray *et al.*, 2016; Van Engen *et al.*, 2014; Van Engen *et al.*, 2017).

A. What accounts for individual differences in audiovisual speech perception?

The reasons for this variability remain poorly understood: Only measures of lipreading ability have been reliably linked to individual differences in audiovisual speech perception (see Bernstein *et al.*, 2022, for review). Some research has also suggested that the degree of acquired hearing loss (HL) in mild-to-moderate hearing-impaired (HI) listeners may predict better lipreading ability (e.g., Bernstein *et al.*, 2000; Suess *et al.*, 2022; Tillberg *et al.*, 1996) and enhanced audiovisual integration for speech perception (e.g., Altieri and Hudock, 2014; Puschmann *et al.*, 2019). However, this effect is not always found (Rosemann and Thiel, 2018; Spehar *et al.*, 2008; Tye-Murray *et al.*, 2007a) and substantial individual differences remain, meaning that too few of those with age-related HL can use visual speech to mitigate the negative consequences of HL (e.g., Punch *et al.*, 2019). Additionally, lipreading ability and audiovisual integration for speech perception are notoriously difficult to train (Preminger and Ziegler, 2008; Richie and Kewley-Port,

[a)] Portions of this work were presented in "Inter-individual variability and correlates of audiovisual speech benefit in behaviour and MEG," in 15th Annual Meeting of the Society for the Neurobiology of Language, Marseille, France, October 2023; "Stable individual differences in audiovisual benefit for speech perception: Exploring the role of perceptual and cognitive abilities," in 15th Speech-in-Noise workshop, Potsdam, Germany, January 2024; "Perceptual and cognitive predictors of stable individual differences in audiovisual and unimodal speech perception," EPS Meeting, Nottingham, U.K., April 2024. A preprint is now available at https://osf.io/preprints/psyarxiv/kj59f (Last viewed January 8, 2025).

[b)] Email: Jacqueline.vonSeth@mrc-cbu.cam.ac.uk

2008). The small improvements in phoneme-level recognition obtained in some lipreading programmes may not generalise to more natural or audiovisual speech stimuli (see Bernstein *et al.*, 2022, for review).

Explanations for individual differences in lipreading and audiovisual speech perception have also been sought in terms of nonspeech cognitive abilities. Feld and Sommers (2009) suggested that processing speed and visuospatial working memory may account for a large amount of the substantial individual variability in lipreading ability in younger and older adults. They argued that if fundamentally stable cognitive traits underlie individual differences in lipreading and audiovisual speech perception, this may explain why training programmes often show limited success. It is well-established that cognitive abilities play a significant role in auditory-only (AO) speech perception in noise (Akeroyd, 2008; Dryden *et al.*, 2017; Heinrich *et al.*, 2015), especially when the signal is degraded (Pichora-Fuller *et al.*, 1995). However, for measures of audiovisual speech perception and audiovisual integration for speech perception, specifically, the picture is less clear. Dual-task demands seem to impair performance on audiovisual speech tasks (Fraser *et al.*, 2010; Alsius *et al.*, 2005; Alsius *et al.*, 2014; Buchan and Munhall, 2011). However, susceptibility to the McGurk effect (McGurk and MacDonald, 1976), which is frequently used as a measure of audiovisual integration for speech perception, is not related to processing speed, working memory, or attentional control (Brown *et al.*, 2018). Similarly, visual enhancement of speech perception (i.e., enhanced report for auditory-visual (AV) compared to AO speech) in school-aged children is also not predicted by performance on cognitive tasks measuring vocabulary knowledge, working memory, or attentional control (Lalonde and McCreery, 2020).

B. Quantifying individual differences in audiovisual benefit

A key challenge in this line of research is the lack of reliable measures of audiovisual integration for speech perception. The lack of correlations among different audiovisual integration measures, including speech- and nonspeech illusions, have been taken to suggest that only measures derived from congruent speech materials may be useful in predicting an individual's ability to use visual cues in ecological conditions (Wilbiks *et al.*, 2022; but see Dong *et al.*, 2024; and Magnotti *et al.*, 2020, for arguments that susceptibility to the McGurk effect may be related to audiovisual speech-in-noise perception). Previous research has most frequently compared unimodal and AV performances at the same level of acoustic clarity or background noise (BN), taking the auditory condition as a baseline (hereafter, *visual enhancement*). The choice of audiovisual integration measure has significant implications regarding the conclusions that may be drawn, for example, concerning the question of whether audiovisual integration for speech perception declines or increases with age (Dias *et al.*, 2021; Sommers *et al.*, 2005; Tye-Murray *et al.*, 2007a). However, even within the same measure, establishing stable individual differences across different speech materials has proven difficult: Visual enhancement of consonant report does not seem to predict visual enhancement for word or sentence report tasks (Grant and Seitz, 1998; Sommers *et al.*, 2005), whereas individual differences in unimodal speech perception are highly related across levels of linguistic structure (Grant *et al.*, 1998; Humes *et al.*, 1994; Sommers *et al.*, 2005; but for lipreading ability, see Bernstein *et al.*, 2000).

These inconsistencies suggest problems for traditional models of audiovisual speech perception, which propose a separate stage of multisensory integration for speech, which should account for a significant amount of variability in audiovisual outcomes (Altieri and Hudock, 2014; Grant *et al.*, 1998; Huyse *et al.*, 2014). If individual differences in audiovisual speech perception are related to a domain-general audiovisual integration ability, measures of visual enhancement should generalise across materials and levels of linguistic structure. In a review of shortcomings of the McGurk effect as a measure of audiovisual speech integration ability, Van Engen *et al.* (2017) suggested a potential explanation for the lack of correlations: Could audiovisual integration rely on different mechanisms at different levels of linguistic structure (e.g., minimal syllables versus meaningful sentences)? In line with this, Sommers (2021) proposed that audiovisual integration for speech perception may not be conceived of as an individual differences measure in the traditional sense—unlike working memory or processing speed, which may be tapped into by different tasks. However, shortcomings of the currently predominant visual enhancement measure provide an alternative explanation for these inconsistent results. It (1) may not adequately capture integration (see Sommers, 2021, for review; Sommers *et al.*, 2005; Tye-Murray *et al.*, 2010); (2) is confounded by differences in intelligibility between conditions; and (3) is susceptible to ceiling and floor effects, truncating the individual variability that is measured. So far, however, the development of more sophisticated capacity or efficiency measures to model individual differences has not yielded promising results in terms of predicting the ability to use visual cues at the sentence-level (Altieri and Hudock, 2014; Blamey *et al.*, 1989; Braida, 1991; Grant and Seitz, 1998; Massaro and Cohen, 1983; Sommers *et al.*, 2005; Wilbiks *et al.*, 2022).

Here, we apply a relatively simple measure of individual differences in audiovisual speech perception, following Aller *et al.* (2022), based on approaches estimating speech reception thresholds at 50% accuracy (e.g., Macleod and Summerfield, 1987). By comparing audiovisual and AO speech perception in materials approximately equated for intelligibility, we avoid confounds introduced by differences in intelligibility between conditions as well as floor and ceiling effects, which appear in the visual enhancement measure depending on which level of acoustic clarity is chosen for experimental conditions. We also assess whether our measure is stable across sessions (and items) and levels of linguistic structure (and speakers).

C. The current study

In recent years, rapid advances in software tools for online experiments (De Leeuw *et al.*, 2023; de Leeuw, 2015; Rodd, 2024) combined with online participant panels allow us to quickly collect reliable data from a balanced sample across age groups and gender. These include older participants with more diverse educational backgrounds who may not be easily targeted by university-based recruitment. This is especially useful for individual differences research, which requires larger samples to achieve sufficient power in testing for cross-condition correlations (for example, comparing audiovisual benefit across levels of linguistic structure in consonant, word, and sentence report tasks). In the current study, we aimed at quantifying the degree of audiovisual benefit using an intelligibility-matched measure in normally hearing (NH), working-age adults (18–60 years old). We measured the relative intelligibility of matched AO and audiovisual speech for materials at three levels of linguistic structure: meaningful sentences, monosyllabic words, and minimal consonant-vowel syllables. Isolating individual variability across speech materials, we tested the degree to which variability in lipreading ability, hearing status, linguistic, and cognitive ability along with demographic variables (age and self-reported gender) explain individual differences in unimodal outcomes and audiovisual benefit in speech perception.

II. METHODS

A. Participants

142 British English native speakers were recruited via Prolific Academic.[1] Participants provided informed consent using an online consent form approved by the Cambridge Psychology Research Ethics Committee (Application No. PRE.2022.056). Participants were screened at the beginning of each session using Wood's headphone test (Woods *et al.*, 2017), excluding 14 participants who were compensated for their time.

Across two sessions, 113 participants successfully completed all audiovisual speech perception tasks, and 103 participants [55 female and 48 male; age range, 18–60 years old; *mean age* ± *standard deviation* (SD) = 38.54 ± 11.55] successfully completed all tasks across three sessions, and did not meet any outlier exclusion criteria. The target sample size ($n = 101$) was estimated using G*Power 3.1.9.4 (Faul *et al.*, 2009) based on pooled effect sizes (weighted for sample size and Hedge's g) from previous studies investigating the reliability of lipreading ability and visual enhancement across speech materials ($g = 0.26$) and Pearson's product correlations of audiovisual benefit and enhancement measures with lipreading ability ($g = 0.76$) and hearing status ($g = 0.33$) to achieve power ≥ 0.8 to test each of our three main hypotheses.

1. Outlier exclusion and data quality

We administered self-report questionnaires of attention, technical difficulties, and task comprehension to identify any issues that might require participants to be excluded after each speech perception task and the cognitive and hearing tests. For any ratings of >3 [on six-point Likert scales for (1) attention, (2) technical difficulties, and (3) clarity of task instructions], typed responses were manually reviewed ($n = 22$). Data from participants with a rating of >3 and no typed responses or responses substantiating difficulties were excluded ($n = 6$), but we decided to retain participants whose responses indicated task comprehension, engagement, and attention (for example, correctly describing task instructions) while acknowledging that they found the task difficult. Additional preset outlier exclusion criteria for the cognitive and speech perception tasks included $<80\%$ in catch trials ($n = 4$) and lapse rate of >0.0625 ($n = 2$), as well as performance of 1.5 interquartile ranges (IQRs) below the first or above the third quartile ($n = 5$). Additional data quality checks leading to the exclusion of individual trials (but not participants) are detailed in individual task descriptions below.

B. Stimuli

1. General description

Consonants in minimal syllables, monosyllabic words, and meaningful sentences were presented to participants in separate audiovisual speech perception tasks across two sessions. Video recordings were drawn from Aller *et al.* (2022), Krason *et al.* (2023b), and Pimperton *et al.* (2019) for sentence-, word-, and consonant-level materials, each produced by a different level. Videos were cropped to show only the face of the speakers, who performed minimal head movements. Example images and links to video recordings can be found in the original publications.

We manipulated the availability of visual speech cues and the degree of acoustic clarity using noise vocoding (Shannon *et al.*, 1995) to create five conditions: visual-only (VO), auditory-only low acoustic clarity (AO_{low}), auditory-only high acoustic clarity (AO_{high}), auditory-visual low acoustic clarity (AV_{low}), and auditory-visual high acoustic clarity (AV_{high}). The availability of visual speech was manipulated by either presenting the originally recorded video or a video of a largely static face produced by repeating frames prior to visual speech onset. Acoustic clarity was manipulated following the same procedure described in Aller *et al.* (2022), which is based on a protocol developed by Zoefel *et al.* (2020), whereby each of the 16 narrowband envelopes env(b), extracted at logarithmically spaced frequency bands b (70–5000 Hz, half-wave rectified, lowpass filtered 30 Hz) is mixed with the broadband envelope *env(broadband)* at proportion p, such that

$$\text{env}_{\text{final}}(b) = \text{env}(b)p + \text{env}(\text{broadband})(1 - p). \qquad (1)$$

The resulting envelopes $\text{env}_{\text{final}}(b)$ were used to modulate noise in each respective frequency band. Recombined signals yielded a mix of 16-/1-channel vocoded speech, ranging from $p = 1$. The level of acoustic clarity was calibrated separately for each task to match two conditions,

AV_{low} and AO_{high}, for intelligibility and ensure that they fall at an intermediate level of intelligibility (40%–60%) to avoid floor or ceiling effects in either the audiovisual or AO condition (see Fig. 1). Mixing proportions p were chosen based on visual inspection and psychometric curves fit to pilot data collected for each task ($n = 9$ for reporting isolated words and forced-choice identification of minimal syllables, eight females and one male, mean age $\pm SD = 31.11 \pm 5.15$) using the *quickpsy* package in *R* (Linares and López-Moliner, 2016). This resulted in $p = 0$ (low clarity) and $p = 1$ (high clarity) for consonants, and $p = 0.47$ (low clarity) and $p = 1$ (high clarity) for words. Additionally, we retained the $p = 0.2$ and $p = 0.7$ conditions for sentences based on a previous pilot study in Aller *et al.* (2022). The difference between these measures was intended to show a mean of zero and a spread of positive and negative values, indicating the degree of audiovisual benefit obtained by individual participants.

2. Item characteristics

a. Consonants. Recordings of 20 consonants in minimal syllables followed by /ə/, spoken by a single female speaker with a neutral (mouth closed) start and end position, were presented in the consonant identification tasks. See the supplementary material Table S1. Participants were instructed to classify the sounds as if they occurred at the start of words, where each consonant is followed by a variation of /æd/, /æt/, /ɛt/, or /ɛd/, or a closely related syllable (with the exception of "thaw" for /θ/), to form a real monosyllabic word. These words, where the initial sound that is highlighted made up the closed-set response options for the task. Clear speech AO recordings of the same sounds produced by a male speaker were presented in the practice phase to familiarise participants with the isolated speech sounds and their corresponding word contexts while preventing participants from learning lip configurations associated with each sound. Identical recordings of the 20 consonants, presented once in each of the 5 conditions, were repeated across both sessions.

b. Words. Video recordings of 200 common, monosyllabic words, spoken in isolation by a single female speaker, were selected from a set of items previously used in Krason *et al.* (2023b) and Krason *et al.* (2023a). See the supplementary material Table S1 for further details. Orthographic responses were cleaned by removing spaces and nonalphabetic symbols, as well as responses where participants indicated that they could not identify the target word (by typing "dk"). Responses were then scored using the Levenshtein ratio, which were calculated using the fast Levenshtein edit distance implemented in the *PanPhon* package for Python (Mortensen *et al.*, 2016). The edit distance measure for each target stimulus–response pair was expressed as a percentage of the length of the longer string and then subtracted from 100 to convert the metric into a ratio measure of word report accuracy. Previous work has indicated that using the Levenshtein distance to automatically score orthographic transcriptions is highly consistent with manually scored responses (e.g., Themistocleous *et al.*, 2020) as spelling errors are not unduly penalised when manual scoring is not feasible because of the large number of responses to be evaluated (see Baese-Berk *et al.*, 2023).

c. Sentences. Recordings of 100 meaningful sentences (number of words, $M \pm SD = 13.97 \pm 2.20$; length, $M \pm SD = 5.11 \pm 0.71$ s), a subset of the stimuli used in Aller *et al.* (2022), were presented across two sessions in the sentence-level word report tasks. Participant responses were cleaned in the same way as responses in the isolated word report task, removing extraneous spaces and symbols. Responses were scored using the token sort ratio (TSR) fuzzy logic string-matching metric (Bosker, 2021) as implemented in the *FuzzyWuzzy* Python package (SeatGeek Inc., 2014). We decided to use fuzzy string-matching metrics to score word report over more conservative item-correct measures as they

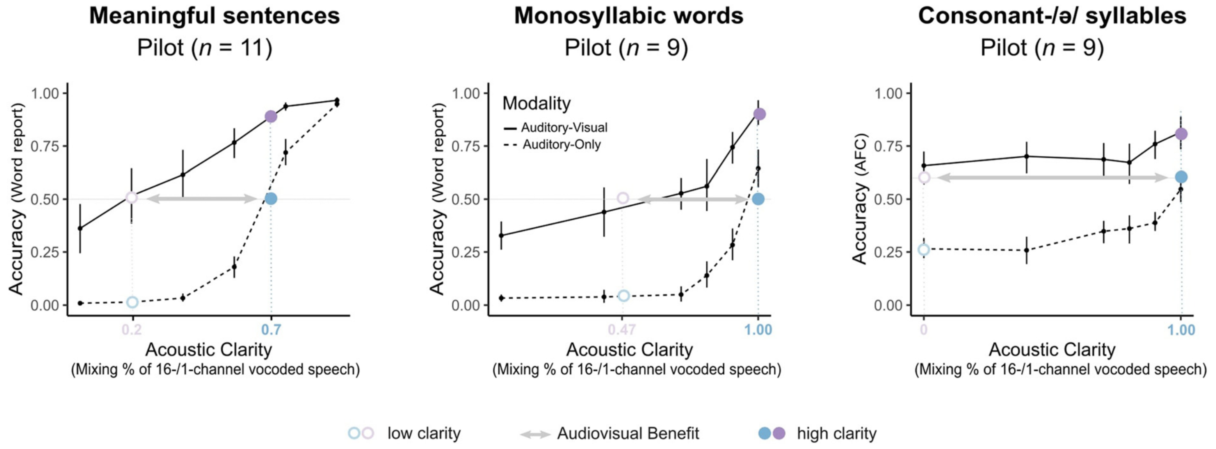

FIG. 1. Pilot data for each task illustrate the acoustic clarity levels chosen in the main experiment, matching intelligibility in intermediate AO (blue) and AV (purple) conditions separately for each level of linguistic structure. Audiovisual benefit is calculated as the difference between intermediate-intelligibility conditions (50% accuracy for sentences and words and 60% accuraxy for consonants). A fifth condition included in the experiment, silent videos (VO), is not illustrated here.

provide a better fine-grained measure of individual differences in perceptual recognition, allowing for partial matches and not unduly penalising spelling errors and homophones (compared to scoring the more stringent percent (%)words correct as originally preregistered; see Bosker, 2021).

