Neural Speech Tracking Highlights the Importance of Visual Speech in Multi-speaker Situations

Chandra L. Haider[1], Hyojin Park[2], Anne Hauswald[1], and Nathan Weisz[1,3]

Abstract

■ Visual speech plays a powerful role in facilitating auditory speech processing and has been a publicly noticed topic with the wide usage of face masks during the COVID-19 pandemic. In a previous magnetoencephalography study, we showed that occluding the mouth area significantly impairs neural speech tracking. To rule out the possibility that this deterioration is because of degraded sound quality, in the present follow-up study, we presented participants with audiovisual (AV) and audio-only (A) speech. We further independently manipulated the trials by adding a face mask and a distractor speaker. Our results clearly show that face masks only affect speech tracking in AV conditions, not in A conditions. This shows that face masks indeed primarily impact speech processing by blocking visual speech and not by acoustic degradation. We can further highlight how the spectrogram, lip movements and lexical units are tracked on a sensor level. We can show visual benefits for tracking the spectrogram especially in the multi-speaker condition. While lip movements only show additional improvement and visual benefit over tracking of the spectrogram in clear speech conditions, lexical units (phonemes and word onsets) do not show visual enhancement at all. We hypothesize that in young normal hearing individuals, information from visual input is less used for specific feature extraction, but acts more as a general resource for guiding attention. ■

INTRODUCTION

As the production of an acoustic speech stream is inevitably accompanied by visible mouth movements in face-to-face speech, it is unsurprising that the brain has developed mechanisms to interpret both the visual input and the auditory input. More importantly, it is also able to integrate both to form a coherent experience. Despite speech processing being dominated by the auditory modality, with the exception of sign language, visual speech plays an important role in the discrimination of neighboring phonemes as demonstrated by the classic McGurk effect (McGurk & Macdonald, 1976) and improves speech comprehension compared to audio-only (A) settings (Sumby & Pollack, 1954). Research in this field has been brought to general attention with the prevalence of face masks during the COVID-19 pandemic and their impact on speech comprehension in everyday situations. However, how does the brain use visual information for speech processing when listening situations are more or less challenging? And how is speech processing affected in such situations, when listeners are deprived of relevant visual input?

Theories about audiovisual (AV) speech integration propose two different paths of speech processing enhancements through congruent visual speech (Van Engen, Dey, Sommers, & Peelle, 2022; Peelle & Sommers, 2015). On the one hand, visual speech acts as a temporal facilitator for the attentional focus on the acoustic information as visual information precedes the acoustic one (Bourguignon, Baart, Kapnoula, & Molinaro, 2020; Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). On the other hand, there is direct information from the mouth and lip area providing information about the phonetic content of speech (e.g., place of articulation), which helps with phoneme discrimination.

To investigate these processes on a neural level, new research possibilities have emerged, which make it possible to investigate the processing of natural continuous speech presentations. Different approaches have been developed that quantify how "faithfully" neural responses track speech features (Hauswald, Lithari, Collignon, Leonardelli, & Weisz, 2018; Crosse, Di Liberto, Bednar, & Lalor, 2016; Park, Kayser, Thut, & Gross, 2016). In general terms, these measures of neural speech tracking implement some association measure between the input signal (e.g., speech envelope) and the output signal (e.g., M/EEG data). The outcome is held to represent the degree to which the respective speech feature (e.g., envelope) is encoded in the brain (Brodbeck et al., 2021; Crosse, Di Liberto, Bednar, et al., 2016).

In the domain of AV speech, past research has revealed that the decoding of the speech envelope from brain data already improves when presenting AV speech in comparison to A speech when the audio input is clear (Crosse,

[1]Paris Lodron Universität Salzburg, [2]University of Birmingham, [3]Paracelsus Medical University Salzburg

Journal of Cognitive Neuroscience 36:1, pp. 128–142
https://doi.org/10.1162/jocn_a_02059

Butler, & Lalor, 2015), but this visual enhancement further increases when the acoustic information is noisy (Crosse, Di Liberto, & Lalor, 2016). We investigated this phenomenon in the context of face masks as commonly used during the COVID-19 pandemic (Haider, Suess, Hauswald, Park, & Weisz, 2022). In that continuous AV speech study, we manipulated the speech signal by adding a face mask and/or a distractor speaker (2 × 2 design). For that study, we explored the features of two categories. First, we looked at acoustic features (i.e., speech envelope, pitch, and formants) and, second, features of lexical segmentation (i.e., phoneme and word onsets) extracted from the speech stimuli. Interestingly, results differed for tracking of acoustic features and lexical segmentation features. On the one hand, tracking of acoustic features was generally affected by masking, whereas on the other hand, tracking of lexical segmentation features was mostly only affected by masking when a distractor was added as well.

However, certain limitations had to be taken into consideration. First, we could not differentiate if the effects on neural tracking via adding a (surgical) face mask were introduced by acoustic distortion of the speech input or by blocking visual speech. Second, we only investigated backward modeling (i.e., stimulus reconstruction), which does not allow to link certain experimental effects to specific regions but gives an overall estimate of stimulus representation in the brain. In addition, the study is limited in its ability to compare the relative contributions of different feature classes, such as acoustic and lexical segmentation, in AV speech processing. A study from our research group, using a speech encoding model on the same data as Haider et al. (2022), showed an increased tracking of lip movement in two-speaker speech compared to single-target speaker speech (Reisinger et al., 2023). This might be indicative of a greater usage of lip movements for focusing attention on the succeeding acoustic information as proposed by Van Engen et al. (2022). Another study found increased AV benefit in noisy conditions compared to clear speech conditions through the addition of lexical features in the form of phonemes by using encoding models and canonical correlation analysis (CCA) (O'Sullivan, Crosse, Di Liberto, de Cheveigné, & Lalor, 2021). That study, however, relied on two separate experiments (one for clear speech and one for noisy speech), which does not allow for a within-participant evaluation of the effect. To this date, it remains unclear on which characteristics (basic lip movements vs. lexical unit features) the enhancement of speech tracking depends on and if the contributions of visual speech can be separated into different functional mechanisms.

With this current study, we want to rule out the more trivial possibility that the negative effect of the face mask is the product of degraded acoustics. Furthermore, we want to separate visual speech benefits generated by basic lip movements and by lexical/phonetic features. For this purpose, we adjusted the experimental design to also include conditions with an A presentation. Now, we can investigate whether the effects of face masks are generated by acoustic distortions or by masked visual cues. An illustration of the new study design can be seen in Figure 1A. To investigate how different brain regions contribute to the found effects, we complemented the backward decoding approach with a forward encoding approach in this study, thereby linking the effects to magnetoencephalography (MEG) channels. In addition, we are now able to compare models with features of different classes directly. We computed forward models trained on acoustic (spectrogram) features and assessed the impact of adding visual (i.e., lip area) and lexical units (i.e., phonemes and word onsets) features on speech tracking in clear and noisy conditions. Hereby, we can evaluate how adding complexity to the model improves outcomes and are able to quantify the visual benefit obtained through these additional features.

METHODS

Participants

Twenty-eight German native speakers (14 women) aged between 20 and 36 years (M = 28.18 years, SD = 3.79 years) took part in our study. As we wanted to extend the findings of our previous study, we opted for a similar sample size as in Haider et al. (2022). The exclusion criteria were nonremovable magnetic objects as well as a history of psychiatric or neurological conditions. Recruitment was done via social media and university lectures. All participants signed an informed consent form and were compensated with € 10 per hour or course credit. The experimental protocol was approved by the ethics committee of the University of Salzburg and was carried out in accordance with the Declaration of Helsinki.

Stimuli

For this study, we used the same stimulus material as in Haider et al. (2022) but with a reduced number of trials. We used excerpts from four different stories for our recording read out in German. "Die Schokoladenvilla—Zeit des Schicksals. Die Vorgeschichte zu Band 3" ["The Chocolate Mansion, The Legacy—A Prequel of Volume 3"] by Maria Nikolai and "Die Federn des Windes" ["The Feathers of the Wind"] by Manuel Timm were read out by a female speaker. "Das Gestüt am See. Charlottes großer Traum" ["The Stud Farm by the Lake. Charlotte's Great Dream"] by Paula Mattis and "Gegen den Willen der Väter" ["Against the Will of Their Fathers"] by Klaus Tiberius Schmidt were read out by a male speaker.