3. Counterbalancing and randomisation

Item-session and item-condition assignments were counterbalanced across participants for the word- and sentence-level audiovisual speech perception tasks. Individual items were randomly assigned to sessions, resulting in five splits of items for each task. Ten item-condition assignments were created across all splits, and each participant was assigned to a split (item-session assignment) and version (item-version assignment). Additionally, in each audiovisual speech perception task, the presentation order of individual items was shuffled for each participant while ensuring an equal number of conditions per block (two blocks in each task for consonants and five blocks in each task for words and sentences) was retained.

C. Procedure

1. General procedure

Participants completed three experimental sessions over a period of 2–3 weeks with at least 7 days between the first two sessions (see the supplementary material Fig. S1). All tasks were coded in *jsPsych* versions 6.3 or 7.3 (De Leeuw *et al.*, 2023). At the start of each session, participants adjusted their volume to a comfortable level and completed a headphone test using antiphase sounds designed by Woods *et al.* (2017) to ensure that they were all wearing binaural headphones.

In the first two sessions, participants performed three audiovisual speech perception tasks with materials at one of three levels of linguistic structure (sentences, words, and consonants in minimal syllables) and items presented in five different conditions, varying in the availability of visual speech cues and acoustic clarity. In both sessions, completion of these tasks was preceded by a period of vocoded speech training in which ten sentences of degraded speech were presented in auditory-visual and AV conditions (mixing proportions p varying from 0.2 to 1), each preceded by a written transcription of the sentence. This was to ensure that the initial rapid perceptual learning that occurs with vocoded speech was completed by the start of the main experiment (Sohoglu and Davis, 2016).

At the start of session 1, participants additionally completed a short language questionnaire which screened for language and hearing difficulties, non-native British English speakers, and collected (voluntarily disclosed) demographic data on age, gender identity, regional accent familiarity, proficiency in languages other than English, as well as significant periods of time (>6 months) spent abroad.

In session 3, participants completed the digits-in-noise (DiN) test (Smits *et al.*, 2013) to assess individual speech-in-noise perception thresholds and section one of the abbreviated profile of hearing aid benefit (APHAB; Cox, 1997). Additionally, the listen-up task (Davis *et al.*, 2019) was administered to assess phonological discrimination thresholds. The matrix reasoning task (MaRs; Chierchia *et al.*, 2019), a non-proprietary version of Raven's progressive matrices test, and the spot-the-word (STW) lexical decision task (Baddeley *et al.*, 1993) were used to assess domain-general and verbal intelligence quotient (IQ), respectively. The order of tasks was randomised for each participant.

Finally, at the end of each audiovisual speech perception task and each of the three sessions, participants were asked to rate their comprehension of the instructions, ability to pay attention, and technical difficulties during the task or session on a scale of 1–6 and provided with a textbox to provide further details of any issues that had occurred.

2. Audiovisual speech perception tasks

For each task, participants first completed a clear speech practice block, introducing the paradigm, and for consonants, participants completed the target stimuli. For the sentence-level and word-level tasks, participants viewed three example items and were asked to type back into a textbox the words that they understood to the best of their abilities. They were encouraged to guess if unsure and warned that some of the trials may appear very difficult. They then completed 5 blocks of word report tasks in sentence- and word-level audiovisual speech tasks consisting of 50 and 100 unique items and trials per session, respectively. Each item (sentence or word) was only presented once to each participant across the entire experiment. At the level of consonants, participants performed a forced-choice task and were asked to select the target consonant out of all 20 possible options, which were presented in the context of a monosyllabic real word. Participants heard a male speaker pronouncing each target consonant during practice trials and were provided with feedback on their responses to ensure participants could correctly match each consonant to the answer options available. After this, they completed 2 blocks of alternative forced-choice (AFC) trials, each of which contained 1 presentation of each item per each of the 5 conditions, for a total number of 100 trials per session.

3. Subjective hearing experience

The first section of the revised form A of the APHAB (Cox, 1997) was administered to assess individual participant's subjective experience of the frequency of hearing and speech-in-noise perception difficulties in everyday life. The APHAB includes 24 items which can be summarised into 4 subscales relating to hearing difficulties in everyday situations: ease of communication (EC), reverberation (RV), BN, and aversiveness (AV). The overall APHAB score for each individual participant was derived from the mean of the first three of these scales. Participants are asked to rate statements such as "I miss a lot of information when I'm listening to a lecture" (BN) from never (1% of the time) to always

(99% of the time). Six items were scored in reverse order, where 99% indicates no difficulties and 1% indicates severe difficulties. Participant's attention was drawn to that fact to ensure that they answer each item carefully. Higher overall scores indicate more substantial hearing difficulties. This task was included as previous work had indicated that when including participants with known HL, scores correlate with individual lipreading ability (Suess *et al.*, 2022). Additionally, including a subjective measure may diverge from an objective measure of hearing difficulties, especially in mild cases or early-onset, while perceived listening effort may predict everyday face-viewing behaviour, which could be linked to lipreading ability or individual audiovisual benefit (Puschmann *et al.*, 2019; Rennig *et al.*, 2020).

4. Speech reception threshold

The DiN test is an established measure of speech-in-noise reception thresholds (SRTs), first introduced by Smits *et al.* (2013) as a task to screen SRTs over the telephone. In our implementation, we followed the procedure described and validated in Smits *et al.* (2013). Digit triplets consisting of a randomly chosen combination sampled from digits 0–9 was presented in long-term average speech-spectrum noise, modulated via a one-up, one-down adaptive procedure with a step size of 2 dB. In each of the 24 trials, participants were asked to report all 3 digits heard using a number pad presented on the screen, and answers were scored as triplets. SRTs are calculated as the mean signal-to-noise ratio (SNR; dB) in the final 20 trials. The DiN was chosen as it has a high test-retest reliability, correlates significantly with pure tone audiometry thresholds, and is highly sensitive to mild-moderate HL (Van den Borre *et al.*, 2021), which is indicated by a SRT of −7.4 dB SNR or above (Smits *et al.*, 2013). Because none of our participants reported a clinical diagnosis of HL and our online version is not yet sufficiently validated, we refer to any differences in hearing measured here as "subclinical."

5. Categorical speech perception

Individual phonological speech discrimination thresholds were measured using the listen-up task (Davis *et al.*, 2019). Monosyllabic, common target words (e.g., "fan" in a female voice) were accompanied by a picture of the target word, followed by presentation of two real-word audio-morphed stimuli using the target word and a minimal-pair foil ("fan" and "van" spoken by a male voice). Participants were asked to indicate which of the two words was closer to the target word. The acoustic difference between both words was progressively reduced, using an adaptive procedure (three down, one up; Levitt, 1971). Trials started with a 100% acoustic difference between the foil stimuli (i.e., resynthesised versions of the original speech), and the acoustic form of each token was reduced by 16% following three correct responses (i.e., 84% and 16% tokens were presented, subsequently, reducing to 68% and 32%, etc.). Step size was reduced by 1/sqrt(2) at each turning point. Therefore, the difficulty of this two-alternative forced-choice task (2AFC) increased progressively throughout the task until step size reached 2% and performance converged on thresholds for distinguishing target and foil spoken words. The outcome measure is the minimum proportion of acoustic difference (PADRI) between speech sounds, allowing an individual to identify the spoken words with 79.4% accuracy (PADRI threshold). Each participant completed two blocks of the listen-up task, and the PADRI threshold was averaged across two blocks. When performance in only one of the blocks met outlier exclusion criteria, the PADRI threshold estimated in the other block was retained. The inclusion of the listen-up task was not originally preregistered, but we decided to include it as an exploratory predictor as it provides a brief complementary test of participant-level variability in speech perception in addition to auditory perceptual acuity.

6. Verbal IQ

Linguistic skill was assessed using the STW lexical decision task, which was developed by Baddeley *et al.* (1993). In the STW task, participants were presented with 60 pairs of words and nonwords and asked to identify each real word in a pair. Real words ranged from frequent to obscure words, whereas nonwords were plausible and followed English orthographic conventions. A practice trial consisting of six word-nonword pairs preceded the task, and participants were instructed to complete each trial page consisting of six word-nonword pairs as quickly as possible. Vocabulary knowledge as a proxy measure of verbal IQ was scored as % correct identification of the real word in nonword-word pairs. Participants were reassured that perfect performance in this task was not expected. Additionally, trials for which reaction times significantly exceeded the expected completion time (1.5 IQRs > $Q3$ across all participants) were excluded from analyses.

7. Nonverbal IQ

Domain-general cognitive abilities were assessed using the MaRs reasoning task and are available online,[2] where individual items are drawn from the open-source MaRs-IB item bank (Chierchia *et al.*, 2019). Each item of the MaRs-IB was made up of a 3 × 3 matrix, where eight cells contain abstract shapes. Participants were asked to "complete the puzzle" by selecting the missing shape from four options presented below within 30 s of trial onset, indicated by a countdown presented for the entire 30 s. Relationships between items may be uni- or three-dimensional, and relate to the colour, shape and positions between cells. Participants saw up to 80 items, depending on how many items they manage to complete within 8 min, at which time the task finished automatically. All participants were shown the same randomly sampled items and distractor types (we used a paired difference strategy for all items) in identical order to ensure that individual differences in task performance did not arise from item-level variation in difficulty (Zorowitz *et al.*, 2024).

Trials with rapid responses (<250 ms) were excluded from analyses. We computed the measures described in Chierchia *et al.* (2019): (i) productivity (absolute number of puzzles completed), (ii) median response time (RT) for correctly completed items, (iii) accuracy (items correct divided by items attempted), and (iv) inverse efficiency (median RTs divided by accuracy). For interpretability and to index accuracy and processing speed separately, our main measures of interest to be included as predictors were reaction time for correctly completed items and accuracy of items attempted.

D. Statistical analysis

The study was preregistered and is available online.[3] Data were preprocessed and scored in *R* (version 4.2.2), MATLAB (version 2020b) and Python (version 3.11), whereas statistical analyses were performed in *R* (version 4.2.2). Anonymised data and analysis scripts are available online.[4]

Linear/logistic mixed-effects models were used to estimate main effects of acoustic clarity, modality (added visual speech), and session on accuracy measures in all three audiovisual speech perception tasks using the *lme4* package in *R*. Audiovisual benefit was calculated by taking the difference between intelligibility-matched audiovisual and AO listening conditions $\mathrm{AV}_{\mathrm{low}} - \mathrm{AO}_{\mathrm{high}}$. For completeness, we estimated three different measures of test-retest reliability across sessions: the Spearman-Brown formula, Cohen's α, and the intra-class correlation coefficient (ICC; even though only two of these measures were preregistered, we report all three; ICC allowed us to compare consistency across three levels in our cross-task comparison). Cronbach's α was computed using the *psych* package in *R* (Revelle, 2024), whereas the ICC was calculated using a two-way mixed-effects model for absolute agreement, treating separate sessions as individual raters, according to

$$\mathrm{ICC} = \frac{\text{Variance between participants}}{\text{Variance between participants} + \text{Error variance} + \text{Variance between sessions}}. \qquad (2)$$

To assess across-task reliability, we calculated correlations which were ceiling corrected for within-level test-retest reliability using the Spearman-Brown formula (adjusted by the square root of the product of the test-retest reliability of both tasks) and estimated consistency across all three tasks using a two-way mixed-effects ICC.

Finally, to isolate condition-specific rather than level-specific variance as the independent variable in the regression analysis and given the significant correlations that we observed across levels, we decided to perform principal component analysis (PCA) on standardised unimodal and audiovisual benefit measures across levels of linguistic structure. PCA scores (isolating variability in audiovisual benefit across tasks) were then predicted using multiple linear regression analysis, including standardised perceptual and cognitive measures, as well as demographic variables (age and self-reported gender) as independent predictors.

Additionally, we performed information transmission analysis (Miller and Nicely, 1955) to explore the role of phonetic feature perception (voicing, manner, and place of articulation) in predicting sentence-level audiovisual benefit. This analysis was not preregistered; therefore, we deem it exploratory here. As a result of the small number of presentations per item per subject, confusion matrices for phonetic features of interest such as voicing, manner, and place of articulation (see Table I for classification scheme), were pooled across participants prior to the calculation of relative transmitted feature information according to

$$\mathrm{IT}_{\mathrm{rel}} = \frac{I(U,V)}{H(U)}. \qquad (3)$$

Here, $I(U,V)$ describes the mutual information between the discrete variables describing presented and identified features, also known as the absolute information transmitted ($\mathrm{IT}_{\mathrm{abs}}$), and $H(U)$ describes the feature entropy of the target variable [see Oosthuizen and Hanekom, 2016, for a detailed methodological description of the classic feature information transmission analysis (FITA) approach). Analyses were conducted using functions from the *entropy* package in *R* (Hausser and Strimmer, 2009). To estimate subject-level variability, a jackknife resampling procedure was used to produce subaverage $\mathrm{IT}_{\mathrm{rel}}$ scores for 115 confusion matrices for $n - 1$. Individual estimates o_i were then retrieved from the set of subaverage scores j_i using the formula (Smulders, 2010)

$$o_i = n\bar{J} - (n-1)j_i. \qquad (4)$$

Here, $\bar{J}$ represents the mean of subaverage scores across n participants. In words, we computed the information transmission for an individual participant as the difference between information transmission for all participants and information transmission for all participants *except* that individual. We refer to this dependent measure as retrieved relative information transmission (retrieved $\mathrm{IT}_{\mathrm{rel}}$) and investigated the effect of

TABLE I. Test-retest reliability measures for within-task audiovisual benefit. $^{***}p < 0.001$.

	Sentences	Words	Consonants
Cronbach's α	0.83 [0.76 0.89][a]	0.68 [0.56 0.80]	0.54 [0.37 0.71]
ICC	0.71 [0.60 0.79]	0.51 [0.36 0.64]	0.37 [0.20 0.52]

[a] Confidence intervals for α were estimated using the Duhachek method (Duhachek and Iacobucci, 2004).

added visual speech (modality) and acoustic clarity on this dependent measure using one-way analyses of variance (ANOVAs) on ranks (Kruskal-Wallis tests, as assumptions of normality were not met; see Sec. III E). We also computed pairwise comparisons of interest and investigated the relationship of retrieved $\mathrm{IT}_{\mathrm{rel}}$ values to sentence-level measures of lipreading ability and audiovisual benefit.

III. RESULTS

A. Substantial individual differences in audiovisual benefit for sentences, words and phonemes

As in previous work (Aller *et al.*, 2022; Grant *et al.*, 1998; Grant and Seitz, 1998; Sommers *et al.*, 2005; Van Engen *et al.*, 2014; Van Engen *et al.*, 2017), we observed substantial inter-individual variability in lipreading ability and audiovisual speech processing across all three audiovisual speech tasks (Fig. 2). In each of the tasks, performance was lowest in the $\mathrm{AO}_{\mathrm{low}}$ condition (sentences, $M \pm \mathrm{SD} = 0.08 \pm 0.09$; words, $M \pm \mathrm{SD} = 0.19 \pm 0.07$; consonants, $M \pm \mathrm{SD} = 0.23 \pm 0.08$), as expected, and increased with added acoustic clarity and added visual speech [see Fig. 3(a)], as indicated by improved fit when including fixed effects of clarity and modality in logistic (for consonants) and linear (for words and sentences) mixed-effects models, according to the following specification (in the Wilkinson notation, Wilkinson and Rogers (1973):

$$\begin{aligned}\text{Accuracy} \sim\ & 1 + \text{Modality} + \text{Clarity} \\ & + \text{Modality} : \text{Clarity} + \text{Session} \\ & + (1 + \text{Modality} + \text{Clarity} \,|\, \text{Participant}) \\ & + (1 |\, \text{Item}). \qquad (5)\end{aligned}$$

For sentences, model comparisons using Kenward-Roger's F-tests suggested that a model including clarity, $F(1{,}112.12) = 1282$, $p < 0.001$, and modality, $F(1{,}112.12) = 667.83$, $p < 0.001$, provided a better fit than models

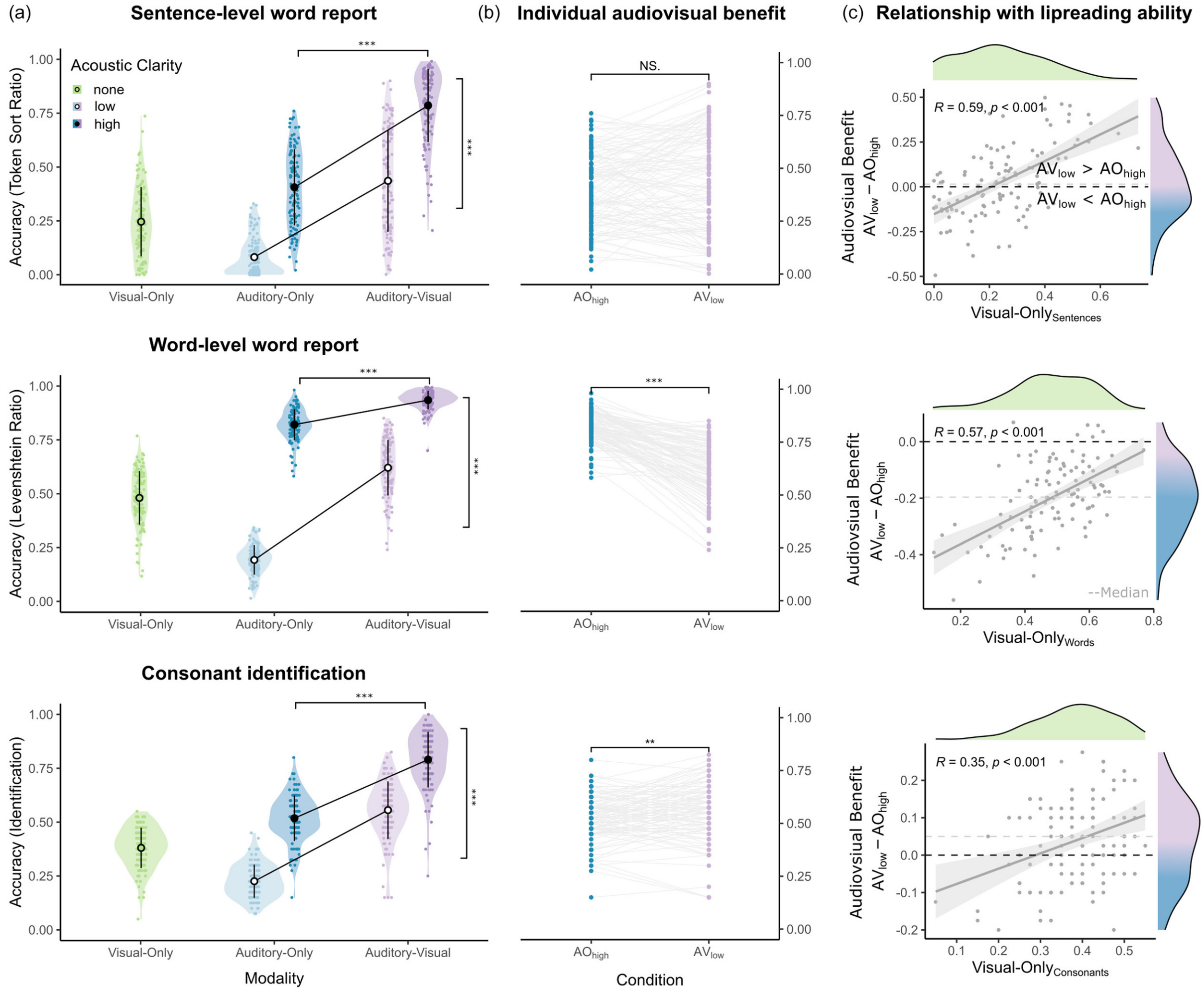

FIG. 2. Individual differences in audiovisual speech perception. (a) Results of the sentence-, word-, and consonant-level audiovisual speech perception tasks [mean ± standard error of the mean (SEM)] as well as marginal probability densities. Asterisks (***) indicate significant main effects of clarity and modality, $p < 0.001$. (b) Audiovisual benefit for individual benefits is calculated as the difference in performance between $\mathrm{AV}_{\mathrm{low}}$ and $\mathrm{AO}_{\mathrm{high}}$ conditions (gray lines). (c) Individual audiovisual benefit is significantly correlated with within-level VO perception, where marginal distributions are displayed at the right and top of the plot, respectively. The dashed gray horizontal line shows the median audiovisual benefit over participants, and the dashed black horizontal line shows zero audiovisual benefit—i.e., equivalent accuracy for $\mathrm{AV}_{\mathrm{low}}$ and $\mathrm{AO}_{\mathrm{high}}$.

without. Examination of the summary output indicated that added acoustic clarity improved word report by 33% [low versus high acoustic clarity, $\beta = 0.36$, standard error (SE) $= 0.006$, $t = 55.246$], whereas the audiovisual modality improved word report by 36% (AO versus AV condition, $\beta = 0.360$, SE $= 0.015$, $t = 24.15$). There was a small but significant interaction between clarity and modality, $F(1,8584) = 4.87$, $p = 0.027$, $\beta = 0.02$, SE $= 0.008$, $t = 2.207$, possibly driven by nonlinearities in the data introduced by floor effects in the AO$_{\text{low}}$ condition. There was also a significant improvement in overall accuracy between sessions, $F(1,8556.59) = 141.90$, $\beta = 0.05$, SE $= 0.004$, $t = 11.912$.

We observed a similar pattern of results for the word-level task with an increased accuracy of word report with added clarity, $F(1,113.43) = 4703.83$, $p < 0.001$, $\beta = 0.62$, SE $= 0.01$, $t = 112.74$, and added visual speech, $F(1,112.98) = 1743.60$, $p < 0.001$, $\beta = 0.42$, SE $= 0.01$, $t = 55.99$, as well as a very small decrease in performance between sessions, $F(1,17548.61) = 13.37$, $p < 0.001$, $\beta = 0.01$, SE $= 0.003$, $t = -3.656$. There was a significant interaction of clarity and modality also in the word-level task, $F(1,17551.58) = 1537.32$, $p < 0.001$, $\beta = -0.30$, SE $= 0.008$, $t = -39.209$, $p < 0.001$. This interaction effect is likely driven by a trend toward the ceiling in the word-level task (AO$_{\text{high}}$).

For the consonant-level task, we used mixed-effects logistic regression to explore the effects of clarity, modality, and session on the binary accuracy measure. As each participant saw each consonant per condition, by-item random slopes were also included in this model for a full random effects structure (see Barr *et al.*, 2013). Model comparisons using likelihood ratio tests indicated that models including acoustic clarity and modality provided the best fit to the data [clarity, $\chi^2(1) = 12.837$, $p < 0.001$, $\beta = 1.78$, SE $= 0.43$, $z = 4.20$; modality, $\chi^2(1) = 17.295$, $p < 0.001$, $\beta = 2.12$, SE $= 0.41$, $z = 5.18$. There was no significant interaction between clarity and modality in the consonant-level task, $\chi^2(1) = 0.01$, $p = 0.939$. Additionally, including session as a fixed effect improved model fit, suggesting a significant improvement in performance between sessions, $\chi^2(1) = 13.26$, $p < 0.001$, $\beta = 0.19$, SE $= 0.02$, $z = 9.746$. Finally, there were substantial individual differences in effects of acoustic clarity and modality on accuracy.

Our measure of audiovisual benefit was calculated as the difference in performance between the intermediate-intelligibility conditions AV$_{\text{low}}$ and AO$_{\text{high}}$. This measure demonstrates substantial differences between individual participants: Some benefitted more from added visual speech than increased acoustic clarity levels and vice versa (as indicated by the slopes of lines in column b of Fig. 2, i.e., positive slopes indicating more benefit from visual speech, whereas negative slope means more benefit from increased acoustic clarity). For example, in the sentence-level task, 59 out of 113 participants benefitted more from added visual speech than increased acoustic clarity, i.e., showed better performance in the AV$_{\text{low}}$ condition than in the AO$_{\text{high}}$ conditions (for words and consonants, these proportions were less balanced with 6 and 68, respectively, out of 113 benefitting more from added visual speech). Over all participants, these measures of audiovisual benefits significantly differed from zero for words and consonants; but this difference was not reliable for sentence-level report [sentences, mean difference (MD) $= 0.030$, $t(112) = 1.57$, $p = 0.119$; words, MD $= -0.19$, $t(112) = -16.72$, $p < 0.001$; consonants, MD $= 0.038$, $t(113) = 3.59$, $p = 0.001$; see Fig. 2(b)]. Nonetheless, all three differences straddle zero and are largely unaffected by floor or ceiling effects in the underlying data. For all three speech tasks, the degree of audiovisual benefit (i.e., difference between AV$_{\text{low}}$ and AO$_{\text{high}}$) was significantly correlated with lipreading ability measured using performance in the VO condition [sentences, $r(111) = 0.59$, $p < 0.001$; words, $r(111) = 0.57$, $p < 0.001$; consonants, $r(111) = 0.35$, $p < 0.001$; see Fig. 2(c)].

B. Audiovisual benefit is stable across sessions and consistent across levels of linguistic structure

To establish whether our measure of audiovisual benefit can be considered to be a stable difference between individuals, we calculated test-retest reliability across two sessions (set at least 1 week apart and containing different items) and consistency across three levels of linguistic structure, adjusted for within-task test-retest reliability (sentences, words, and consonants as produced by different speakers). According to standard interpretations of test-retest metric values (Hedge *et al.*, 2018), audiovisual benefit for the sentence-level task shows good test-retest reliability, whereas the word-level task shows moderate and the consonant-level task shows poor-moderate test-retest reliability [see Table I for a comparison of different measures and Fig. 3(a)].

To assess whether audiovisual benefit is also consistent across the three tasks, we calculated pairwise ceiling-corrected Spearman-Brown correlations [accounting for within-task test-retest reliability; see Fig. 3(b)] as well as Cronbach's α and ICC$_{3,k}$ across all three levels to assess consistency (two-way mixed for consistency rather than absolute agreement due to magnitude differences between different scores, using the average across two sessions, making ICC identical to alpha), which yielded values of $\alpha = 0.65$, 95% confidence interval (CI) [0.53 0.75].

Further assessing pairwise correlations between these three measures of audiovisual benefit demonstrates moderate consistency between monosyllabic words and sentences ($\alpha = 0.69$ or $\varrho = 0.75$, where ϱ is corrected for within-task test-retest reliability) and poor to moderate consistency of audiovisual benefit measures for words and sentences with the consonant-level measure (words, $\alpha = 0.35$, $\varrho = 0.31$; sentences, $\alpha = 0.49$, $\varrho = 55$). Low correlations may be driven by the moderate test-retest reliability of the consonant-level task (due to the relatively small number of trials presented). Nonetheless, these results show that across different items, stimulus types, and speakers, we find meaningful correlations in the magnitude of audiovisual benefit individuals derive: A linear regression model including word- ($\beta = 0.693$, SE $= 0.121$, $p < 0.001$) and consonant-level

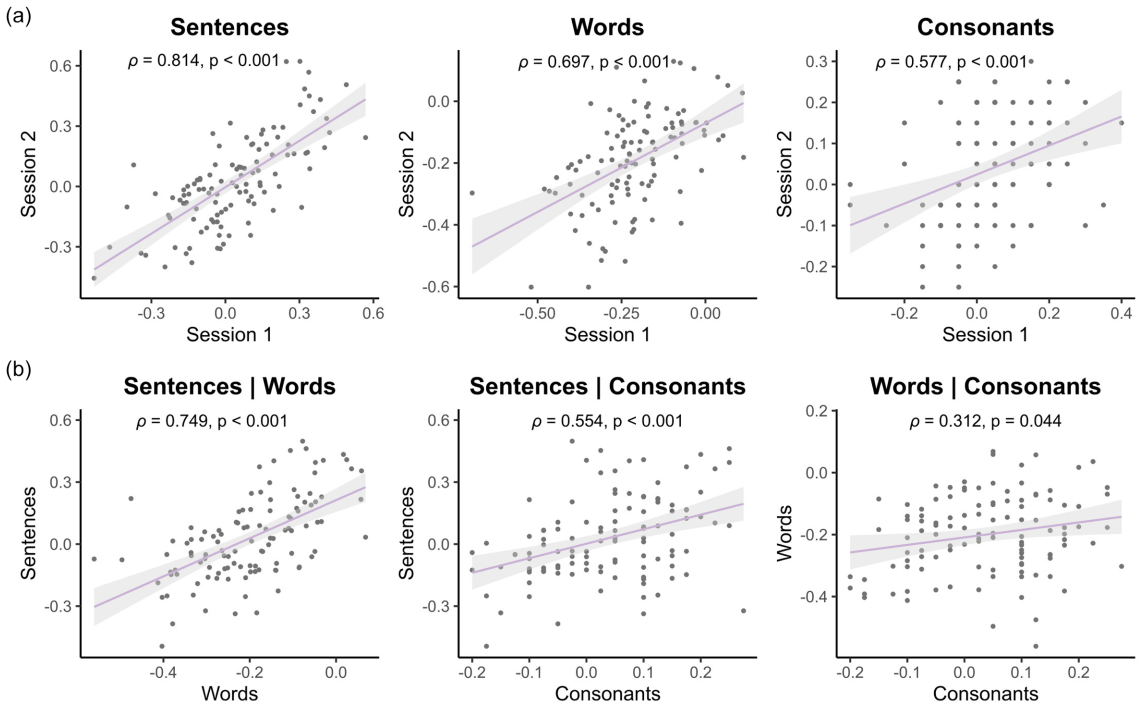

FIG. 3. Test-retest reliabilities of audiovisual benefits (a) across sessions and (b) across levels of linguistic structure are shown. Values depicted represent the Spearman-Brown measure, which is ceiling corrected for within-task re-test reliability in the across-task measure in (b).

audiovisual benefit ($\beta = 0.370$, SE $= 0.125$, $p < 0.001$) explained 33% of the variance in sentence-level audiovisual benefit, $F(2,112) = 28.32$, $p < 0.001$, $R^2 = 0.324$.

Previous studies have provided inconsistent results regarding whether measures of lipreading ability calculated for materials at different levels of linguistic structure are positively correlated (Bernstein *et al.*, 2000). Exploring this here, we observed that lipreading ability was reliably and positively correlated across tasks [sentences-words, $r(111) = 0.57$, $p < 0.001$; sentences-consonants, $r(111) = 0.43$, $p < 0.001$; words-consonants, $r(111) = 0.53$, $p < 0.001$).