Stimuli were recorded using a Sony FS100 camera with a sampling rate of 25 Hz for video and a Rode NTG 2 microphone with a sampling rate of 48 kHz for audio. We aimed at a duration for each story of approximately 10 min, which were cut into 10 videos of around 1 min each (range: 56–76 sec, M = 63 sec, SD = 5.0 sec). All stories were

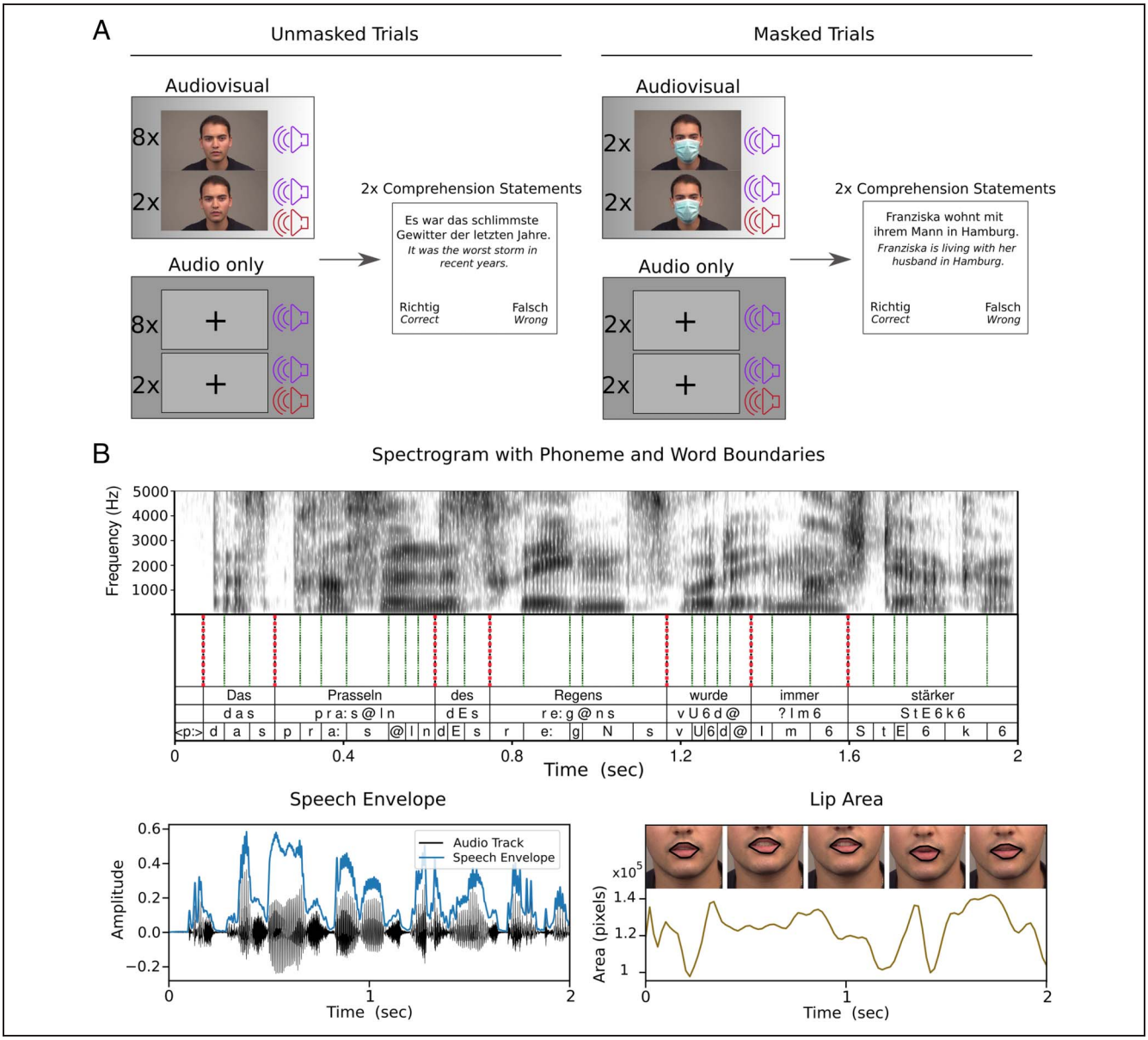

Figure 1. Experimental procedure and speech features. (A) The condition design for the experiment with the male speaker as an example. Every participant was presented with 30 trials of ~1 min each. We conducted AV trials (14 trials in total), A trials (14 trials in total), and V trials (two trials in total; V trials not depicted in figure). For A and AV trials, we used 10 unmasked trials each. In eight of these trials, only a single target speaker was presented. In two of those trials, we added a second same-sex (distractor) speaker (denoted by the second sound icon). We used four masked trials for A and AV speech, respectively. In two of those trials, again only a single target speaker was presented, and in two trials, we added a distractor speaker. The design was unbalanced to obtain sufficient training data to compute the temporal response function models (see TRF Model Fitting section). After each A and AV trial, we prompted two "true or false" statements to measure comprehension and to keep participants focused. Participants answered via a button press (left or right button). (B) The investigated speech features. The spectrogram is shown alongside the investigated speech units, phonemes, and words underneath (top row: orthographic word; middle row: phonetic word; bottom row: phoneme). The speech envelope can be seen on the bottom left of the figure. On the bottom right of the figure, the extracted lip area can be seen with the presentation of the mouth outline corresponding to the 1st, 13th, 25th, 38th, and 50th frames (i.e., 0, 0.5, 1, 1.5, and 2 sec). All depictions are based on the same 2-sec-long speech interval. The spectrogram and speech envelope are depicted before downsampling for illustrative purposes.

recorded twice, once without the speaker wearing a surgical face mask and once with the speaker wearing a surgical face mask (Type IIR, three-layer single-use medical face mask; see Figure 1A). After cutting, all videos were approximately 1 min in length (80 AV recordings in total). Thirty of those were presented to each participant (15 with a female speaker, 15 with a male speaker) to rule out the sex-specific effects of the stimulus material. The audio track was extracted and stored separately. The audio files were then normalized using the Python function *ffmpeg-normalize*. Prerecorded audiobooks read out by different speakers (one female, one male) were used for the distractor speaker and normalized using the same method. The syllable rate was analyzed using a Praat script (de Jong & Wempe, 2009; Boersma & Weenink,

2001). The target speakers' syllable rates in the 30 trials varied between 3.65 and 4.57 Hz ($M = 4.04$ Hz, $SD = 0.25$ Hz). Target and distractor stimuli were all played to the participant at the same volume, which was individually set to a comfortable level at the start of the experiment by using an example audiobook with the target female speaker.

Experimental Procedure

Before the start of the experiment, we performed a standard clinical audiometry using an AS608 Basic (Interacoustics) to assess participants' individual hearing ability. Of all participants, four participants showed hearing impairment in higher frequencies (≥3000 Hz) but reported no-to-mild subjective impairment. Afterward, participants were prepared for MEG (see Data Acquisition section).

We started the MEG measurement with 5 min of resting-state activity (not included in this article). We then adjusted the stimulation volume by presenting an exemplary audiobook and adjusted the volume until participants stated that it was clearly audible and comfortable. One block consisted of six trials of ~1-min length. The condition design is depicted in Figure 1. As an extension to Haider et al. (2022), we added audio-only modality (A) and visual-only modality (V) conditions in addition to the AV modality conditions. For the A trials, we used the same stimuli as the AV trials but did not show the corresponding video and instead presented a fixation cross. This results in 14 AV, 14 A, and two V trials. For A and AV conditions respectively, we still wanted to keep the 2 × 2 design from Haider et al. (2022). Therefore, eight trials per modality consisted of clear speech (i.e., no mask and no distractor), whereas in the trials where speakers wore a face mask, a second same-sex speaker was added or both (i.e., two trials per one of these conditions). This design was chosen to have sufficient data to train the temporal response function (TRF) models on the one hand and to have a reasonable experiment duration on the other hand. All conditions had an equal amount of female and male speaker trials. In conditions with a second speaker (distractor), the second speaker only started 5 sec after the first (target) speaker to give participants time to focus on the to-be-attended speaker. Within the first four blocks, the story presentation followed a consistent storyline across trials. The fifth block was a mixed speaker block (both female and male speakers) containing continuations from the story "Gegen den Willen der Väter" (male speaker) and "Die Federn des Windes" (female speaker; three trials each). Some participants also performed six extra V trials, which resulted in an additional sixth block. For this, three trials from the story "Die Schokoladenvilla—Zeit des Schicksals. Die Vorgeschichte zu Band 3" and three trials from "Das Gestüt am See. Charlottes großer Traum" were used. Each story segment was only presented once per participant. As not all participants participated in this sixth block, this was not included in the analysis. With the exception of the V trials, two unstandardized true or false statements regarding semantic content were asked at the end of trials to assess comprehension performance and keep participants focused (Figure 1A). In addition, participants rated subjective difficulty twice per condition on a 5-point Likert scale for A and AV trials. The participants' answers were given via button presses. The condition design was shuffled across trials to rule out any influence of the specific stimuli material on the results. As an exception to this, we always assigned a trial without a mask and without a distractor as the starting trial of each block to let participants adapt to a (possible) new target speaker. Videos were back-projected on a translucent screen with a screen diagonal of 74 cm via a PROPixx DLP projector (VPixx Technologies) ~110 cm in front of the participants. It was projected with a refresh rate of 120 Hz and a resolution of 1920 × 1080 pixels. Including preparation, the experiment took about 2 hr per participant. The experiment was coded and conducted with the Psychtoolbox-3 (Kleiner et al., 2007; Brainard, 1997; Pelli, 1997) with an additional class-based library (Objective Psychophysics Toolbox, o_ptb) on top of it (Hartmann & Weisz, 2020).

Data Acquisition

We recorded brain data with a sampling rate of 1 kHz at 306 channels (204 first-order planar gradiometers and 102 magnetometers) with a TRIUX MEG system (MEGIN). The acquisition was performed in a magnetically shielded room (AK3B, Vacuumschmelze). Online bandpass filtering was performed from 0.1 to 330 Hz. Before the acquisition, cardinal head points (nasion and preauricular points) were digitized with a Polhemus FASTRAK digitizer (Polhemus) along with around 300 points on the scalp to assess individual head shapes. Using a signal space separation algorithm provided by the MEG manufacturer (Maxfilter, Version 2.2.15), we filtered noise resulting from sources outside the head and realigned the data to a standard head position, which was measured at the beginning of each block.

Speech Feature Extraction

All the speech features investigated are depicted in Figure 1B. The speech envelope was extracted using the Chimera toolbox. By using the default options, the speech signal was filtered forward and in reverse with a fourth-order Butterworth bandpass filter at nine different frequency bands equidistantly spaced on the cochlear map between 100 and 10000 Hz (Smith, Delgutte, & Oxenham, 2002). Then, a Hilbert transformation was performed to extract the envelopes from the resulting signals. These nine bands were then summed up to one general speech envelope and normalized. Finally, an 80-Hz low-pass filter was applied on the envelope.

The spectrogram was computed similarly. We only adjusted the lower end of the frequency range to 50 Hz, as this then also includes the speech pitch/fundamental frequency (especially for the male speaker). In the end,

the resulting nine bands were not summed up but used together as the speech spectrogram.

Phonemes and word onset values were generated using forced alignment with MAUS Web services (Kisler, Reichel, & Schiel, 2017; Schiel, 1999) to obtain a measure for speech segmentation. We generated two time series at 50 Hz with binary values indicating an onset of a phoneme or word, respectively.

To use these binary features for backward modeling, we smoothed the time series of binary values using a Gaussian window with a width of 20 msec, as 20 msec is the smallest possible window at a sampling rate of 50 Hz. In addition, we did not distinguish between different phonemes in the case of backward modeling to generate a single time series for reconstruction (see Haider et al., 2022). These last two steps were skipped when features were analyzed using forward models because, in this case, binary features can be investigated and a multivariate input (i.e., different individual phonemes) is beneficial. In this case, we ended up with 63 time series, one for each phoneme, containing only zeros and ones.

The lip area was extracted from the video recordings using a MATLAB function by Park et al. (2016). Through this, we generated a time series of the lip area in pixels with a sampling rate of 25 Hz (i.e., the frame rate of the video recording).

In the end, all features were sampled to 50 Hz to match the sampling rate of the corresponding brain signal, as most speech-relevant signals present themselves below 25 Hz (Crosse et al., 2021).

MEG Preprocessing

The raw data were analyzed using MATLAB R2020b (The MathWorks) and the FieldTrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011). As part of our standard pipeline, we first computed 50 independent components to remove eye and heart artifacts. We removed, on average, 2.29 components per participant (SD = 0.86). We further filtered the data using a sixth-order zero-phase Butterworth bandpass filter between 0.1 and 25 Hz. Finally, we downsampled our data to 50 Hz for more efficient computation while still preserving sufficient information from our data (Crosse et al., 2021).

TRF Model Fitting

To prepare the data for fitting the TRF models, we first z scored the MEG data of all 306 channels and rescaled the speech envelope and lip area to values between 0 and 1. This rescaling was preferred over z scoring, as this does not introduce negative values into only positive sign features. We used the mTRF Toolbox (Crosse, Di Liberto, Bednar, et al., 2016) to reconstruct stimulus features (in the case of backward modeling) or to predict brain data (in the case of forward models). This was done similarly to Haider et al. (2022). For a more detailed discussion about this method, please refer to Crosse et al. (2021) and Crosse, Di Liberto, Bednar, et al. (2016). We trained two separate models on AV clear speech (i.e., AV speech without a face mask and without a distractor) and on A clear speech. We then used these models to reconstruct the stimulus feature or predict brain activity for their respective modality (AV model for AV conditions and A model for A conditions) but across other conditions (i.e., for trials with masks and with distractors). As the second distractor speaker only starts after 5 sec into the trial, we removed the first 5 sec of each distractor trial from the analysis. For both forward and backward modeling, we used time lags from −50 to 500 msec to train our models. Regularization parameter (lambda) between 10^{-6} and 10^{6} was determined using a sevenfold leave-one-out cross-validation (Willmore & Smyth, 2003).

To completely rule out the possibility that a single outlier trial influences the model training, we again made use of a leave-one-out procedure. We used seven clear speech trials for model training and the eighth clear speech trials for model testing. We then looped across the trials so that every trial was once used for testing and otherwise used for training. In the end, we averaged the Fisher z-transformed values across each iteration together. Overall, we used ∼7 min for training our A and AV models and ∼1 min of testing for the clear speech condition in each iteration. For all other three conditions (i.e., No Mask + Distractor, Mask + No Distractor, Mask + Distractor), we used all trials for testing, which was ∼2 min each.

Backward Modeling

With this approach, we are able to map the brain response back to the stimulus feature (e.g., the speech envelope) to acquire a measure of how well a certain feature is encoded in the brain. This model has the benefit that the whole-brain activity is taken into account when reconstructing a feature, which holds information about a general representation of a certain stimulus feature in the brain. By this approach, one obtains a single correlation value (Pearson's r) as the measure of how well a feature can be reconstructed, which makes it easily interpretable. This, however, comes with the downside that the reconstructed features are all modeled independently of each other. Therefore, the amount of shared information between the features is not taken into account. It is therefore more commonly used on 1-D features (e.g., speech envelope) than on multidimensional features (e.g., spectrogram), as in the latter case, each dimension is reconstructed independently, which makes claims on how well a certain feature is tracked as a whole difficult. Furthermore, by using this approach, spatial information on which brain regions are involved in the processing of certain stimulus characteristics is lost. To solve both of these issues, one can use forward modeling as a complementary analysis.

Forward Modeling

By using this approach, we are mapping the stimulus (feature) "forward" to the individual MEG channels. We therefore acquire a measure of how well a certain stimulus feature is encoded in the individual channels. In contrast to backward models, this model cannot give a measure of how well a feature is generally encoded in the whole brain, as it does not account for intercorrelations between the channels. However, as we get a prediction accuracy measure for each individual channel, it is possible to acquire a spatial distribution in the brain and test hypotheses about the contribution of different brain regions. In addition, we can use multivariate inputs to predict brain data (e.g., spectrogram), which allows us to evaluate more complex (multidimensional) features. By adding a feature to a baseline model and subsequently comparing the new model to the baseline, one is able to obtain the contribution of the added feature above the simpler model. This is especially important in speech research, as speech features are highly correlated.

Statistical Analysis

For the analysis of backward models and separately for the A and AV conditions, we performed 2 × 2 repeated-measures ANOVA with the factors Mask (no face mask vs. face mask) and Distractor (no distractor speaker vs. distractor speaker) and the obtained Fisher z-transformed correlation coefficients (i.e., reconstruction accuracy) as dependent variables. For evaluating the effects of Modality (A vs. AV), we additionally performed a 2 × 2 × 2 repeated-measures ANOVA with the within-factors Modality, Mask, and Distractor.

For the behavioral results (comprehension performance and subjective difficulty), we also used a repeated-measures ANOVA with the same factors Modality, Mask, and Distractor. We used comprehension performance scores (i.e., the percentage of correct answers) and averaged subjective difficulty ratings respectively as dependent variables. For one participant, behavioral data were missing, resulting in 27 participants for this analysis.

We analyzed the forward models with one-tailed dependent sample cluster-based permutation tests with 10,000 permutations (Maris & Oostenveld, 2007). We used the "maxsum" method implemented in FieldTrip for the calculation of the t statistic. We furthermore investigated the effect size of each cluster by averaging the Cohen's d values across all significant channels in that cluster. A one-tailed test was chosen because we tested for differences in simple versus additive models, whose performance must be at least on the level of the simple model. We restricted our analysis to combined gradiometers, as these allow for a better spatial resolution. For that, the correlation coefficients of each gradiometer pair were averaged.