C. Audiovisual benefit is independently predicted by relatively poorer hearing and better lipreading ability

Having confirmed that our measure of audiovisual benefit shows sufficient convergent validity, we performed PCA using the *principal* function in the *psych* package on standardised audiovisual benefit scores to isolate participant-level variability across tasks. All three benefit measures loaded on one component, which explained 60% of the variance in the data, with loading strengths of 0.88 for sentences, 0.80 for words, and 0.63 for consonants. Multiple linear regression was then used to predict variability in this PCA score from cognitive and hearing measures, as well as demographic variables (see Table II).

We decided to include RTs (for correct items) and accuracy (for attempted items) for the MaRs as separate predictors to improve interpretability (compared to a trade-off measure, such as inverse efficiency) and index processing speed (RT) and reasoning ability (accuracy). RT and accuracy (%) in the MaRs were weakly correlated with age [RT, $r(101) = 0.26$, $p = 0.011$; %, $r(101) = 0.20$, $p = 0.038$], indicating declines in reasoning speed but increases in reasoning accuracy with age (but neither of those correlations survived correction for multiple comparisons; see Fig. 4). Both of these measures of matrix reasoning ability were positively associated with performance on the STW task [RT, $r(101) = 0.34$, $p < 0.001$; %, $r(101) = 0.35$, $p < 0.001$]. As expected, STW performance was moderately positively correlated with age [$r(101) = 0.49$, $p < 0.001$], in line with previous observations of increased verbal IQ in older individuals (Hartshorne and Germine, 2015). There were no significant relationships between the hearing or speech perception measures, and none of these measures were correlated with age (see Fig. 4).

A series of model comparison *F*-tests suggested that only poorer hearing, as indicated by higher speech reception

TABLE II. Summary of results for cognitive and perceptual measures. The measures and participants included in the multiple regression analysis are contained herein.

Measure	*N*	*M*	SD
Vocabulary knowledge (STW accuracy)	103	0.675	0.191
Matrix reasoning (RT correct in s)	103	8.283	3.650
Matrix reasoning (accuracy for items attempted)	103	0.805	0.084
Frequency of hearing difficulties (%APHAB)	103	0.323	0.049
Speech reception threshold (DiN dB SNR)	103	−10.233	0.933
PADRI threshold (listen up)	103	0.165	0.056
Age	103	38.544	11.549

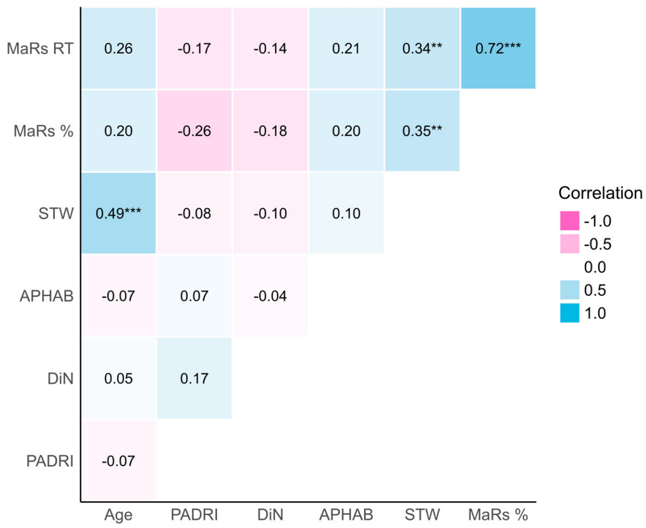

FIG. 4. Correlation matrix of cognitive, perceptual, and demographic predictors included in multiple linear regression analyses. ***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$, corrected for multiple comparisons using the Holm-Bonferroni adjustment. %, accuracy; APHAB, subjective hearing; DiN, objective hearing; PADRI, listen up.

thresholds (worse performance) estimated using the DiN test [$F(1) = 4.298$, $p = 0.041$] and better lipreading ability, measured as the mean of (standardised) VO performance across all three tasks [$F(1) = 87.845$, $p < 0.001$] predicted individual differences in audiovisual benefit, $F(9,93) = 12.33$, $R^2 = 49.9\%$; dropping either of these but none of the other predictors, significantly affected model fit [see the Fig. 5(a) and 5(b) and supplementary material Table S3)].

Previously, it has been suggested that speech-in-noise perception may explain some variability in lipreading ability (Bernstein, 2018; Watson *et al.*, 1996). However, variance partitioning analysis indicated that lipreading ability and speech reception thresholds independently predicted audiovisual benefit. Lipreading ability uniquely explained 45% [$F(1,100) = 77.448$, $p < 0.001$] of the variance in audiovisual benefit, whereas speech perception thresholds explained 7% [$F(1,100) = 8.644$, $p = 0.006$] with only 2% of variance shared between the two predictors [Fig. 5(c)]. A third variable included in this analysis, age, which has been associated previously with a decline in lipreading ability and speech reception thresholds and trended toward significance in the full regression model, did not explain any joint or unique variance in audiovisual benefit, $F(1,100) = 0.294$, $p = 0.589$.

D. Unimodal speech perception is associated with demographic variables and domain-general cognitive abilities

We also explored whether hearing status, verbal and nonverbal cognitive abilities, age, and gender explained any of the variability observed in performance in two unimodal conditions not used to calculate audiovisual benefit: VO and AO_{low} (see Fig. 6 and the supplementary material Fig. S2).

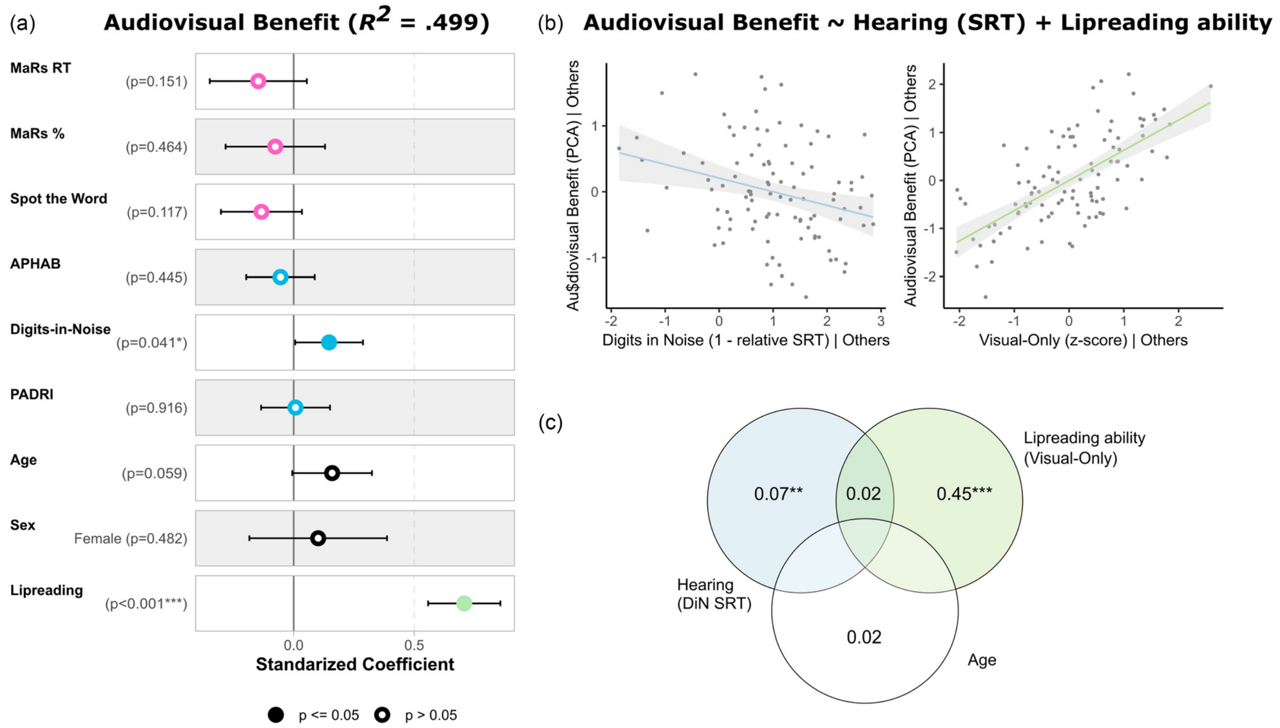

FIG. 5. Results of multiple linear regression analysis predicting audiovisual benefit. (a) Forest plot illustrates results for the full regression model, predicting audiovisual benefit PCA scores across levels. Filled circles indicate a significant predictor. (b) Partial regression plot is shown for the two predictors which significantly contribute to model fit. (c) Variance partitioning results indicate that hearing and lipreading ability independently explain variability in audiovisual benefit. Significance of partitions was tested using regularized discriminant analysis (RDA) across 999 permutations.***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$.

We, again, extracted variability across levels using PCA to isolate modality-specific, level-independent variability. AO_{low} performance loaded on one component explains 53% of the overall variance with loading strengths of 0.76 for sentences, 0.82 for words, and 0.57 for consonants. Model comparisons (see the supplementary material Table 4) revealed that better performance in the AO modality was associated with younger age [$F(1) = 12.724$, $p < 0.001$] and predicted by matrix reasoning accuracy [$F(1) = 5.106$, $p = 0.026$] and performance in the STW task [$F(1) = 9.085$, $p = 0.003$], explaining 31.6% of the variance in performance, $F(8,94) = 6.886$, $p < 0.001$. As a result of some (but not substantial) indication of multicollinearity of this model (variance inflation factor = 2.3), we also performed variance partitioning here to explore potential associations between age and measures of verbal and nonverbal IQ. This suggested that matrix reasoning accounted for 16% of the variance [$F(1,100) = 21.047$, $p = 0.001$], where 9% variance is shared between the two cognitive measures and a nonsignificant unique contribution of STW performance [$F(1,100) = 2.949$, $p = 0.087$], whereas age independently explained 8% of the variance in AO word report, $F(1,100) = 5.05$, $p = 0.020$, see Fig. 6(a).

All three measures of lipreading ability were loaded on a single PCA component, explaining 67% of the variance in the data, with loading strengths of 0.81 for sentences, 0.86 for words, and 0.79 for consonants. VO perception was related only to our two demographic measures, $F(8,94) = 2.662$, $p = 0.011$, $R^2 = 11.5\%$. Model comparisons suggested that lipreading ability decreased with age [$F(1) = 5.812$, $p = 0.017$], and participants identifying as female were better overall at lipreading than those identifying as male [$F(1) = 4.675$, $p = 0.033$], see Fig. 6(b). None of the other measures predicted individual differences in lipreading ability. See the supplementary material Table S5. Unlike that for audiovisual benefit, speech reception thresholds estimated in the DiN task did not meaningfully contribute to model fit, $F(1) = 0.939$, $p = 0.335$. Overall, these results suggest that whereas audiovisual benefit is related to perceptual abilities, unimodal speech perception (audio only and VO) is associated with demographic variables (such as age/gender) and (for auditory speech perception) measures of domain-general cognition.

E. Exploratory: Sentence-level audiovisual benefit is predicted by visual perception of place and manner of articulation features

Our previous analyses have indicated that perceptual rather than non-signal-related cognitive variables predict individual differences in audiovisual benefit for all three levels of linguistic structure tested. We, therefore, embarked on exploratory analyses to identify the perceptual cues that are most relevant to audiovisual speech perception and may be better exploited by participant showing enhanced audiovisual benefit. Our focus here is on perception of specific articulatory features that might explain variability in consonant identification. To this end, we used a classic information theoretic approach to quantify transmission of phonetic cues in unimodal and audiovisual perception of consonants: FITA (Files *et al.*, 2015; Grant *et al.*, 1998; Jesse and Massaro, 2010; Lalonde and Werner, 2019; Miller and Nicely, 1955; Walden *et al.*, 1975).

Consistent with the previous literature (e.g., Grant *et al.*, 1998), we expected that place of articulation would be most easily transmitted in the visual or audiovisual modality, whereas voicing and manner would be more easily recognised in the auditory modality. We statistically assessed two comparisons of interest: (1) AO_{low} compared to VO (i.e., which features are better transmitted in two low intelligibility conditions that convey only auditory or only visual information); (2) $AOnly_{high}$ compared to $AVisual_{low,}$ (i.e., which features are better transmitted in intermediate-intelligibility, AO, and AV conditions; see Fig. 7). Additionally, We conducted a factorial analysis to investigate main effects of auditory clarity (low/high) and visual information (absent/present) for the four auditory conditions (excluding VO). As the data analysed are retrieved IT_{rel} values, assumptions of normality are violated (as confirmed by Shapiro-Wilk tests; voicing, $W = 0.821$, $p < 0.001$; manner, $W = 0.778$, $p < 0.001$; place, $W = 0.705$, $p < 0.001$), therefore, nonparametric tests were used for statistical analysis.

Kruskal-Wallis H tests indicated that there was a main effect of modality [$\chi^2(1) = 16.543$, $p < 0.001$), as well as a main effect of clarity [$\chi^2(1) = 231.299$, $p < 0.001$] for transmission of the voicing feature [see Fig. 7(a)]. Transmission was significantly better in the AO_{low} condition than in the VO condition (MD = 0.00871, $z = 4403$, $p = 0.001$ corrected for multiple comparisons using the Holm method) and better in the AO_{high} condition compared to the AV_{low} condition (MD = 0.0348, $z = 6101$, $p < 0.001$). These observations suggest an overall auditory advantage for transmission of voicing information. These findings are consistent with the existing literature, indicating that voicing information is typically considered to be absent in visual speech signals (Lisker *et al.*, 1977; but see Raphael, 1972, 1975; and Van Son *et al.*, 1994). Nonetheless, the presence of a main effect of modality is interesting as it suggests that visual information can enhance perception of consonantal voicing contrasts when combined with auditory signals—for instance, because visual information can signal the timing of closure for stop consonants.

For manner of articulation, the picture was more complex: There was a main effect of modality [$\chi^2(1) = 85.506$, $p < 0.001$] and clarity [$\chi^2(1) = 206.027$, $p < 0.001$] on relative information transmission [see Fig. 7(b)]. Manner cues were better transmitted in the VO than the AO_{low} condition (MD = 0.0434, $z = 537$, $p < 0.001$) but more easily transmitted in the AO_{high} than in the AV_{low} condition (MD = 0.0274, $z = 5472$, $p < 0.001$), suggesting that although some manner information is available in the VO condition, at higher levels of acoustic clarity, the auditory modality contains more reliable cues to the manner feature.

For place of articulation, there was a main effect of modality [$\chi^2(1) = 251.617$, $p < 0.001$] as well as a main

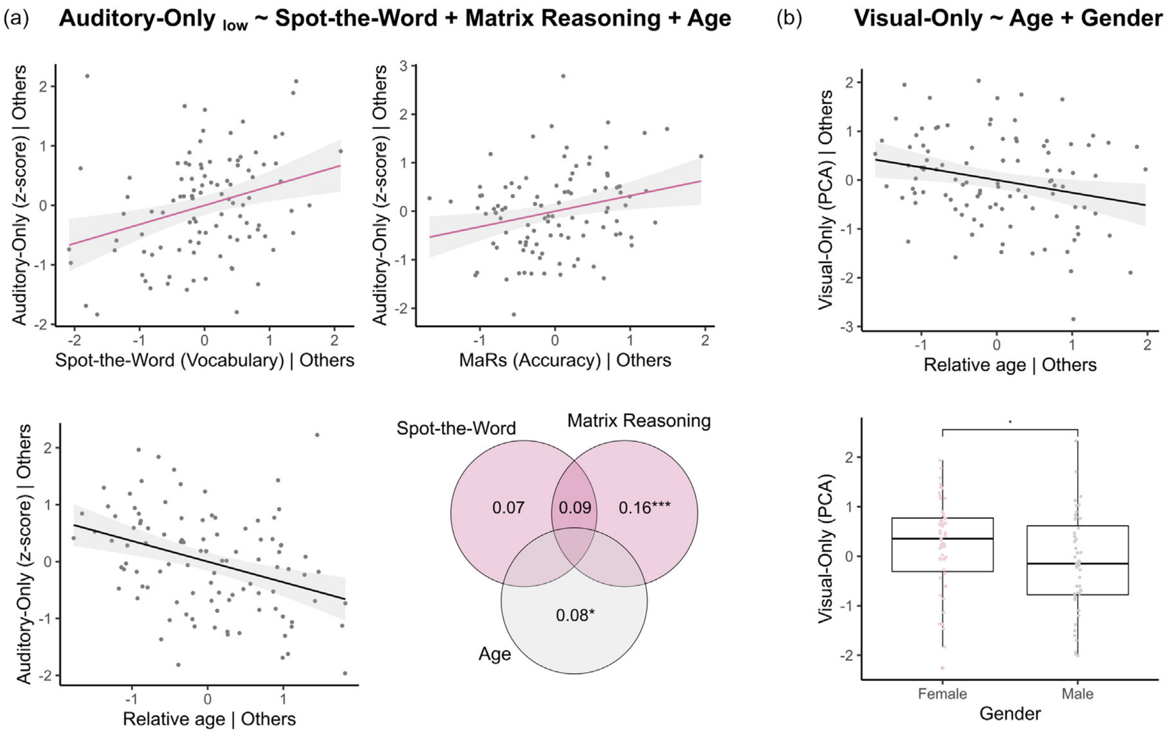

FIG. 6. Partial regression plots for cognitive and demographic predictors of unimodal speech perception. (a) Partial regression plots and variance partitioning results for the regression analysis predicting AO_{low}. (b) Predictors of VO speech perception.$^{***}p < 0.001$ and $^{*}p < 0.05$.