RESULTS

Behavioral Results

For comprehension performance, we computed the percentage of correctly answered questions for each condition for each participant. We analyzed the data according to our 2 × 2 × 2 design with the factors Modality (A vs. AV), Mask (no mask vs. mask) and Distractor (no distractor vs. distractor). The only significant main effect was for the Distractor, revealing a significant decrease in comprehension performance, $F(1, 26) = 5.73, p < .024, \eta_p^2 = .18$. The interaction between Modality and Mask was also significant, showing only a decrease through the mask of comprehension performance in the AV conditions, $F(1, 26) = 6.28, p < .019, \eta_p^2 = .20$. Please note that, for most conditions (all except clear speech conditions), we only prompted four comprehension statements per participant, with a guess rate of 50%. Therefore, these results should be interpreted with care. All other results were not significant and are shown in Figure 2A. Simple effects for the Mask effect are shown in Appendix Table A1.

The subjective difficulty was calculated by averaging the ratings (as indicated on a 5-point Likert scale by the participant) per condition. Results show a significant effect of Modality, $F(1, 26) = 11.84, p = .002, \eta_p^2 = .31$; Mask, $F(1, 26) = 8.71, p = .007, \eta_p^2 = .25$; and Distractor, $F(1, 26) = 109.68, p < .001, \eta_p^2 = .81$. The interactions between Modality and Mask were also significant, $F(1, 26) = 7.23$, $p = .012, \eta_p^2 = .22$, as well as between Modality and Distractor, $F(1, 26) = 5.04, p = .034, \eta_p^2 = .16$. Participants reported more difficulty in face mask trials than in no-face-mask conditions only in the AV condition, whereas it stayed the same for the A conditions. Moreover, AV stimuli had a buffering effect on the influence of the distractor by reducing its impact as indicated by the significant interaction between Modality and Distractor. All other results were not significant and are shown in Figure 2A. Again, simple effects for the Mask effect are shown in Appendix Table A1.

Backward Modeling Is Only Affected by Face Masks in AV Conditions not in Audio-only Conditions

By using the mTRF Toolbox in the backward direction (Crosse, Di Liberto, Bednar, et al., 2016), we calculated one correlation coefficient per condition and per participant for each investigated feature. The features used here were the speech envelope (as an acoustic feature), phoneme and word onsets (as lexical segmentation features), and lip area (as a visual feature). The correlation coefficients were then Fisher z transformed and averaged across each condition.

For statistical analysis, we used two 2 × 2 ANOVA models separately for A and AV conditions and performed a 2 × 2 × 2 ANOVA to investigate the influence of Modality and the interactions. In general, all reconstruction accuracies investigated were heavily influenced by the addition

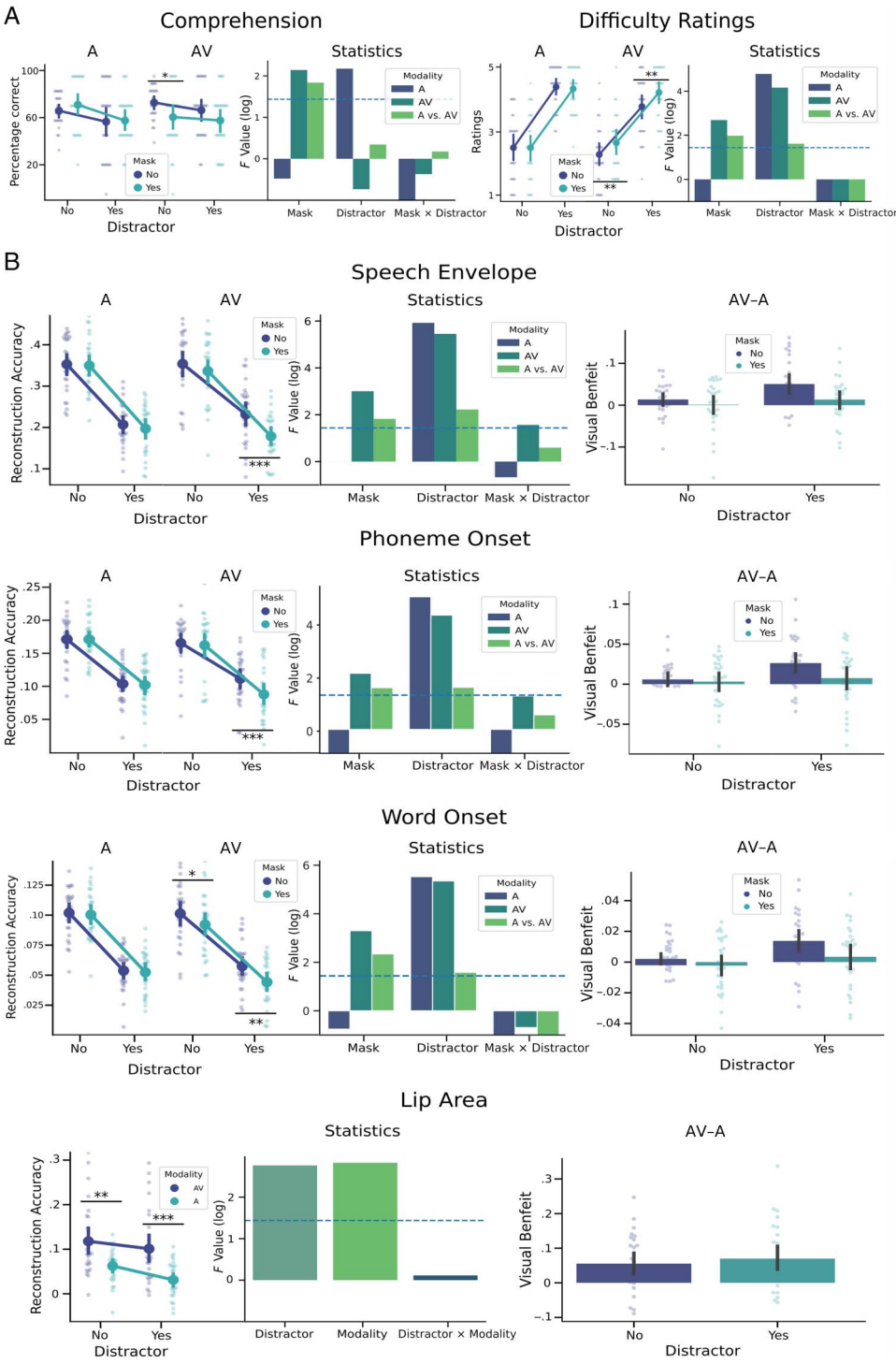

Figure 2. Visual illustration of the results of the behavioral results and backward modeling. (A) Visual illustration of the behavioral results of comprehension and difficulty ratings split up for A and AV conditions. The left graphs show the overall values, whereas right to it, the statistical evaluation of the effects is depicted. The dashed horizontal blue line represents the critical *F* value ($\alpha = 5\%$) of 4.21 (1.44 in log space). The light green bar represents the interaction with Modality (i.e., the difference between A and AV) for each effect. (B) Visual illustration of the results of the backward modeling for the investigated features speech envelope, phoneme onset, word onset, and lip area. On the left side, the raw reconstruction accuracy is depicted split up for A and AV conditions, except for lip area. Next to it on the right, the statistical evaluation of the effects of Mask and Distractor and their interaction are shown with log-scaled *F* values. On the right-hand side, the visual benefit for each participant is shown (i.e., AV–A). For the lip area feature, it was not possible to evaluate the effects of the face mask; therefore, only the effects of Distractor and Modality can be seen. The dashed horizontal blue line represents the critical *F* value ($\alpha = 5\%$) of 4.20 (1.44 in log space). The light green bar represents the interaction with Modality (i.e., the difference between A and AV) for each effect. All error bars in the figure represent 95% confidence intervals calculated by *seaborn*. $^{\circ}p < .1$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Not significant simple effects are not shown here for visual clarity. All conducted simple effect tests can be seen in Appendix Tables A1 and A2.

of a distractor speaker. Therefore, the main effect of Distractor will not be reported here but is visually depicted in Figure 2B. Furthermore, simple effects for the Mask effect can be seen in Appendix Table A2.

The visual benefit was calculated by subtracting the reconstruction accuracies in the A conditions from the AV conditions. Results are shown in Figure 2. For the conditions when the speakers wore a face mask, we could not compute the lip area. However, we investigated the lip area tracking in the AV conditions and also the tracking of the unseen lip area in the A condition. Note that, with backward modeling, we cannot take shared information between features into account. However, differences in how the stimulus reconstruction of different features is affected by the experimental manipulation are still informative. Nonetheless, a complementary analysis with forward modeling is performed later, which allows to control for shared information.

Speech Envelope

In the AV conditions, the reconstruction of the speech envelope showed impairment through a face mask, $F(1, 27) = 20.15, p < .001$ $\eta_p^2 = .42$, as well as an interaction with a distractor, $F(1, 27) = 4.79, p = .037$ $\eta_p^2 = .15$. For the A conditions, both these effects are missing [Mask: $F(1, 27) = 0.999, p = .327, \eta_p^2 = .04$; Mask × Distractor: $F(1, 27) = 0.51, p = .48, \eta_p^2 = .02$]. The effect of modality shows a significant main effect, $F(1, 27) = 10.42, p = .003$, $\eta_p^2 = .28$, as well as a significant interaction with Mask, $F(1,27) = 6.20, p = .019, \eta_p^2 = .19$, and Distractor, $F(1, 27) = 9.24, p = .005$ $\eta_p^2 = .26$. The threefold interaction (Modality × Mask × Distractor), however, is not significant, $F(1, 27) = 1.82, p = .189, \eta_p^2 = .06$. Investigating the confidence intervals for each condition individually reveals that only for the distractor condition without a face mask the visual benefit is significantly above zero (95% CI [.0295, .0717]). Hereby, we can confirm the idea that visual speech has a buffering effect on tracking of the speech envelope especially in conditions where the audio input is unclear. Face masks only affect tracking in AV conditions but do not show any significant impact on tracking in A conditions. Similar results can be seen for the lexical segmentation features of phoneme onset and word onset.

Phoneme Onset and Word Onset

In the AV conditions, the reconstruction of the phoneme onsets showed impairment through a face mask, $F(1, 27) = 10.49, p < .001, \eta_p^2 = .003$, and a trend for the interaction with the distractor condition, $F(1, 27) = 4.07, p = .054, \eta_p^2 = .13$. For the A conditions, both these effects are not significant [Mask: $F(1, 27) = 0.06, p = .809, \eta_p^2 = .00$; Mask × Distractor: $F(1, 27) = 0.04, p = .843, \eta_p^2 = .00$]. The effect of Modality shows a significant main effect, $F(1, 27) = 9.03, p = .005, \eta_p^2 = .25$, as well as a significant interaction with Mask, $F(1, 27) = 5.75, p = .024, \eta_p^2 = .18$, and Distractor, $F(1, 27) = 5.84, p = .023, \eta_p^2 = .18$. The threefold interaction (Modality × Mask × Distractor), however, is not significant, $F(1, 27) = 1.84, p = .186, \eta_p^2 = .06$. Again, the only condition with significant visual benefit was the distractor condition without a face mask (95% CI [.0147, .0381]).

In the AV conditions, the reconstruction of the word onsets showed impairment through a face mask, $F(1, 27) = 27.15, p < .001, \eta_p^2 = .50$, but no interaction with the distractor condition, $F(1, 27) = .50, p = .487, \eta_p^2 = .02$. For the A conditions, both these effects are again not significant [Mask: $F(1, 27) = 0.50, p = .499, \eta_p^2 = .02$; Mask × Distractor: $F(1, 27) = 0.00, p = .978, \eta_p^2 = .00$]. The effect of Modality shows a significant main effect, $F(1, 27) = 8.57$, $p = .007, \eta_p^2 = .24$, as well as a significant interaction with Mask, $F(1, 27) = 10.44, p = .003, \eta_p^2 = .28$, and Distractor, $F(1, 27) = 4.88, p = .036, \eta_p^2 = .15$. As for the other features, the threefold interaction (Modality × Mask × Distractor) is not significant, $F(1, 27) = .26$, $p = .616, \eta_p^2 = .01$. In this case, visual benefit is significant in both conditions without a mask, that is, without a distractor (95% CI [.0007, .0079]) and with a distractor (95% CI [.0068, .0208]).

For both lexical segmentation features, results are similar to the effects observed for tracking of the speech envelope. Again, face masks only have an impact in AV conditions, not in A conditions. The visual benefit is largest in the condition with a distractor speaker but only when the speaker does not wear a face mask. For the tracking of the word onsets, both conditions without a face mask show significant visual benefit (despite being much smaller in the clear speech conditions).

Lip Area

As we could not compute the lip area for the trials in which speakers wore a face mask (as the mouth is occluded in the video), we computed a 2 × 2 ANOVA with the factors Modality (A vs. AV) and Distractor (no distractor vs. distractor). Both main effects are significant [Modality: $F(1, 27) = 16.90, p \leq .001, \eta_p^2 = .39$; Distractor: $F(1, 27) = 15.86, p \leq .001, \eta_p^2 = .37$], whereas the interaction was not significant, $F(1, 27) = 1.12, p = .299, \eta_p^2 = .04$. As expected, both conditions show above-zero visual benefit (no distractor: 95% CI [.0255, .0852], distractor: 95% CI [.0356, .1041]). These results highlight the importance of controlling for effects generated by correlated audio features when investigating visual speech in an AV paradigm. This is further supported by a strong correlation between audio-only lip area tracking values and audio-only speech envelope tracking values [no distractor: $r(26) = .605, p < .001$, distractor: $r(26) = .472, p = .01$]. Simple effects show a significant effect between AV and A only in when no distractor was present, $t(26) = 3.57, p = .001, d = 0.69$, as well as in the presence of a distractor speaker, $t(26) = 3.93, p < .001, d = 0.76$.

Forward Modeling Reveals Improved Tracking Across Conditions Through Modulations of the Lip Area and Increased Visual Benefit Through Lexical Features Only in Clear Speech

By using the mTRF Toolbox in the forward direction (Crosse, Di Liberto, Bednar, et al., 2016), we calculated one correlation coefficient per condition, per channel, and per participant for each investigated feature. In contrast to backward modeling, we can now investigate features of multidimensional scale, such as the speech spectrogram or individual phonemes. As speech features are proven to be strongly intercorrelated, the use of additive models is highly beneficial to investigate differences and potential gains through adding additional features to the models. For this analysis, we computed a (base) acoustic model (spectrogram), a spectrogram + lip model (spectrogram + lip area), and a spectrogram + lexical model (spectrogram + phonemes + word onsets). The resulting correlation coefficients from the models were Fisher z transformed and averaged within conditions. As we have shown for the backward models, the effect of the mask is only affecting features in conditions with visual input significantly. Therefore, we did not include conditions with masks in this following result section but instead focused on the visual benefit