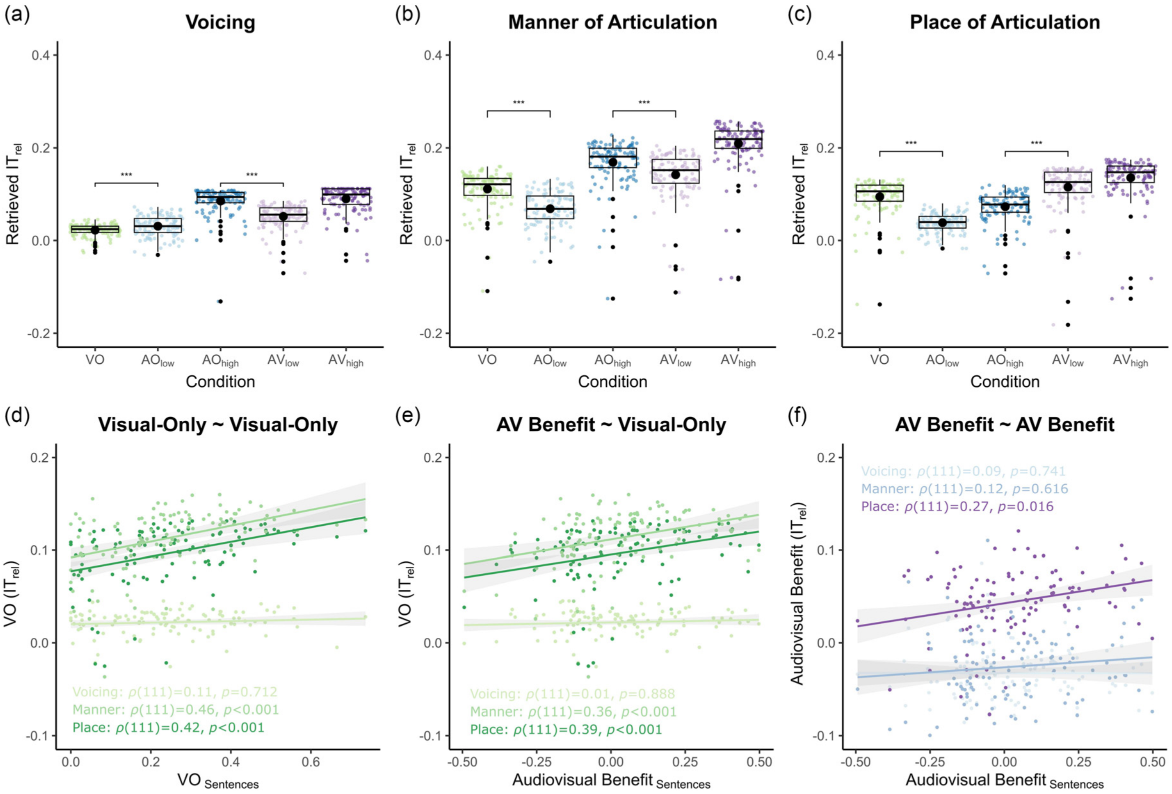

FIG. 7. Results of the FITA. Retrieved values reflect relative information transmitted for (a) voicing, (b) manner of articulation and (c) place of articulation features across five conditions, including significance levels for pairwise comparisons for conditions of interest, $^{***}p < 0.001$. (d) Correlations of sentence-level lipreading ability (accuracy in the VO condition) with visual transmission of voicing, manner, and place of articulation features; (e) relationship of sentence-level audiovisual benefit with visual feature transmission; and (f) correlation of audiovisual benefit calculated using retrieved IT_{rel} with sentence-level audiovisual benefit are shown. Correlations are corrected for multiple comparisons using the Holm adjustment.

effect of clarity [$\chi^2(1) = 37.419$, $p < 0.001$; see Fig. 7]. Transmission was lowest in the AO_{low} condition, which was significantly worse than transmission in the VO condition according to a Wilcoxon sign-rank test ($MD = 0.0597$, $z = 177$, $p < 0.001$). Finally, transmission of the place feature was better in the AV_{low} than in the AO_{high} condition ($MD = 0.0467$, $z = 513$, $p < 0.001$), indicating an overall advantage of the visual modality for transmitting place of articulation information. As for voicing, this is largely consistent with the existing literature, showing that visual speech provides valuable cues to place of articulation. Previous studies have pointed to the importance of place of articulation extraction for audiovisual speech perception (e.g., Grant *et al.*, 1998). For instance, ability to extract place information is a significant predictor of individual susceptibility to the McGurk effect (Brown *et al.*, 2018; Strand *et al.*, 2014)

Finally, we explored a possible relationship between our retrieved IT_{rel} values for lipreading ability (VO) and audiovisual benefit [retrieved $IT_{rel}(AV_{low})$ − retrieved $IT_{rel}(AO_{high})$] at the consonant-level for each feature and sentence-level lipreading and audiovisual benefit measures [see Figs. 7(d)–7(f)]. These analyses suggested that the ability to extract manner and place of articulation features in the visual modality predicted individual differences in sentence-level lipreading ability [manner, $\varrho(111) = 0.46$, $p < 0.001$; place, $\varrho(111) = 0.42$, $p < 0.001$] and audiovisual benefit [manner, $\varrho(111) = 0.36$, $p < 0.001$; place, $\varrho(111) = 0.39$, $p < 0.001$], whereas the visual transmission of voicing did not explain any variability in either measure at the sentence-level [VO, $\varrho(111) = 0.11$, $p = 0.712$; benefit, $\varrho(111) = 0.01$, $p = 0.888$]. Furthermore, the relative transmission of place of articulation information in the matched AV_{low} and AO_{high} conditions was meaningfully related to individual differences in sentence-level audiovisual benefit [$\varrho(111) = 0.27$, $p = 0.016$], whereas this was not the case for audiovisual benefit for either voicing [$\varrho(111) = 0.09$, $p = 0.741$] or manner [$\varrho(111) = 0.12$, $p = 0.616$] feature transmission. Therefore, despite the small number of presentations that our estimates are based on, we find a reliable relationship between perception of consonantal place and manner features and sentence-level measures of individual differences in visual and audiovisual speech. Overall, these results indicate that ability to extract manner and place of articulation cues visually and the ability to extract place information audiovisually, relative to a participant's AO performance when identifying individual consonants, is related to audiovisual benefit for sentence-level speech.

IV. DISCUSSION

Not all listeners can benefit equally from visual information to enhance speech perception (Grant *et al.*, 1998). Here, we investigated individual differences in audiovisual speech perception using a matched, intermediate-intelligibility measure of audiovisual benefit. Macleod and Summerfield (1987) similarly compared SRTs at 50% accuracy in AO and AV conditions, respectively. Measuring the relative intelligibility of matched AO and audiovisual speech, rather than comparing changes in intelligibility due to added visual cues better avoids floor and ceiling effects and confirms that audiovisual benefit is stable across time. Crucially, unlike previous studies using more conventional visual enhancement measures (Grant *et al.*, 1998; Sommers *et al.*, 2005; Tye-Murray *et al.*, 2010), we found that this audiovisual benefit measure is correlated across different speech materials (sentences, words, and consonants), suggesting that audiovisual integration relies on common mechanisms across levels of linguistic structure.

Isolating participant-level variability across levels of linguistic structure, we found that individual differences in audiovisual benefit were predicted by perceptual rather than cognitive abilities: Better lipreading abilities and higher DiN SRTs (relatively poorer hearing) independently predicted enhanced audiovisual benefit. Conversely, unimodal speech perception was associated with cognitive measures (matrix reasoning and vocabulary) and demographic variables (age and gender). Using information transmission analyses, we further showed that visual speech perception and audiovisual benefit for sentence perception are predicted by individual differences in the perception of place of articulation (and to a lesser-degree, manner of articulation) features during a consonant identification task. These findings point to common speech perception mechanisms that support audiovisual benefit in speech listening.

A. A common mechanism underlying audiovisual benefit across levels of linguistic structure

In the present study, we find reliable correlations for our measure of audiovisual benefit across speech materials probed at different levels of linguistic structure. That is, the degree of benefit obtained at the level of minimal syllables predicts the relative magnitude of benefit obtained at the level of monosyllabic words and meaningful sentences (each of which were additionally produced by a different speaker). Previous work has most commonly not been able to establish such a relationship, for example, using the visual enhancement (VE) measure (Grant and Seitz, 1998; Sommers *et al.*, 2005). Grant *et al.* (1998) found no reliable correlations between consonant- and sentence-level VE in older HI listeners, whereas Sommers *et al.* (2005) found only one moderate correlation between word- and sentence-level VE in younger, NH but not older (NH and HI) listeners, and no statistically reliable association could be established for consonant-level VE to higher-level measures. These limited or null findings for AV speech are surprising given that in unimodal conditions, similar cross-task correlations are typically reliable (Bernstein *et al.*, 2000; Grant *et al.*, 1998; Humes *et al.*, 1994; Sommers *et al.*, 2005).

A potential explanation for this lack of correlations proposed previously (Sommers, 2021; Sommers *et al.*, 2005; Van Engen *et al.*, 2017) is that audiovisual integration for speech perception may rely on different mechanisms across levels of linguistic structure. In a multistage model of audiovisual speech perception (Peelle and Sommers, 2015), mechanisms relying on the complementarity of audiovisual information at the level of phonetic features (e.g., Summerfield *et al.*, 1997) or whole words (e.g., auditory and visual neighbourhoods, Tye-Murray *et al.*, 2007b) could be differentially engaged depending on the linguistic complexity of the speech materials presented. This account might also extend to sentence perception—for example, if visually mediated cortical entrainment is a mechanism that enhances sensitivity to upcoming, quasi-rhythmic continuous speech (Peelle and Sommers, 2015), this might not easily apply to isolated syllables or single words.

However, other speech perception mechanisms—e.g., predictive processing of mouth-leading speech—more plausibly operate at multiple levels of linguistic structure (Chandrasekaran *et al.*, 2009; Karas *et al.*, 2019). Mouth-leading speech refers to cases in which visual cues precede corresponding acoustic speech in time. For example, when articulating the phoneme "m," a preparatory gesture of closing the lips can provide a visual speech cue before auditory cues to place are apparent. This is sometimes referred to as a "visual speech head start" (Karas *et al.*, 2019) that may facilitate audiovisual speech perception (van Wassenhove *et al.*, 2005). However, the frequency of these mouth-leading events in natural speech remains unclear (Schwartz and Savariaux, 2014). By this view, perception of visual articulation activates phonological representations, which can support speech perception when auditory cues are degraded or absent. This shared mechanism, relying on simple phonetic representations, may explain common sources of variability that we observe when combining multiple levels of linguistic structure and our finding of a link between perception of consonantal features and audiovisual benefit. That our observations also generalise across different speakers (which may introduce additional noise in across-task comparisons, e.g., Hazan *et al.*, 2010; Heald and Nusbaum, 2014) is striking and suggests that similar effects might also be observed in ecological listening situations.

Of course, the amount of lexical and semantic context available to listeners may impact speech recognition in unimodal and audiovisual conditions (e.g., Iverson *et al.*, 1998; Smayda *et al.*, 2016). In our work, this is evident in the pilot data, explaining why our intermediate conditions of interest are created using different levels of acoustic clarity. Measuring audiovisual benefit based on matched, intermediate-intelligibility conditions, thus, alleviates intelligibility confounds in comparing across modalities and tasks. Unlike other measures used in studying audiovisual speech perception, our intelligibility-matched measure of AV benefit shows moderate or good test-retest reliability. Previous work has noted the apparent lack of success of audiovisual speech training at the group level (e.g., Preminger and Ziegler, 2008) and commented that audiovisual benefit may implicitly be assumed to be stable within an individual. However, this assumption has not explicitly been tested, especially in research specifically designed to investigate individual differences (Grant *et al.*, 1998; Sommers *et al.*, 2005; Tye-Murray *et al.*, 2007a; Tye-Murray *et al.*, 2016). Here, we show that our measure of individual differences generalises across time when participants are tested on different items and, furthermore, exhibit reliable cross-task correlations, even for tests assessing different levels of linguistic structure, with measures adjusted for within-task test-retest reliability. This approach, of (a) estimating audiovisual benefit at comparable, intermediate levels of acoustic clarity across materials, (b) avoiding floor and ceiling effects, (c) estimating test-retest reliability, and (d) taking task reliability into account for cross-task correlations was successful in showing consistent AV benefits for speech materials. We, therefore, encourage future studies of audiovisual speech training to consider the methods proposed here when testing for changes in the use of visual speech to support degraded speech perception.

A natural next step for this work will be to test whether individual differences in AV benefit are similarly apparent in HI individuals. However, in populations with more variable hearing abilities (e.g., in HL, where acoustic degradation levels similar to those used here are likely to introduce floor effects), we recommend the investigation of individual speech perception thresholds determined using an adaptive procedure to quantify the relative visual benefit (Macleod and Summerfield, 1987). Alternatively, researchers should consider testing a range of levels (guided by pilot experiments) rather than limiting their experiment to one or two levels of degradation determined *a priori*. Where automatic scoring methods are more challenging—for example, when working with sentence-level word reports tasks in online experiments (but see Borrie *et al.*, 2019; and Bosker, 2021, for recent advances in this area) —sampling at multiple levels would present an alternative option to prevent floor and ceiling effects.

B. The role of cognitive, perceptual, and demographic variables in explaining individual differences

Having confirmed that individual differences in audiovisual benefit are stable over time and consistent across levels of linguistic structure, we set out to investigate the role of cognitive, perceptual, and demographic variables in explaining these individual differences. Importantly, we do not attempt to isolate the integration stage here but instead take a holistic approach to understanding individual differences in audiovisual speech benefit across levels of linguistic complexity. This represents a more ecological approach: assessing audiovisual speech perception, in general, rather than understanding audiovisual integration as a discrete, separable part of the process. To understand the role of linguistic and cognitive abilities as well as auditory perceptual acuity, we administered several well-established psychometric tests in our final session. We also recruited a balanced sample across the adult age range (18–60 years old, mean = 38 years old), and recorded participant's self-identified gender.

Whereas auditory speech perception was predicted by demographic (age and gender) and domain-general cognitive abilities, we found that only unimodal perceptual abilities predicted individual differences in audiovisual benefit. This is in line with previous research: It is well-established that cognitive abilities correlate with performance on speech perception tasks (especially at the sentence-level, Heinrich *et al.*, 2015), and auditory speech perception and lipreading ability decline with age (Tye-Murray *et al.*, 2010; Tye-Murray *et al.*, 2016). A common finding in previous work is a lipreading advantage for female participants (which may be due to differences in strategy or gaze behaviour, e.g., see Bernstein, 2018), which we also find here. Finally, the idea that individual differences in audiovisual enhancement is a consequence of differences in unimodal perceptual abilities has been suggested previously (Sommers, 2021; Tye-Murray *et al.*, 2016). We extend these previous findings by using a number of speech task-external measures, capturing different aspects of auditory perceptual acuity. Additionally, we explored the role of phonetic feature information across modalities and how individual variability in their transmission at the consonant-level generalises to higher levels of linguistic structure. We will discuss our findings with regard to each of these factors in turn.