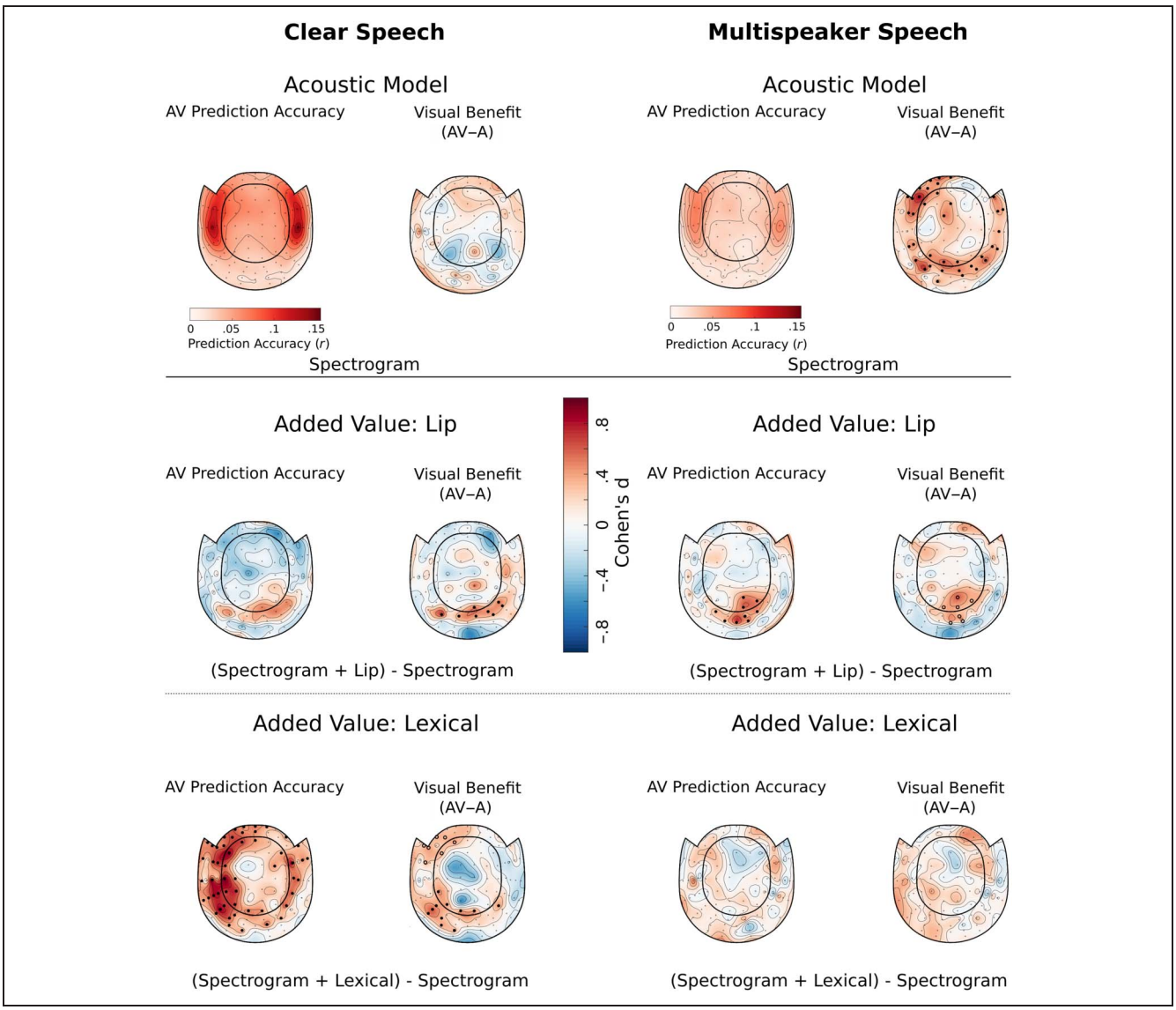

Figure 3. Visual illustration of the results of the forward modeling and added value. Results of our forward modeling analysis. The first row shows the prediction accuracy and visual benefit for all combined gradiometers of the (baseline) acoustic model based on the spectrogram. On the left side, these values were acquired from brain data of participants listening to clear speech; on the right side, they were listening to a multi-speaker (distractor) speech. Statistical analysis for the visual benefit (i.e., AV–A) for the acoustic model was calculated. The second and third rows show the added value of visual features (lip area) and lexical (unit) features (phonemes and word onsets), respectively. Specifically, they represent the difference between the spectrogram model and the additive model (i.e., spectrogram + lip or spectrogram + lexical) for prediction accuracies and visual benefit. This again is done for clear speech and distractor speech. Filled dots represent channels with significant differences computed by a one-tailed cluster-based permutation test (α = 5%). Circles represent trends ($p < .10$).

in clear speech and speech with a second (distractor) speaker. To investigate the effects statistically, we used a one-tailed dependent-sample cluster-based permutation test, as described in the Methods section.

Spectrogram Model

We used the spectrogram model as our baseline model, shown in Figure 3. For this model, we only estimated statistical contrasts between the A and AV speech for clear speech and distractor speech. In clear speech, we found a cluster on trend level in the right temporal ($t_{sum} = 16.05$, $p = .069$, $d = .66$) and in occipital sensors ($t_{sum} = 13.34$, $p = .099$, $d = .69$). We found an increased tracking in AV compared to A speech in the distractor condition spreading over frontal temporal and occipital sensors ($t_{sum} = 130.73$, $p < .001$, $d = 1.12$). This effect corresponds well to the observation of the backward models, which confirms that visual benefit is beneficial in the conditions with an unclear speech acoustic.

Added Value: Spectrogram + Lip Model

The additive model of spectrogram and lip area showed a significant improvement in the clear speech condition over the acoustic-only spectrogram model for overall prediction accuracy in occipital sensors ($t_{sum} = 22.36$, $p = .037$, $d = .68$). Further, a significant increase in visual benefit in the clear speech conditions can be observed over occipital regions ($t_{sum} = 16.34$, $p = .048$, $d = .50$) with an additional trend over frontal regions ($t_{sum} = 13.65$, $p = .070$, $d = 49$). For multi-speaker speech, we found no significant cluster for overall prediction accuracy as well as visual benefit.

Added Value: Spectrogram + Lexical Model

For the lexical unit model, we observe an increase in prediction accuracy in the clear speech condition over right temporal and frontal regions ($t_{sum} = 73.42$, $p < .001$, $d = 1.00$) as well as left temporal and frontal regions ($t_{sum} = 42.08$, $p = .016$, $d = .70$). No increase in visual benefit is found for the clear speech conditions.). For multi-speaker speech, we observe a significant cluster over right and left temporal and frontal regions ($t_{sum} = 119.90$, $p < .001$, $d = .68$) showing a general increase in prediction accuracy as well as a trend for a cluster over occipital regions ($t_{sum} = 14.61$, $p = .084$, $d = .67$). We found no significant effects for visual benefit over the spectrogram model.

Please note that although a combination of all feature classes (i.e., acoustic [spectrogram], visual features [lip movements], and lexical units [phonemes and word onsets]) would yield the highest predictive power, it does not additionally highlight the different contributions of lexical unit features and lip movements. Therefore, we opted for not including this in the article.

DISCUSSION

In previous work, we showed that face masks impair the reconstruction of acoustic features and of features of lexical segmentation in the presence of a distractor speaker (Haider et al., 2022). As we, however, only investigated the effects of face masks in AV speech, we could not differentiate the contribution of (blocked) visual speech and the distortion of the acoustic signal to this face mask effect. Furthermore, our study only employed a backward modeling approach, which limits our ability to investigate the contribution of speech features to specific brain regions. In this study, we aimed to answer these remaining questions. The backward modeling approach revealed that the negative effects of face masks on reconstruction accuracy for the investigated speech features can be primarily attributed to the absence of visual information, whereas the acoustic distortion by surgical face masks appears to have a negligible impact on speech tracking. We can further add to this that additional visual information in general has a buffering effect on the decrease in tracking when the audio input becomes unclear, which points to an increase in visual contribution in this situation (Crosse, Di Liberto, & Lalor, 2016; Crosse et al., 2015; Golumbic, Cogan, Schroeder, & Poeppel, 2013).

With forward modeling, we can show improved speech tracking of an acoustic feature (i.e., spectrogram) especially in difficult listening situations through visual speech. We can provide evidence that, beyond acoustic tracking, visual benefit is presenting itself in two separate ways: When listening is relatively unchallenging, visual speech improves neural speech tracking through more complex lexical features. When speech becomes more challenging, the brain only relies on tracking the modulations of the speaker's lip area, possibly to focus attention on the target.

Face Masks Negatively Affect Speech Reconstruction, Comprehension, and Subjective Difficulty Only in an AV not in an A Presentation

As hypothesized in our previous study (Haider et al., 2022), the effect face masks have on tracking is only relevant in AV settings. This is shown by significant effects of the face mask for all features in the AV speech condition, with all these effects being absent in the A speech conditions. This supports the idea that, despite the changes to the acoustic signal (Haider et al., 2022; Homans & Vroegop, 2021; Corey, Jones, & Singer, 2020), the negative impact of face masks is indeed produced by the missing visual speech input. This holds also true for the behavioral results of comprehension performance and participants' subjective difficulty ratings, which measure the constraints people face more directly. This study again shows the negative effects of face masks on speech processing, as quantified on both neural and behavioral

levels. Whereas the negative impact of a decreasing level of speech comprehension is obvious, an increased subjective difficulty has been shown to lead to social withdrawal in some individuals (Hughes, Hutchings, Rapport, McMahon, & Boisvert, 2018). However, our results cannot directly demonstrate that the found effects are directly linked to speech processing. For example, the increased reconstruction accuracy in AV versus A conditions might be because of responses of the visual system to visual information correlated with acoustic information (e.g., lip movements being correlated with acoustics). With our analysis, we cannot definitively answer if this information is processed in parallel or if the visual information is integrated with the acoustic information. Nonetheless, our results are in line with previous literature, and we can show evidence for the proposed inverse effectiveness (i.e., increased multisensory integration if one condition is noisy; Meredith & Stein, 1986).

In comparison to our previous study, where face masks had the most detrimental effects in a distractor condition, the impact was more general in the present study. Especially, the effects for the reconstruction of the word onset do not show the same pattern. In the previous study, we found an interaction of the face mask with the distractor, hinting at a stronger effect of face masks in difficult listening situations, which is missing in the present study. Despite reanalyzing the old data and adjusting the analysis to be identical to this study (i.e., manual resampling of binary feature, smoothing with 20-msec window, adjusting time lags to [−50, 500], and, most importantly, restricting model training to only ~7 min), the effect remained stable. It can only be attributed to (uncontrolled) differences in population or by the fact that participants in the former study had fewer experimental conditions but more trials per condition and could therefore adjust better to each individual listening situation (e.g., listening to a target speaker in multi-speaker speech).