1. Cognitive and linguistic abilities

Our results suggest that measures of language and domain-general cognition are associated with individual differences in auditory but not visual speech perception or audiovisual benefit. This is in line with previous work: It is well-established that individual differences in cognitive measures and language proficiency predict performance on AO speech recognition tasks, even after accounting for individual differences in audibility in consonant-, word-, and sentence-level tasks (Akeroyd, 2008; Besser *et al.*, 2013; Humes *et al.*, 1994; Moradi *et al.*, 2013; Moradi *et al.*, 2014). Although more specific measures have previously been linked to speech recognition (specifically, measures of working memory such as *n*-back tasks; see Besser *et al.*, 2013), we find that shared, domain-general, variance between our cognitive and linguistic tasks predicts better AO performance across levels of linguistic structure. Specifically, after accounting for the unique variance in the MaRs and shared variance with the STW vocabulary measure, vocabulary knowledge itself no longer significantly explained variability in AO performance. This finding could reflect influences of fluid intelligence on performance in working memory tasks (Harrison *et al.*, 2015; Wiley *et al.*, 2011) or the influence of domain-general neural mechanisms shown by impaired perception of degraded speech in individuals with brain lesions affecting fluid intelligence (MacGregor *et al.*, 2022). Alternatively, scoring word report tasks using more granular string-matching metrics (Bosker, 2021) might have attenuated the influence of linguistic knowledge on relative performance (Stevenson *et al.*, 2015).

Previous studies, however, have been less clear on whether cognitive and linguistic abilities predict individual differences in audiovisual enhancement. One aspect of this debate concerns whether the addition of visual speech leads to increased or decreased computational demands in speech recognition tasks (Fraser *et al.*, 2010; Moradi *et al.*, 2013; Moradi *et al.*, 2017). In a dual-task paradigm, Fraser *et al.* (2010) found that compared to intelligibility-matched AO speech, performance in an audiovisual speech recognition task was more disrupted by the presence of a secondary task. Like our intelligibility-matched method, this approach avoids a pitfall of traditional methods based on comparing conditions where the audiovisual task is naturally more intelligible, i.e., less cognitively demanding. However, for Fraser *et al.* (2010), this effect was only apparent in RTs but not in accuracy scores or subjective listening effort ratings. In the current study, we also compare matched conditions to avoid intelligibility-related confounds and found no evidence of any relationship between audiovisual benefit and domain-general cognitive or linguistic abilities.

It might be that variability in visual speech perception, and by extension, audiovisual benefit, relies on more domain-specific cognitive abilities such as visuo-verbal or visuospatial working memory and processing speed, which had previously been linked to lipreading ability (Feld and Sommers, 2009; Lyxell and Holmberg, 2000; Tye-Murray *et al.*, 2014). However, here, we found no evidence of a relationship to either of the MaRs measures, suggesting that—to the extent that processing speed and perceptual synthesis (Watson *et al.*, 1996) are measured by this nonverbal reasoning task—then neither was related to lipreading ability or audiovisual benefit. Another explanation of the somewhat conflicting findings in the literature could be that proposals tying (audio)visual speech perception to higher-level cognitive and linguistic abilities might be specific to research with special populations (i.e., school-aged children or individuals with HL that occurred early in life; Lyxell and Holmberg, 2000; Lyxell and Rönnberg, 1989). More recent work generally confirms the idea that individual differences in cognitive or linguistic abilities are not correlated with audiovisual enhancement, even in these populations (Lalonde and McCreery, 2020). In a sample of school-aged children with and without HL, Lalonde and McCreery (2020) found no relationship of vocabulary, working memory, and executive function measures with audiovisual enhancement in a sentence-recognition task: Only the degree of HL predicted the magnitude of audiovisual enhancement, which is consistent with our findings.

2. Unimodal speech perception abilities

We found that relatively poorer hearing and better lipreading ability independently predicted individual differences in audiovisual benefit across levels of linguistic structure. As expected (Bernstein *et al.*, 2022, for review), lipreading ability itself accounted for a large amount of the variance in audiovisual benefit. The idea that variability in unimodal perceptual abilities explains individual differences in the audiovisual speech advantage is not a new proposition: Tye-Murray *et al.* (2016), for example, conducted a PCA on a closed-set word identification task in 11 conditions of AO, VO, and AO speech, which returned only two components, suggesting that variability was entirely explained by two unimodal variability factors rather than requiring a third, distinct integration ability. Importantly, when using the term "perceptual" here, we refer to "speech perception," which involves modality-specific phonetic categorisation (e.g., Holt and Lotto, 2010). Our use of this term is, therefore, not limited to prelinguistic perceptual processes. This is to distinguish accounts which consider AV integration to be a distinct ability (similar to working memory or processing speed, as addressed in Tye-Murray *et al.*, 2016), which might be associated with supramodal cognitive abilities.

A key result from our regression analysis is that variability in SRTs estimated in the DiN test (Smits *et al.*, 2013) predicted individual differences in audiovisual benefit. It is well-established that hearing impairment is associated with improved lipreading ability (likely the result of early developmental experiences, Auer and Bernstein, 2007; Bernstein *et al.*, 2000; Tye-Murray *et al.*, 2014). Older adults with age-related HL generally also show increased audiovisual enhancement (Altieri and Hudock, 2014; Moradi *et al.*, 2017; Puschmann *et al.*, 2019, but see Rosemann and Thiel, 2018; Spehar *et al.*, 2008; and Tye-Murray *et al.*, 2007a). Because we find that mild differences in hearing predict audiovisual benefit *independently* of lipreading ability, we interpret this in line with a re-weighting of visual perceptual cues during audiovisual speech processing as information conveyed through the auditory modality becomes less reliable (even though visual cues in isolation are not necessarily more reliably identified, as we find no evidence of a relationship here). This is in line with causal inference models of audiovisual perception (Körding *et al.*, 2007; Ma *et al.*, 2009), whereby the sensory uncertainty introduced by poorer hearing induces shifts in perceptual weighting. For example, relatedly, a recent study has suggested that children with developmental dyslexia (DD) may increasingly rely on the visual modality to compensate for (auditory) phonological processing difficulties compared to children without DD (Gijbels *et al.*, 2024).

Of course, the cross-sectional nature of our study and lack of longitudinal data limits the strength of the conclusions that we can draw. We do not find, for instance, that individual differences in our hearing measures are correlated with age and, thus, cannot draw any conclusions regarding the onset and length of relative difficulties in speech-in-noise perception or, by extension, any role of cross-modal plasticity, which has been proposed to underlie increased audiovisual enhancement in age-related HL (Campbell and Sharma, 2014; Puschmann and Thiel, 2017). Nonetheless, it is interesting that even mild differences in speech-in-noise perception are reliably associated with enhanced audiovisual benefit.

At the same time, we find no evidence that audiovisual benefit is related to subjective experiences of hearing impairment (which may result in intentional changes in gaze behaviour, e.g., Rennig *et al.*, 2020) or phonetic perceptual gradiency specifically (see also Brown *et al.*, 2018, for a similar lack of evidence that categorical perception accounts for individual differences in the McGurk effect). It might be that additional self-report measures, such as the more extensive speech, spatial, and qualities of hearing scale (SSQ; Gatehouse and Noble, 2004), would provide a more reliable score. Alternatively, perhaps such mild differences in hearing are not subjectively noticeable by participants. Suess *et al.* (2022) found that subjective hearing impairment measured using the APHAB predicted enhanced lipreading abilities in a sample including participants with moderate HL.

To better understand the role of phonetic feature recognition in explaining variability in lipreading ability and audiovisual benefit, we conducted exploratory information transmission analyses.Overall, our results are in line with previous work (Grant *et al.*, 1998; Lalonde and Werner, 2019) in showing an auditory advantage for voicing and manner, as well as increased transmission of place information in the visual modality, and we also successfully identify a directly relationship between phonetic feature recognition and sentence-level listening benefit. We find that while place of articulation is predominantly transmitted through the visual modality, manner is more easily transmitted through the auditory modality, which is in line with the classic VPAM framework (Binnie *et al.*, 1974; Summerfield, 1979), emphasising the complementarity of visual and auditory cues to phonetic perception. This is also in agreement with what we know from incongruent contexts: In minimal syllables, Lalonde and Werner (2019) showed that consonant identification was most likely determined by auditory manner and voicing information or visual place information. Interestingly, we see a main effect of added visual speech, even for voicing, for which transmission in the visual modality alone was negligible. We explain this effect, which is super-additive in nature, via temporal cues to voice-onset time (Raphael, 1972, 1975), transmitted in combination by combining visual cues for the timing of stop-release with auditory cues to voicing.

In line with the VPAM framework, Grant *et al.* (1998) set out to show that cue complementarity, i.e., the ability to extract manner information in the auditory modality, and place information in the visual modality explains a significant amount of variance in individual differences in audiovisual speech perception. Although this was true for nonsense syllables, it failed to generalise to their measure of audiovisual enhancement at the level of sentences. By contrast, we found that ability to extract manner and place of articulation cues positively predicted individual differences in lipreading ability and audiovisual benefit at the sentence-level. This suggests that better perception of place and manner cues in visual speech (independent of speech perception differences in AO conditions) generalises to improved lipreading and the use of visual cues at the sentence-level. Interestingly, only for place information did relative feature transmission at the consonant-level ($AV_{low} > AO_{high}$) predict individual differences in sentence-level audiovisual benefit. This suggests that the complementary nature of visual relative to auditory cue extraction for place of articulation plays a role in an individual's ability to benefit from visual cues in more ecological listening conditions. Our findings, therefore, confirm that place of articulation perception is an important avenue for audiovisual speech rehabilitation. Item transmission analysis is usually performed on datasets containing a large number of presentations in a small sample (e.g., Lalonde and Werner, 2019, $n = 9$). Here, we show that based on only two presentations of each item, we can retrieve sufficiently reliable measures of individual differences in feature transmission to be predictive of audiovisual performance in more ecological speech stimuli.

3. Demographic variables and clinical implications

As expected, in our speech perception task, we replicate the well-documented age-related decline in auditory speech perception (Füllgrabe *et al.*, 2015; Gordon-Salant, 2014) and lipreading (Feld and Sommers, 2009; Tye-Murray *et al.*, 2007a; Tye-Murray *et al.*, 2016). A combination of sensory, perceptual, and cognitive changes likely contribute to this effect (Füllgrabe *et al.*, 2015; Roberts and Allen, 2016). In our sample, older adults were no more likely to have poorer speech reception thresholds than younger adults and did not provide subjective reports of a higher incidence of hearing difficulties. This may be because we intentionally recruited participants without known hearing difficulties and recruited a younger sample (aged $\leq$ 60 years old) than studies explicitly aimed at investigating age-related changes in (audiovisual) speech perception. This may also explain why we found matrix reasoning performance to remain largely intact in older individuals, whereas previous studies including older adults reported a linear decline in scores (Der *et al.*, 2009; Salthouse, 1993). However, as expected, linguistic skills (performance in the STW task) improved with age (Hartshorne and Germine, 2015). The age-related decline in unimodal speech perception that we observe here may, therefore, be a combined consequence of perceptual and cognitive changes throughout the lifespan, including changes to more domain-specific cognition, for instance, working memory. Füllgrabe *et al.* (2015), for example, found that performance in digit span tests accounted for some age-related deficits in speech-in-noise identification).

Previous suggestions that deficits in unimodal speech perception could be compensated for by an audiovisual integration capacity (Freiherr *et al.*, 2013; Laurienti *et al.*, 2006), in agreement with the principle of inverse effectiveness (Stein and Meredith, 1993), have been of great clinical interest. However, this proposal has not been substantiated by the literature on audiovisual speech perception (Spehar *et al.*, 2008; Stevenson *et al.*, 2015; Tye-Murray *et al.*, 2007a; Winneke and Phillips, 2011, but see Dias *et al.*, 2021) or found in the current study. We found no age-related changes in audiovisual benefit. At similar, intermediate levels of intelligibility for AO and audiovisual speech, it seems, therefore, that unlike unimodal speech perception, audiovisual benefit is generally preserved with age.

Understanding determiners of individual differences in lipreading and audiovisual benefit can help to further our comprehension of the mechanisms underlying audiovisual speech perception (Kidd *et al.*, 2018) and also identify potential rehabilitative strategies to restore speech communication in HL by helping us grasp what participants with better recognition "do right." Our results support the notion that improvements in visual phonemic perception can have substantial positive implications for sentence-level speech perception. Recent advances in lipreading training are especially promising in this regard, even though lipreading training has traditionally been notoriously challenging (Bernstein *et al.*, 2022; Bernstein *et al.*, 2023). Files *et al.* (2015) suggest that while sub-visemic contrasts are not usually processed, they are available to participants, i.e., NH adults can discriminate between phonemes that are usually grouped into the same viseme class, such as /ʒa/ and /da/, suggesting this as a potential avenue for training, which can generalise to natural speech (e.g., Schmitt *et al.*, 2023). Additionally, lipreading training targeted specifically at the phonemic contrasts that are increasingly degraded in the auditory modality may be especially beneficial in supporting speech communication in multimodal environments.

Additionally, in our study, we observed a slight lipreading advantage for female participants, which had been reported previously (Bernstein, 2018; Johnson *et al.*, 1988; Watson *et al.*, 1996); but see Auer and Bernstein, 2007 and Tye-Murray *et al.*, 2007a). The source of these gender differences and whether they can provide insights for potential avenues for rehabilitation, however, remain elusive. Bernstein (2018) speculates that differences in response strategies—specifically, increased guessing—may underlie better lipreading scores. Gender differences in face processing may also underlie this effect: Women tend to rate faces as more salient (Proverbio, 2017), which may, in turn, affect face-viewing behaviour, which is in line with the idea that "social tuning," a measure of the frequency of mouth and eye fixations, predicts enhanced visual speech identification in children with and without HL (Worster *et al.*, 2018). When speech is degraded, participants have a general tendency to fixate the mouth (Rennig *et al.*, 2020), therefore, simply encouraging mouth-looking behaviour in adults is unlikely to make a sustained difference. Overcoming selection bias (Awh *et al.*, 2012) toward prioritising the encoding of social rather than phonetic information from the face may also play a role (Bernstein *et al.*, 2023). Finally, the effect of visual acuity on audiovisual speech perception remains unclear and understudied, despite well-documented age-related declines in visual abilities (Andersen, 2012). Mild differences in general visual acuity in older adults do not seem to predict audiovisual speech (Hickson *et al.*, 2004), whereas early visual deprivation in congenital cataract patients permanently impairs lipreading ability (Putzar et al., 2010). In our study, we do not explicitly investigate the role of visual acuity and its relationship with lipreading ability and audiovisual speech perception. Future work employing a similar paradigm, measuring the relative intelligibility of auditory and audiovisual speech, combined with assessment of domain-general visual abilities, would be well-suited to address this question.

4. Effects of different types of acoustic clarity manipulations

In our work, we used a form of artificially degraded speech—noise-vocoded speech—that allows for careful matching of intelligibility (necessary for our comparison of visual and audiovisual speech perception) and provides an approximate simulation of speech transduced by a cochlear implant (Shannon *et al.*, 1995). However, our use of this form of artificial degradation leaves unanswered important questions concerning the relationship between our findings and effects of BN or competing talkers on audiovisual speech perception in ecological listening situations. Visual speech can (a) provide information about the content of speech (object formation) and (b) aid the listener in segregating target speech from BN or distractor speakers (object selection; e.g., Devergie *et al.*, 2011). Our use of noise-vocoded speech focused predominantly on the first case, whereas different types of BN may introduce additional task demands related to sound source segregation and selective attention to the target sound source.