Regardless of these inconsistencies, our results have important implications for the use of face masks in critical contexts, such as hospitals. After our findings, we would advocate in favor of face masks, which allow visual speech still to be processed. As transparent face shields proved to be of little use in stopping the transmission of viruses (Lindsley, Blachere, Law, Beezhold, & Noti, 2021) and transparent face masks are distorting the acoustic signal severely (Brown, Van Engen, & Peelle, 2021; Corey et al., 2020), we cannot make final conclusions here. Using transparent face masks might be a valuable solution in clinical settings when communicating with hearing-impaired individuals (Kratzke, Rosenbaum, Cox, Ollila, & Kapadia, 2021; Atcherson et al., 2017).

Visual Benefits are Especially Strong in Multi-speaker Speech, While the Influence of Specific Visual Features is Limited

With the current study we extend the findings of Haider et al. (2022) by demonstrating visual benefits of speech processing on the individual sensor level. By contrasting audiovisual and audio-only conditions, we can further corroborate these benefits. We can show that tracking of the spectrogram is generally higher in AV conditions than in A-only conditions, but especially in multi-speaker conditions. Furthermore, we can show that lip movements further increase this tracking in clear speech conditions. A past study has already convincingly demonstrated the visual benefit of lip movements in a neural tracking context (Reisinger et al., 2023). However, many studies have only focused on visual tracking in a V speech paradigm. Hauswald et al. (2018) and Bourguignon et al. (2020) demonstrate that tracking of lip movements can be first observed in the visual cortex. This information is then extracted in the right angular gyrus and forwarded to auditory cortices. Here, they are mapped to the predicted corresponding acoustic speech. Although this study focused on separate V and A speech, the conclusion that visual speech is predicting expected acoustic input to auditory cortices seems nonetheless plausible, because in natural speech, the visual information precedes the acoustic one (Bourguignon et al., 2020). Despite an overwhelming amount of behavioral evidence for visual enhancement of speech processing, only little research has been published that demonstrates this on a neural level. Reisinger et al. (2023) showed an increase in the tracking of lip movements when listening to a target speaker while a distractor speaker was also present. They also found that participants differ substantially in the extent to which they incorporate visual speech for speech processing, a result previously demonstrated by Aller, Økland, MacGregor, Blank, and Davis (2022). However, Reisinger et al. (2023) corroborate this finding further by showing that people with this higher reliance on visual speech show worse comprehension performance and higher subjective difficulty ratings in conditions when the mouth area was occluded.

Similarly to the (backward) reconstruction of the lip area, we do not demonstrate this specific added value of lip movements in multi-speaker speech conditions in the current study. However, as can be seen in the visual benefit for tracking of the spectrogram, we see a strong visual benefit not only in temporal and frontal regions but importantly also in occipital regions suggesting involvement of visual processing regions. Intuitively, one could assume that lip movements are involved in this. However lip movements do not yield an additional benefit in multi-speaker conditions - similarly to the backward reconstruction. This general visual benefit therefore is unlikely to be specific for lip movements but rather to visual speech in general as other facial cues also play a role in audiovisual speech processing (Thomas & Jordan, 2004). Additionally, as acoustic and visual speech features are strongly correlated (Chandrasekaran et al., 2009), extraction of information from the lip area might already be encoded in the spectrogram.

As pointed out previously, we can not directly demonstrate that this increase is related to an improvement to

speech processing. We cannot exclude the possibility that this increase in tracking in occipital areas is just related to responses of the visual system to visual input (i.e. lip movements) with no direct influence on speech processing.

The increased tracking in frontal sensors in AV compared to A-only speech in a multi-speaker context, could additionally be attributed to top-down influences: By tracking the target speaker in areas associated with higher level function such as attention, the brain is able to segregate target and distractor speakers more easily. This is done by a process of enhancing the target speaker stream (over the distractor speaker stream) in order to make subsequent information extraction easier (Orf et al., 2022). Lexical units show strong widespread enhancement in clear and multi-speaker speech. However, visual benefits through lexical units seem to be lacking or small. The topography shows increases primarily in frontal and temporal regions in both hemispheres indicative of higher level processes beyond the auditory cortex. We do however not observe a lateralisation as proposed in previous models (Hickok & Poeppel, 2007; Zatorre, Evans, Meyer, & Gjedde, 1992). Further, the observed trend over occipital regions is in line with findings by Nidiffer and colleagues (2021), showing processing of lexical units on the level of occipital cortex.

Similarly to our study, O'Sullivan et al. (2021) investigated AV benefit (i.e., AV > A + V) through lexical features in clear and noisy speech via EEG. In contrast to our study, they found an increase in AV benefit in the noisy condition compared to the clear speech condition for lexical features, while controlling for acoustic features. A few differences to our study have to be taken into account. First, they investigated the AV benefit, whereas we can only make statements about visual benefits for speech processing. Second, two separate experiments with different participants were conducted in clear and noisy speech, respectively, which does not allow for a within-participant comparison of effects. Third, they created a difficult listening situation by adding noise compared to adding a second speaker in our study, which might involve different processes.

Taken together, we observe strong enhancements in speech tracking through visual speech especially in multi-speaker situations, again confirming the notion of inverse effectiveness (Meredith & Stein, 1986). However in the context of young normal hearing individuals, listeners might rely to a lesser extent on specific visually encoded information such as lip movements or lexical units. Instead, they might use visual speech (i.e. lip but also jaw movements) as a more general cue and attentional resource in order to facilitate subsequent acoustic feature extraction.

Future Directions

The impact of face masks on speech tracking can indeed be traced back to missing visual speech. However, we observed mainly young normal-hearing participants in this and our past study, which leaves the question open on how older individuals or individuals with hearing loss are affected by face masks.

As previous studies have shown, the interindividual variability for the incorporation of visual information for speech processing is relatively high (Reisinger et al., 2023; Aller et al., 2022). One interesting participant group in that regard are again individuals having hearing impairment because a past study has already shown that these participants do profit more from added visual information compared to normal-hearing participants (Puschmann et al., 2019). These proposed benefits were, however, never quantified on a neural level. Moreover, results from this study and the study from O'Sullivan et al. (2021) do not show the same pattern. However, differences in approach might be the reason for this. A future within-participant study that makes it also possible to investigate AV benefit (AV > A + V) in clear and distractor speech might be able to reconcile the two results. In addition, a systematic investigation on how different degradations of the speech signal (e.g., multi-speaker speech, adding noise, or vocoding) affect tracking results might be beneficial in completing the picture of speech tracking in difficult listening situations.

Conclusion

Although some differences could be observed to our previous results, we show here that the negative impact of face masks on neural speech tracking (in this case, reconstruction accuracy) can be traced to a blocking of the visual speech of the speaker, whereas acoustic degradations through (surgical) face masks seem to have minimal impact. Using forward modeling, we show that visual speech improves tracking in occipital, temporal and frontal sensors especially in multi-speaker speech. However, the added value of visually tracked lip movements and lexical units seems to be limited. Our sample of young normal hearing individuals might rely to a lesser extent on the tracking of specific visual features such as lip movements or visemes, but uses visual input more as a general attentional resource.

APPENDIX

Table A1. Simple Effects of Behavioral Data for Mask Effect Across Distractor and Modality Conditions

	Audio Only						*AV*					
	No Distractor			*Distractor*			*No Distractor*			*Distractor*		
	No Mask vs. Mask			*No Mask vs. Mask*			*No Mask vs. Mask*			*No Mask vs. Mask*		
	t(26)	*p*	*d*	*t(26)*	*p*	*d*	*t(26)*	*p*	*d*	*t(26)*	*p*	*d*
Comprehension	−1.36	.184	−0.27	−0.14	.892	−0.03	**3.15**	**.004**	**0.62**	1.03	.316	0.20
Difficulty	−0.11	.912	−0.02	0.55	.589	0.11	**−2.88**	**.008**	**−0.57**	**−2.87**	**.008**	**−0.56**

Table A2. Simple Effects of Stimulus Reconstruction for Mask Effect Across Distractor and Modality Conditions

	Audio Only						*AV*					
	No Distractor			*Distractor*			*No Distractor*			*Distractor*		
	No Mask vs. Mask			*No Mask vs. Mask*			*No Mask vs. Mask*			*No Mask vs. Mask*		
	t(27)	*p*	*d*	*t(27)*	*p*	*d*	*t(27)*	*p*	*d*	*t(27)*	*p*	*d*
Envelope	0.44	.667	0.08	1.10	.280	0.21	1.63	.115	0.31	**4.45**	**<.001**	**0.86**
Phoneme onset	0.01	.99	0.00	0.26	.797	0.05	0.47	.645	0.09	**3.77**	**<.001**	**0.73**
Word onset	0.60	.56	0.12	0.38	.708	0.07	**2.67**	**.013**	**0.51**	**3.52**	**.002**	**0.68**

Acknowledgments

Sound icons were made by *Smashicon* from www.flaticon.com. We would like to thank the whole research team and especially Juliane Schubert for her help as a speaker as well as giving direct and strict but most often correct feedback.