It is important to consider whether our results might also apply to more typical listening challenges such as speech in noise. It is possible that audiovisual integration for speech perception may not follow the same principles when it is needed to segregate target speech from BN (Blackburn *et al.*, 2019; Micula *et al.*, 2024). Future work is needed to determine how demands related to object selection in ecological settings are affected by the presence of visual speech in listeners with HL or CIs—and its association with supramodal abilities, including attention. However, energetic masking introduced by BN or distractor speakers also leads to the physical degradation of the target signal (Brungart, 2001), which can be compensated for by visual speech. For this aspect (i.e., object formation), we believe that our results should hold true independent of the type of auditory manipulation. Additionally, our work is not inconsistent with conclusions drawn from studies investigating audiovisual speech-in-noise perception (see Sommers, 2021, for review). Importantly, ceiling effects can introduce confounds when studying predictors of individual differences regardless of the type of acoustic manipulation. It is, therefore, an important detail that floor/ceiling effects did not contribute to our critical audiovisual benefit measure.

V. CONCLUSION

Substantial individual differences in lipreading and audiovisual speech perception exist in the general population. Hence, only for some can the negative consequences of HL be alleviated substantially via increased reliance on visual cues. In our study, added visual cues improved speech scores at the sentence-level by an average of 36% (AV > AO) with a range of 0%–86%. However, rather than investigating individual differences using these estimates, we employed the relative benefit obtained by comparing matched, intermediate-intelligibility AO and AV speech ($AV_{low} > AO_{high}$). We found that this measure of audiovisual benefit was stable across sessions and correlated across speech materials: meaningful sentences, monosyllabic words, and minimal syllables. This suggests that if we avoid intelligibility confounds, we find evidence that audiovisual speech benefit relies on a shared mechanism across levels of linguistic structure. Information transmission analysis suggested that even at the sentence-level, audiovisual benefit relies fundamentally on the ability to perceive simple articulatory features (such as place of articulation) in visual speech.

Additionally, we found that individual differences in audiovisual benefit were predicted by better lipreading ability and subclinical indicators of poorer hearing (speech reception thresholds in the DiN task). Overall, this is in agreement with the idea that variability in unimodal perceptual abilities underlies individual differences in audiovisual speech processing. Future research exploring how best to support older individuals with declining hearing would be served best by focussing on supporting their declining unimodal perceptual abilities (specifically those most relevant in ecological, multimodal contexts as can be identified using, for example, information transmission analysis). Rather than resulting from a simple linear combination of auditory and visual speech perception skills (e.g., Tye-Murray *et al.*, 2016), however, it seems that individuals with mild speech-in-noise recognition difficulties are more adept at using visual cues in audiovisual context. This was independent of any improvements in lipreading ability. We interpret this in line with a causal inference framework, which has previously been applied to explain perception of incongruent AV stimuli. Although we do not find any other behavioural correlates of this enhanced audiovisual benefit, our conclusions are, perhaps, limited by the cross-sectional nature of this study. Future research adopting a longitudinal approach or carefully controlled and validated neuroimaging measures (considering intelligibility explicitly as a potentially confounding variable) may be better suited to identifying strategies to aid multimodal speech communication in individuals with HL.

SUPPLEMENTARY MATERIAL

See the supplementary material for further details on and summary statistics of consonant- and word stimuli, an illustration of the procedure, and detailed statistics on Secs. III C and III D including model comparisons performed via stepwise deletion. Table S1 cites Miller and Nicely (1955) and Jesse and Massaro (2010). Table S2 cites Balota *et al.* (2007), Lund and Burgess (1996), Luce and Pisoni (1998), Kuperman *et al.* (2012), and Krason *et al.* (2023b).

ACKNOWLEDGMENTS

This research was funded by a Medical Research Council Doctoral Training Partnership (MRC DTP) Studentship (Reference No. MR/N013433/1) and a Cambridge Trust Scholarship to J.v.S., and MRC funding of the Cognition and Brain Sciences Unit (Reference No. MC_UU_00030/6, supporting M.H.D. and M.A.). The authors would like to thank Thu Ngan Dang, Shangqiguo Wang, and Tobias Goehring for providing stimuli for the DiN task; Anna Krason, Ye Claudia Zhang, Gabriella Vigliocco, Elizabeth Buchanan-Worster, and Mairéad MacSweeney for sharing video recordings for the speech stimuli. Finally, we would like to thank Lucy MacGregor, Adam Attahari, and Harriet J. Smith for sharing jsPsych scripts for the STW and listen-up tasks. Further details on the speech stimuli used in this study are available in the original publications for which they were recorded. Sentence-level stimuli have been made available by Aller *et al.* (2022).[5] Clear speech recordings from which our word-level stimuli were drawn (Krason *et al.*, 2023a; Krason *et al.*, 2023b) are available online.[6] Examples of consonant-level stimuli have been printed in Pimperton *et al.* (2019) and Buchanan-Worster *et al.* (2021).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to declare.

Ethics Approval

Ethics approval was obtained by the Cambridge Psychology Research Ethics Committee (Application No. PRE.2022.056).

DATA AVAILABILITY

The data that support the findings of this study are openly available in https://osf.io/j56y4/ at http://doi.org/10.17605/OSF.IO/J56Y4 (von Seth, 2024).

[1] See www.prolific.com (Last viewed January 8, 2025).
[2] See https://sites.google.com/site/blakemorelab/research/mars-ib (Last viewed January 8, 2025).
[3] See https://aspredicted.org/34C_Z4L (Last viewed January 8, 2025).
[4] See https://osf.io/j56y4/ (Last viewed January 8, 2025).
[5] See https://osf.io/st6fe/ (Last viewed January 8, 2025).
[6] See https://osf.io/gudj6/ (Last viewed January 8, 2025).

Akeroyd, M. A. (**2008**). "Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults," Int. J. Audiol. **47**, S53–S71.

Aller, M. (**2022**). "Differential auditory and visual phase-locking are observed during audio-visual benefit and silent lip-reading for speech perception," https://doi.org/10.17605/OSF.IO/ST6FE.

Aller, M., Økland, H. S., MacGregor, L. J., Blank, H., and Davis, M. H. (**2022**). "Differential auditory and visual phase-locking are observed during audio-visual benefit and silent lip-reading for speech perception," J. Neurosci. **42**, 6108–6120.

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., and Tiippana, K. (**2014**). "Effect of attentional load on audiovisual speech perception: Evidence from ERPs," Front. Psychol. **5**, 727.

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (**2005**). "Audiovisual integration of speech falters under high attention demands," Curr. Biol. **15**, 839–843.

Altieri, N., and Hudock, D. (**2014**). "Assessing variability in audiovisual speech integration skills using capacity and accuracy measures," Int. J. Audiol. **53**, 710–718.

Andersen, G. J. (**2012**). "Aging and vision: Changes in function and performance from optics to perception," WIRES Cognit. Sci. **3**, 403–410.

Auer, E. T., and Bernstein, L. E. (**2007**). "Enhanced visual speech perception in individuals with early-onset hearing impairment," J. Speech. Lang. Hear. Res. **50**, 1157–1165.

Awh, E., Belopolsky, A. V., and Theeuwes, J. (**2012**). "Top-down versus bottom-up attentional control: A failed theoretical dichotomy," Trends Cogn. Sci. **16**, 437–443.

Baddeley, A., Emslie, H., and Nimmo-Smith, I. (**1993**). "The spot-the-word test: A robust estimate of verbal intelligence based on lexical decision," Br. J. Clinic. Psychol. **32**, 55–65.

Baese-Berk, M. M., Levi, S. V., and Van Engen, K. J. (**2023**). "Intelligibility as a measure of speech perception: Current approaches, challenges, and recommendations," J. Acoust. Soc. Am. **153**, 68–76.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treian, R. (**2007**). "The English Lexicon Project," Behav. Res. Methods **39**, 445–459.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Mem. Lang. **68**, 255–278.

Bernstein, L. E. (**2018**). "Response errors in females' and males' sentence lipreading necessitate structurally different models for predicting lipreading accuracy," Lang. Learn. **68**, 127–158.

Bernstein, L. E., Auer, E. T., and Eberhardt, S. P. (**2023**). "Modality-specific perceptual learning of vocoded auditory versus lipread speech: Different effects of prior information," Brain Sci. **13**, 1008.

Bernstein, L. E., Jordan, N., Auer, E. T., and Eberhardt, S. P. (**2022**). "Lipreading: A review of its continuing importance for speech recognition with an acquired hearing loss and possibilities for effective training," Am. J. Audiol. **31**, 453–469.

Bernstein, L. E., Tucker, P. E., and Demorest, M. E. (**2000**). "Speech perception without hearing," Percept. Psychophys. **62**, 233–252.

Besser, J., Koelewijn, T., Zekveld, A. A., Kramer, S. E., and Festen, J. M. (**2013**). "How linguistic closure and verbal working memory relate to speech recognition in Nnoise—A review," Trends Amplif. **17**, 75–93.

Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (**1974**). "Auditory and visual contributions to the perception of consonants," J. Speech Hear. Res. **17**, 619–630.

Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., and Stacey, P. C. (**2019**). "Visual speech benefit in clear and degraded speech depends on the auditory intelligibility of the talker and the number of background talkers," Trends Hear. **23**, 2331216519837866.

Blamey, P. J., Cowan, R. S., Alcantara, J. I., Whitford, L. A., and Clark, G. M. (**1989**). "Speech perception using combinations of auditory, visual, and tactile information," J. Rehabil. Res. Dev. **26**, 15–24.

Borrie, S. A., Barrett, T. S., and Yoho, S. E. (**2019**). "Autoscore: An open-source automated tool for scoring listener perception of speech," J. Acoust. Soc. Am. **145**, 392–399.

Bosker, H. R. (**2021**). "Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies," Behav. Res. Methods **53**, 1945–1953.

Braida, L. D. (**1991**). "Crossmodal integration in the identification of consonant segments," Q. J. Exp. Psychol. Sect. A **43**, 647–677.

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., and Strand, J. F. (**2018**). "What accounts for individual differences in susceptibility to the McGurk effect?," PLoS One **13**, e0207160.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Buchan, J. N., and Munhall, K. G. (**2011**). "The Influence of selective attention to auditory and visual speech on the integration of audiovisual speech information," Perception **40**, 1164–1182.

Buchanan-Worster, E., Hulme, C., Dennan, R., and MacSweeney, M. (**2021**). "Speechreading in hearing children can be improved by training," Dev. Sci. **24**, e13124.

Campbell, J., and Sharma, A. (**2014**). "Cross-modal re-organization in adults with early stage hearing loss," PLoS One **9**, e90594.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (**2009**). "The natural statistics of audiovisual speech," PLoS Comput. Biol. **5**, e1000436.

Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., and Blakemore, S.-J. (**2019**). "The matrix reasoning item bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults," R. Soc. Open Sci. **6**, 190232.

Cox, R. M. (**1997**). "Administration and application of the APHAB," Hear. J. **50**, 35–48.

Davis, M. H., Evans, S., McCarthy, K., Evans, L., Giannakopoulou, A., and Taylor, J. S. H. (**2019**). "Lexical learning shapes the development of speech perception until late adolescence," PsyArXiv, 10.31234/osf.io/ktsey.

de Leeuw, J. R. (**2015**). "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," Behav. Res. Methods **47**, 1–12.

de Leeuw, J. R., Gilbert, R. A., and Luchterhandt, B. (**2023**). "jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments," J. Open Source Softw. **8**, 5351.

Der, G., Allerhand, M., Starr, J. M., Hofer, S. M., and Deary, I. J. (**2009**). "Age-related changes in memory and fluid reasoning in a sample of healthy old people," Aging Neuropsychol. Cogn. **17**, 55–70.

Devergie, A., Grimault, N., Gaudrain, E., Healy, E. W., and Berthommier, F. (**2011**). "The effect of lip-reading on primary stream segregation," J. Acoust. Soc. Am. **130**, 283–291.

Dias, J. W., McClaskey, C. M., and Harris, K. C. (**2021**). "Audiovisual speech is more than the sum of its parts: Auditory-visual superadditivity compensates for age-related declines in audible and lipread speech intelligibility," Psychol. Aging **36**, 520–530.

Dong, C., Noppeney, U., and Wang, S. (**2024**). "Perceptual uncertainty explains activation differences between audiovisual congruent speech and McGurk stimuli," Hum. Brain Mapp. **45**, e26653.

Dryden, A., Allen, H. A., Henshaw, H., and Heinrich, A. (**2017**). "The association between cognitive performance and speech-in-noise perception for adult listeners: A systematic literature review and meta-analysis," Trends Hear. **21**, 2331216517744675.

Duhachek, A., and Iacobucci, D. (**2004**). "Alpha's standard error (ASE): An accurate and precise confidence interval estimate," J. Appl. Psychol. **89**, 792–808.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (**2009**). "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses," Behav. Res. Methods **41**, 1149–1160.

Feld, J. E., and Sommers, M. S. (**2009**). "Lipreading, processing speed, and working memory in younger and older adults," J. Speech. Lang. Hear. Res. **52**, 1555–1565.

Files, B. T., Tjan, B. S., Jiang, J., and Bernstein, L. E. (**2015**). "Visual speech discrimination and identification of natural and synthetic consonant stimuli," Front. Psychol. **6**, 878.

Fraser, S., Gagné, J.-P., Alepins, M., and Dubois, P. (**2010**). "Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues," J. Speech. Lang. Hear. Res. **53**, 18–33.

Freiherr, J., Lundström, J. N., Habel, U., and Reetz, K. (**2013**). "Multisensory integration mechanisms during aging," Front. Hum. Neurosci. **7**, 863.

Füllgrabe, C., Moore, B. C. J., and Stone, M. A. (**2015**). "Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition," Front. Aging Neurosci. **6**, 347.

Gatehouse, S., and Noble, W. (**2004**). "The speech, spatial and qualities of hearing scale (SSQ)," Int. J. Audiol. **43**, 85.

Gijbels, L., Lee, A. K. C., and Yeatman, J. D. (**2024**). "Children with developmental dyslexia have equivalent audiovisual speech perception performance but their perceptual weights differ," Dev. Sci. **27**, e13431.

Gordon-Salant, S. (**2014**). "Aging, hearing loss, and speech recognition: Stop shouting, I can't understand you," in *Perspectives on Auditory Research*, edited by A. N. Popper and R. R. Fay (Springer, New York), pp. 211–228.

Grant, K. W., and Seitz, P. F. (**1998**). "Measures of auditory–visual integration in nonsense syllables and sentences," J. Acoust. Soc. Am. **104**, 2438–2450.

Grant, K. W., Walden, B. E., and Seitz, P. F. (**1998**). "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration," J. Acoust. Soc. Am. **103**, 2677–2690.

Harrison, T. L., Shipstead, Z., and Engle, R. W. (**2015**). "Why is working memory capacity related to matrix reasoning tasks?," Mem. Cogn. **43**, 389–396.

Hartshorne, J. K., and Germine, L. T. (**2015**). "When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span," Psychol. Sci. **26**, 433–443.

Hausser, J., and Strimmer, K. (**2009**). "Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks," J. Mach. Learn. Res. **10**, 1469–1484.

Hazan, V., Kim, J., and Chen, Y. (**2010**). "Audiovisual perception in adverse conditions: Language, speaker and listener effects," Speech Commun. **52**, 996–1009.

Heald, S. L. M., and Nusbaum, H. C. (**2014**). "Talker variability in audio-visual speech perception," Front. Psychol. **5**, 698.

Hedge, C., Powell, G., and Sumner, P. (**2018**). "The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences," Behav. Res. Methods **50**, 1166–1186.

Heinrich, A., Henshaw, H., and Ferguson, M. A. (**2015**). "The relationship of speech intelligibility with hearing sensitivity, cognition, and perceived hearing difficulties varies for different speech perception tests," Front. Psychol. **6**, 782.

Hickson, L., Hollins, M., Lind, C., Worrall, L., and Lovie-Kitchin, J. (**2004**). "Auditory-visual speech perception in older people: The effect of visual acuity," Aust. N. Z. J. Audiol. **26**, 3–11.

Holt, L. L., and Lotto, A. J. (**2010**). "Speech perception as categorization," Atten. Percept. Psychophys. **72**, 1218–1227.

Humes, L. E., Watson, B. U., Christensen, L. A., Cokely, C. G., Halling, D. C., and Lee, L. (**1994**). "Factors associated with individual differences in clinical measures of speech recognition among the elderly," J. Speech. Lang. Hear. Res. **37**, 465–474.

Huyse, A., Leybaert, J., and Berthommier, F. (**2014**). "Effects of aging on audio-visual speech integration," J. Acoust. Soc. Am. **136**, 1918–1931.

Iverson, P., Bernstein, L. E., and Auer, E. T., Jr. (**1998**). "Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition," Speech Commun. **26**, 45–63.

Jesse, A., and Massaro, D. W. (**2010**). "The temporal distribution of information in audiovisual spoken-word identification," Atten. Percept. Psychophys. **72**, 209–225.

Johnson, F. M., Hicks, L. H., Goldberg, T., and Myslobodsky, M. S. (**1988**). "Sex differences in lipreading," Bull. Psychon. Soc. **26**, 106–108.

Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., and Beauchamp, M. S. (**2019**). "The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech," eLife **8**, e48116.

Kidd, E., Donnelly, S., and Christiansen, M. H. (**2018**). "Individual differences in language acquisition and processing," Trends Cogn. Sci. **22**, 154–169.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (**2007**). "Causal inference in multisensory perception," PLoS One **2**, e943.

Krason, A., Fenton, R., Varley, R., and Vigliocco, G. (**2023a**). "The role of iconic gestures and mouth movements in face-to-face communication," available at http://osf.io/gudj6.

Krason, A., Zhang, Y., Man, H., and Vigliocco, G. (**2023b**). "Mouth and facial informativeness norms for 2276 English words," Behav. Res. Methods **56**, 4786–4801.

Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (**2012**). "Age-of-acquisition ratings for 30,000 English words," Behav. Res. Methods **44**, 978–990.

Lalonde, K., and McCreery, R. W. (**2020**). "Audiovisual enhancement of speech perception in noise by school-age children who are hard of hearing," Ear Hear. **41**, 705–719.

Lalonde, K., and Werner, L. A. (**2019**). "Perception of incongruent audiovisual English consonants," PLoS One **14**, e0213588.

Laurienti, P. J., Burdette, J. H., Maldjian, J. A., and Wallace, M. T. (**2006**). "Enhanced multisensory integration in older adults," Neurobiol. Aging **27**, 1155–1163.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Linares, D., and López-Moliner, J. (**2016**). "quickpsy: An *R* package to fit psychometric functions for multiple groups," R J. **8**, 122–131.

Lisker, L., Liberman, A. M., Erickson, D. M., Dechovitz, D., and Mandler, R. (**1977**). "On pushing the voice-onset-time (Vot) boundary about," Lang. Speech **20**, 209–216.

Luce, P. A., and Pisoni, D. B. (**1998**). "Recognizing spoken words: The neighborhood activation model," Ear Hear. **19**(1), 1–36.

Lund, K., and Burgess, C. (**1996**). "Producing high-dimensional semantic spaces from lexical co-occurrence," Behav. Res. Methods Instrum. Comput. **28**, 203–208.

Lyxell, B., and Holmberg, I. (**2000**). "Visual speechreading and cognitive performance in hearing-impaired and normal hearing children (11–14 years)," Br. J. Educ. Psychol. **70**, 505–518.

Lyxell, B., and Rönnberg, J. (**1989**). "Information-processing skill and speech-reading," Br. J. Audiol. **23**, 339–347.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., and Parra, L. C. (**2009**). "Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space," PLoS One **4**, e4638.

MacGregor, L. J., Gilbert, R. A., Balewski, Z., Mitchell, D. J., Erzinçlioğlu, S. W., Rodd, J. M., Duncan, J., Fedorenko, E., and Davis, M. H. (**2022**). "Causal contributions of the domain-general (multiple demand) and the language-selective brain networks to perceptual and semantic challenges in speech comprehension," Neurobiol. Lang. **3**, 665–698.

Macleod, A., and Summerfield, Q. (**1987**). "Quantifying the contribution of vision to speech perception in noise," Br. J. Audiol. **21**, 131–141.

Magnotti, J. F., Dzeda, K. B., Wegner-Clemens, K., Rennig, J., and Beauchamp, M. S. (**2020**). "Weak observer-level correlation and strong stimulus-level correlation between the McGurk effect and audiovisual speech-in-noise: A causal inference explanation," Cortex **133**, 371–383.

Massaro, D. W., and Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception," J. Exp. Psychol. Hum. Percept. Perform. **9**, 753–771.

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

Micula, A., Holmer, E., Ning, R., and Danielsson, H. (**2024**). "Relationships between hearing status, cognitive abilities, and reliance on visual and contextual cues," Ear Hear. (published online).

Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Moradi, S., Lidestam, B., Danielsson, H., Ng, E. H. N., and Rönnberg, J. (**2017**). "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," J. Speech. Lang. Hear. Res. **60**, 2687–2703.

Moradi, S., Lidestam, B., and Rönnberg, J. (**2013**). "Gated audiovisual speech identification in silence vs. noise: Effects on time and accuracy," Front. Psychol. **4**, 359.

Moradi, S., Lidestam, B., Saremi, A., and Rönnberg, J. (**2014**). "Gated auditory speech perception: Effects of listening conditions and cognitive capacity," Front. Psychol. **5**, 531.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. (**2016**). "PanPhon: A resource for mapping IPA segments to articulatory feature vectors," in *Proceedings of COLING 2016, the 26th International Conference Computational Linguistics Technical Papers*, edited by Y. Matsumoto and R. Prasad, December 11–17, Osaka, Japan (The COLING 2016 Organizing Committee, Osaka, Japan), pp. 3475–3484.

Oosthuizen, D. J. J., and Hanekom, J. J. (**2016**). "Information transmission analysis for continuous speech features," Speech Commun. **82**, 53–66.

Peelle, J. E., and Sommers, M. S. (**2015**). "Prediction and constraint in audiovisual speech perception," Cortex **68**, 169–181.

Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. (**1995**). "How young and old adults listen to and remember speech in noise," J. Acoust. Soc. Am. **97**, 593–608.

Pimperton, H., Kyle, F., Hulme, C., Harris, M., Beedie, I., Ralph-Lewis, A., Worster, E., Rees, R., Donlan, C., and MacSweeney, M. (**2019**). "Computerized speechreading training for deaf children: A randomized controlled trial," J. Speech. Lang. Hear. Res. **62**, 2882–2894.

Preminger, J. E., and Ziegler, C. H. (**2008**). "Can auditory and visual speech perception be trained within a group setting?," Am. J. Audiol. **17**, 80–97.

Proverbio, A. M. (**2017**). "Sex differences in social cognition: The case of face processing," J. Neurosci. Res. **95**, 222–234.

Punch, J. L., Hitt, R., and Smith, S. W. (**2019**). "Hearing loss and quality of life," J. Commun. Disord. **78**, 33–45.

Puschmann, S., Daeglau, M., Stropahl, M., Mirkovic, B., Rosemann, S., Thiel, C. M., and Debener, S. (**2019**). "Hearing-impaired listeners show increased audiovisual benefit when listening to speech in noise," NeuroImage **196**, 261–268.

Puschmann, S., and Thiel, C. M. (**2017**). "Changed crossmodal functional connectivity in older adults with hearing loss," Cortex **86**, 109–122.

Putzar, L., Goerendt, I., Heed, T., Richard, G., Büchel, C., and Röder, B. (**2010**). "The neural basis of lip-reading capabilities is altered by early visual deprivation," Neuropsychologia **48**, 2158–2166.

Raphael, L. J. (**1972**). "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," J. Acoust. Soc. Am. **51**, 1296–1303.

Raphael, L. J. (**1975**). "The physiological control of durational differences between vowels preceding voiced and voiceless consonants in English," J. Phonet. **3**, 25–33.

Rennig, J., Wegner-Clemens, K., and Beauchamp, M. S. (**2020**). "Face viewing behavior predicts multisensory gain during speech perception," Psychon. Bull. Rev. **27**, 70–77.

Revelle, W. (**2024**). "psych: Procedures for psychological, psychometric, and personality research," available at https://cran.r-project.org/web/packages/psych/index.html (Last viewed January 8, 2025).

Richie, C., and Kewley-Port, D. (**2008**). "The effects of auditory-visual vowel identification training on speech recognition under difficult listening conditions," J. Speech. Lang. Hear. Res. **51**, 1607–1619.

Roberts, K. L., and Allen, H. A. (**2016**). "Perception and cognition in the ageing brain: A brief review of the short- and long-term links between perceptual and cognitive decline," Front. Aging Neurosci. **8**, 39.

Rodd, J. M. (**2024**). "Moving experimental psychology online: How to obtain high quality data when we can't see our participants," J. Mem. Lang. **134**, 104472.

Rosemann, S., and Thiel, C. M. (**2018**). "Audio-visual speech processing in age-related hearing loss: Stronger integration and increased frontal lobe recruitment," NeuroImage **175**, 425–437.

Salthouse, T. A. (**1993**). "Influence of working memory on adult age differences in matrix reasoning," Br. J. Psychol. **84**, 171–199.

SeatGeek Inc. (**2014**). "{fuzzywuzzy}: Fuzzy string matching in Python," available at https://github.com/seatgeek/fuzzywuzzy (Last viewed January 8, 2025).

Schmitt, R., Meyer, M., and Giroud, N. (**2023**). "Improvements in naturalistic speech-in-noise comprehension in middle-aged and older adults after 3 weeks of computer-based speechreading training," npj Sci. Learn. **8**, 1–12.

Schwartz, J.-L., and Savariaux, C. (**2014**). "No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag," PLOS Comput. Biol. **10**, e1003743.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Smayda, K. E., Engen, K. J. V., Maddox, W. T., and Chandrasekaran, B. (**2016**). "Audio-visual and meaningful semantic context enhancements in older and younger adults," PLoS One **11**, e0152773.

Smits, C., Theo Goverts, S., and Festen, J. M. (**2013**). "The digits-in-noise test: Assessing auditory speech recognition abilities in noise," J. Acoust. Soc. Am. **133**, 1693–1706.

Smulders, F. T. Y. (**2010**). "Simplifying jackknifing of ERPs and getting more out of it: Retrieving estimates of participants' latencies," Psychophysiology **47**, 387–392.

Sohoglu, E., and Davis, M. H. (**2016**). "Perceptual learning of degraded speech by minimizing prediction error," Proc. Natl. Acad. Sci. U.S.A. **113**, E1747–E1756.

Sommers, M. S. (**2021**). "Santa Claus, the tooth fairy, and auditory-visual integration," in *The Handbook of Speech Perception*, edited by J. S. Pardo, L. C. Nygaard, R. E. Remez, and D. B. Pisoni (Wiley, New York), pp. 517–539.

Sommers, M. S., Tye-Murray, N., and Spehar, B. (**2005**). "Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults," Ear Hear. **26**, 263–275.

Spehar, B. P., Tye-Murray, N., and Sommers, M. S. (**2008**). "Intra- versus intermodal integration in young and older adults," J. Acoust. Soc. Am. **123**, 2858–2866.

Stein, B. E., and Meredith, M. A. (**1993**). *The Merging of the Senses* (MIT Press, Cambridge, MA), 231 pp.

Stevenson, R. A., Nelms, C. E., Baum, S. H., Zurkovsky, L., Barense, M. D., Newhouse, P. A., and Wallace, M. T. (**2015**). "Deficits in audiovisual speech perception in normal aging emerge at the level of whole-word recognition," Neurobiol. Aging **36**, 283–291.

Strand, J., Cooperman, A., Rowe, J., and Simenstad, A. (**2014**). "Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity," J. Speech. Lang. Hear. Res. **57**, 2322–2331.

Suess, N., Hauswald, A., Zehentner, V., Depireux, J., Herzog, G., Rösch, S., and Weisz, N. (**2022**). "Influence of linguistic properties and hearing impairment on visual speech perception skills in the German language," PLoS One **17**, e0275585.

Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.

Summerfield, Q. (**1979**). "Use of visual information for phonetic perception," Phonetica **36**, 314–331.

Summerfield, Q., Bruce, V., Cowey, A., Ellis, A. W., and Perrett, D. I. (**1997**). "Lipreading and audio-visual speech perception," Philos. Trans. R. Soc. London, Ser. B: Biol. Sci. **335**, 71–78.

Themistocleous, C., Neophytou, K., Rapp, B., and Tsapkini, K. (**2020**). "A tool for automatic scoring of spelling performance," J. Speech Lang. Hear. Res. **63**, 4179–4192.

Tillberg, I., Rönnberg, J., Svärd, I., and Ahlner, B. (**1996**). "Audio-visual speechreading in a group of hearing aid users the effects of onset age, handicap age, and degree of hearing loss," Scand. Audiol. **25**, 267–272.

Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., and Sommers, M. S. (**2014**). "Lipreading in school-age children: The roles of age, hearing status, and cognitive ability," J. Speech. Lang. Hear. Res. **57**, 556–565.

Tye-Murray, N., Sommers, M. S., and Spehar, B. (**2007a**). "Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing," Ear Hear. **28**, 656.

Tye-Murray, N., Sommers, M., and Spehar, B. (**2007b**). "Auditory and visual lexical neighborhoods in audiovisual speech perception," Trends Amplif. **11**, 233–241.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., and Hale, S. (**2010**). "Aging, audiovisual integration, and the principle of inverse effectiveness," Ear Hear. **31**, 636–668.

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., and Sommers, M. (**2016**). "Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration," Psychol. Aging **31**, 380–389.

Van den Borre, E., Denys, S., van Wieringen, A., and Wouters, J. (**2021**). "The digit triplet test: A scoping review," Int. J. Audiol. **60**, 946–963.

Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., and Chandrasekaran, B. (**2014**). "Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker," J. Speech. Lang. Hear. Res. **57**, 1908–1918.

Van Engen, K. J., Xie, Z., and Chandrasekaran, B. (**2017**). "Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect," Atten. Percept. Psychophys. **79**, 396–403.

Van Son, N., Huiskamp, T. M. I., Bosman, A. J., and Smoorenburg, G. F. (**1994**). "Viseme classifications of Dutch consonants and vowels," J. Acoust. Soc. Am. **96**, 1341–1355.

van Wassenhove, V., Grant, K. W., and Poeppel, D. (**2005**). "Visual speech speeds up the neural processing of auditory speech," Proc. Natl. Acad. Sci. U.S.A. **102**, 1181–1186.

von Seth, J. (**2024**). "Individual differences in audiovisual benefit for acoustically degraded speech," https://doi.org/10.17605/OSF.IO/J56Y4

Walden, B. E., Prosek, R. A., and Worthington, D. W. (**1975**). "Auditory and audiovisual feature transmission in hearing-impaired adults," J. Speech Hear. Res. **18**, 272–280.

Watson, C. S., Qiu, W. W., Chamberlain, M. M., and Li, X. (**1996**). "Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition," J. Acoust. Soc. Am. **100**, 1153–1162.

Wilbiks, J. M. P., Brown, V. A., and Strand, J. F. (**2022**). "Speech and non-speech measures of audiovisual integration are not correlated," Atten. Percept. Psychophys. **84**, 1809–1819.

Wiley, J., Jarosz, A. F., Cushen, P. J., and Colflesh, G. J. H. (**2011**). "New rule use drives the relation between working memory capacity and Raven's Advanced Progressive Matrices," J. Exp. Psychol. Learn. Mem. Cogn. **37**, 256–263.

Wilkinson, G. N., and Rogers, C. E. (**1973**). "Symbolic description of factorial models for analysis of variance," J. R. Stat. Soc. Ser. C Appl. Stat. **22**, 392–399.

Winneke, A. H., and 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," Psychol. Aging **26**, 427–438.

Woods, K. J. P., Siegel, M. H., Traer, J., and McDermott, J. H. (**2017**). "Headphone screening to facilitate web-based auditory experiments," Atten. Percept. Psychophys. **79**, 2064–2072.

Worster, E., Pimperton, H., Ralph-Lewis, A., Monroy, L., Hulme, C., and MacSweeney, M. (**2018**). "Eye movements during visual speech perception in deaf and hearing children," Lang. Learn. **68**, 159–179.

Zoefel, B., Allard, I., Anil, M., and Davis, M. H. (**2020**). "Perception of rhythmic speech is modulated by focal bilateral transcranial alternating current stimulation," J. Cogn. Neurosci. **32**, 226–240.

Zorowitz, S., Chierchia, G., Blakemore, S.-J., and Daw, N. D. (**2024**). "An item response theory analysis of the matrix reasoning item bank (MaRs-IB)," Behav. Res. Methods **56**, 1104–1122.