Corresponding author: Chandra L. Haider, Department of Psychology, Paris Lodron Universität Salzburg, Hellbrunner Straße 34, 5020 Salzburg, Austria, or via e-mail: chandraleon.haider@plus.ac.at.

Data Availability Statement

The code to reproduce statistical analyses can be found at gitlab.com/CLH96/visualspeechenhancements. Further code can be supplied upon request.

Author Contributions

Chandra L. Haider: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Validation; Visualization; Writing—Original draft; Writing—Review & editing. Hyojin Park: Conceptualization; Supervision; Writing—Review & editing. Anne Hauswald: Conceptualization; Supervision; Writing—Review & editing. Nathan Weisz: Conceptualization; Funding acquisition; Resources; Supervision; Writing—Review & editing.

Funding Information

This work was supported by the Austrian Science Fund (https://dx.doi.org/10.13039/501100002428), grant number P34237 ("Impact of Face Masks on Speech Comprehension").

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider

gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Aller, M., Økland, H. S., MacGregor, L. J., Blank, H., & Davis, M. H. (2022). Differential auditory and visual phase-locking are observed during audio-visual benefit and silent lip-reading for speech perception. *Journal of Neuroscience*, *42*, 6108–6120. https://doi.org/10.1523/JNEUROSCI.2476-21.2022, PubMed: 35760528

Atcherson, S. R., Mendel, L. L., Baltimore, W. J., Patro, C., Lee, S., Pousson, M., et al. (2017). The effect of conventional and transparent surgical masks on speech understanding in individuals with and without hearing loss. *Journal of the American Academy of Audiology*, *28*, 58–67. https://doi.org/10.3766/jaaa.15151, PubMed: 28054912

Boersma, P., & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *Journal of Neuroscience*, *40*, 1053–1065. https://doi.org/10.1523/JNEUROSCI.1101-19.2019, PubMed: 31889007

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436. https://doi.org/10.1163/156856897X00357

Brodbeck, C., Das, P., Gillis, M., Kulasingham, J. P., Bhattasali, S., Gaston, P., et al. (2021). Eelbrain: A Python toolkit for time-continuous analysis with temporal response functions. *bioRxiv*. https://doi.org/10.1101/2021.08.01.454687

Brown, V. A., Van Engen, K. J., & Peelle, J. E. (2021). Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. *Cognitive Research: Principles and Implications*, *6*, 49. https://doi.org/10.1186/s41235-021-00314-0, PubMed: 34275022

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*, e1000436. https://doi.org/10.1371/journal.pcbi.1000436, PubMed: 19609344

Corey, R. M., Jones, U., & Singer, A. C. (2020). Acoustic effects of medical, cloth, and transparent face masks on speech signals. *Journal of the Acoustical Society of America*, *148*, 2371–2375. https://doi.org/10.1121/10.0002279, PubMed: 33138498

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*, 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015, PubMed: 26490860

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604. https://doi.org/10.3389/fnhum.2016.00604, PubMed: 27965557

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, *36*, 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016, PubMed: 27656026

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A., Molholm, S., & Lalor, E. C. (2021). Linear modeling of neurophysiological responses to naturalistic stimuli: Methodological considerations for applied research. *PsyArXiv*. https://doi.org/10.31234/osf.io/jbz2w

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*, 385–390. https://doi.org/10.3758/BRM.41.2.385, PubMed: 19363178

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, *33*, 1417–1426. https://doi.org/10.1523/JNEUROSCI.3675-12.2013, PubMed: 23345218

Haider, C. L., Suess, N., Hauswald, A., Park, H., & Weisz, N. (2022). Masking of the mouth area impairs reconstruction of acoustic speech features and higher-level segmentational features in the presence of a distractor speaker. *Neuroimage*, *252*, 119044. https://doi.org/10.1016/j.neuroimage.2022.119044, PubMed: 35240298

Hartmann, T., & Weisz, N. (2020). An introduction to the objective psychophysics toolbox. *Frontiers in Psychology*, *11*, 585437. https://doi.org/10.3389/fpsyg.2020.585437, PubMed: 33224075

Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., & Weisz, N. (2018). A visual cortical network for deriving phonological information from intelligible lip movements. *Current Biology*, *28*, 1453–1459. https://doi.org/10.1016/j.cub.2018.03.044, PubMed: 29681475

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*, 393–402. https://doi.org/10.1038/nrn2113, PubMed: 17431404

Homans, N. C., & Vroegop, J. L. (2021). The impact of face masks on the communication of adults with hearing loss during COVID-19 in a clinical setting. *International Journal of Audiology*, *61*, 365–370. https://doi.org/10.1080/14992027.2021.1952490, PubMed: 34319825

Hughes, S. E., Hutchings, H. A., Rapport, F. L., McMahon, C. M., & Boisvert, I. (2018). Social connectedness and perceived listening effort in adult cochlear implant users: A grounded theory to establish content validity for a new patient-reported outcome measure. *Ear and Hearing*, *39*, 922–934. https://doi.org/10.1097/AUD.0000000000000553, PubMed: 29424766

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*, 1–16.

Kratzke, I. M., Rosenbaum, M. E., Cox, C., Ollila, D. W., & Kapadia, M. R. (2021). Effect of clear vs standard covered masks on communication with patients during surgical clinic encounters. *JAMA Surgery*, *156*, 372–378. https://doi.org/10.1001/jamasurg.2021.0836, PubMed: 33704389

Lindsley, W. G., Blachere, F. M., Law, B. F., Beezhold, D. H., & Noti, J. D. (2021). Efficacy of face masks, neck gaiters and face shields for reducing the expulsion of simulated cough-generated aerosols. *Aerosol Science and Technology*, *55*, 449–457. https://doi.org/10.1080/02786826.2020.1862409, PubMed: 35924077

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024, PubMed: 17517438

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. https://doi.org/10.1038/264746a0, PubMed: 1012311

Meredith, M. A., & Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research*, *365*, 350–354. https://doi.org/10.1016/0006-8993(86)91648-3, PubMed: 3947999

Nidiffer, A. R., Cao, C. Z., O'Sullivan, A., & Lalor, E. C. (2021). A linguistic representation in the visual system underlies

successful lipreading. *bioRxiv*. https://doi.org/10.1101/2021.02.09.430299
O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., de Cheveigné, A., & Lalor, E. C. (2021). Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects. *Journal of Neuroscience*, *41*, 4991–5003. https://doi.org/10.1523/JNEUROSCI.0906-20.2021, PubMed: 33824190
Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 156869. https://doi.org/10.1155/2011/156869, PubMed: 21253357
Orf, M., Wöstmann, M., Hannemann, R., & Obleser, J. (2022). Auditory neural tracking reflects target enhancement but not distractor suppression in a psychophysically augmented continuous-speech paradigm. *bioRxiv*. https://doi.org/10.1101/2022.06.18.496558
Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife*, *5*, e14521. https://doi.org/10.7554/eLife.14521, PubMed: 27146891
Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006, PubMed: 25890390
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442. https://doi.org/10.1163/156856897X00366
Puschmann, S., Daeglau, M., Stropahl, M., Mirkovic, B., Rosemann, S., Thiel, C. M., et al. (2019). Hearing-impaired listeners show increased audiovisual benefit when listening to speech in noise. *Neuroimage*, *196*, 261–268. https://doi.org/10.1016/j.neuroimage.2019.04.017, PubMed: 30978494
Reisinger, P., Gillis, M., Suess, N., Vanthornhout, J., Haider, C. L., Hartmann, T., et al. (2023). Neural speech tracking benefit of lip movements predicts behavioral deterioration when the speaker's mouth is occluded. *bioRxiv*. https://doi.org/10.1101/2023.04.17.536524
Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. In J. J. Ohala (Ed.), *Proceedings of the XIVth International Congress of Phonetic Sciences* (pp. 607–610). https://doi.org/10.5282/ubm/epub.13682
Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*, 87–90. https://doi.org/10.1038/416087a, PubMed: 11882898
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215. https://doi.org/10.1121/1.1907309
Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 873–888. https://doi.org/10.1037/0096-1523.30.5.873, PubMed: 15462626
Van Engen, K. J., Dey, A., Sommers, M. S., & Peelle, J. E. (2022). Audiovisual speech perception: Moving beyond McGurk. *Journal of the Acoustical Society of America*, *152*, 3216–3225. https://doi.org/10.1121/10.0015262, PubMed: 36586857
Willmore, B., & Smyth, D. (2003). Methods for first-order kernel estimation: Simple-cell receptive fields from responses to natural scenes. *Network: Computation in Neural Systems*, *14*, 553–577. https://doi.org/10.1088/0954-898X_14_3_309, PubMed: 12938771
Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, *256*, 846–849. https://doi.org/10.1126/science.1589767, PubMed: 1589767