Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Delta-band cortical speech tracking predicts audiovisual speech-in-noise benefit from natural and simplified visual cues

Enrico Varano [a],[1], Mike Thornton [b],[1], Dorothea Kolossa [c], Steffen Zeiler [c], Tobias Reichenbach [b],[*]

[a] *Linguistic Research Infrastructure, University of Zurich, Andreasstrasse 15, 8050 Zürich, Switzerland*

[b] *Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberger Strasse 74, 91052 Erlangen, Germany*

[c] *Electronic Systems of Medical Engineering, Technische Universität Berlin, Einsteinufer 17, 10587 Berlin, Germany*

ARTICLE INFO

Dataset link: https://zenodo.org/record/6855794, https://github.com/Mike-boop/varano-av-analysis

Keywords:
Multisensory integration
Audiovisual speech
Temporal response functions
Electroencephalography

ABSTRACT

Humans comprehend speech in noisy environments more effectively when they can see the talker's facial movements. While the benefits of audiovisual (AV) speech are well established, the specific visual features that support this enhancement and its underlying neural mechanisms remain unclear. Here, we examine how simplified facial signals that preserve structural and dynamic information affect AV speech-in-noise comprehension as well as neural speech tracking. In a behavioural experiment, participants viewed natural or progressively simplified facial videos while listening to short sentences in background noise. Visual stimuli included natural facial recordings, coarse facial outlines, and a simple geometric analogue of visual speech—a disk whose radius oscillated with the speech envelope. In an EEG experiment, we assessed how the progressively simplified visual signals influenced cortical tracking of the speech envelope during continuous AV speech. Behaviourally, we found that comprehension improved with increasing visual detail, while the disk provided no AV benefit, underscoring the importance of dynamic facial cues. For the EEG experiment, only the most natural visual signals enhanced delta-band (1–4 Hz) temporal response functions (TRFs) relative to audio-only stimulation, peaking around 180 ms. This neural enhancement correlated with behavioural benefit across participants. Theta-band effects were weaker and less consistent, suggesting a more limited role in AV integration. Together, these findings highlight the importance of facial detail in AV speech perception, with natural visual input driving stronger delta-band tracking and potentially reflecting alignment of auditory processing with word-level visual cues.

1. Introduction

Understanding speech in noisy environments is aided by integrating auditory information with visual cues from a speaker's facial movements. Audiovisual (AV) integration enhances speech comprehension, especially under difficult listening conditions—a phenomenon known as inverse effectiveness (Sumby and Pollack, 1954; Ross et al., 2007; Grant and Seitz, 2000). People with auditory processing difficulties, such as older adults and those who are hard of hearing, rely more on visual cues and benefit more from them than individuals with typical hearing (Sommers et al., 2005; Puschmann et al., 2019). Yet, the specific visual features and neural mechanisms supporting AV integration remain poorly understood.

1.1. Behaviour and facial features

During audiovisual speech perception, visual *features* from a talker's face must be processed into *cues* that complement the auditory signal. These cues can be temporal — relating to speech rhythm and timing — or categorical, aiding in distinguishing phonemic classes such as place and, to a lesser extent, manner of articulation (Summerfield, 1992; Peelle and Sommers, 2015).

Early studies sought to identify facial regions, such as the lips, jaw, tongue, and other articulators, that convey key visual cues. Summerfield (1979) found that viewing the lips alone provided an AV benefit, but a simplified point-light analogue with just four lip dots did not. Later, Rosenblum et al. (1996) showed that increasing the

[*] Corresponding author.
E-mail address: tobias.j.reichenbach@fau.de (T. Reichenbach).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neuroimage.2025.121654
Received 26 June 2025; Received in revised form 3 December 2025; Accepted 15 December 2025
Available online 16 December 2025

number of lip dots restored the benefit, and systematically adding point-lights to other oral and extra-oral areas further enhanced it. These findings suggest that the most important visual features are concentrated in the oral region, and the ability to form a mouth-area percept is especially crucial. Nevertheless, extra-oral facial features still contribute to AV benefit: subsequent research has indicated that overall facial configuration could be an important feature for AV speech perception (Rosenblum et al., 2000; Thomas and Jordan, 2004).

Surface-level features have also been examined in the literature. McCotter and Jordan (2003), Jordan et al. (2000) found that colour had no effect on AV speech enhancement, and luminance inversion produces only a small reduction in benefit. Erber (1979) explored texture by presenting participants with AV speech in which the talker's face was positioned behind a Plexiglas sheet. The distance between the face and the sheet was varied to parametrically control the degree of blurring observed by participants. As blurring increased, lipreading performance in quiet conditions rapidly deteriorated, a result the authors attributed to the fact that only gross oral kinematics (and thus gross temporal cues) were available under high levels of blur.

The Plexiglas manipulation in Erber (1979) degraded all facial texture, including edges — sharp transitions in luminance or contrast — that delineate facial structure. Vision science has long highlighted edges as crucial for spatial and object perception (Gibson, 1950; Troscianko et al., 2008). In visual speech, they may help delineate articulators (e.g., lips, jawline) to support place-of-articulation cues, and enhance temporal precision by making movement onsets more salient. Despite their theoretical importance, facial edges have yet to be systematically isolated in AV speech perception research.

Despite the adverse effects of heavy blurring in Erber (1979), a substantial AV benefit remained for both adults with typical hearing and children with severe hearing loss. This was likely due to availability of gross oral kinematics even under heavy blur, which could have provided cues for the timing of speech. Appealing to this idea, some researchers have explored whether abstract geometric visualisations of the timing of speech can result in an AV benefit (Yuan et al., 2020, 2021b,a; Summerfield, 1979; Benz et al., 2025b). In these studies, the visual signals were modulated by the acoustic amplitude envelope, which has been shown to be moderately-to-strongly-correlated with the area of mouth opening and could serve as a suitable proxy for this aspect of visual speech (Grant and Seitz, 2000; Chandrasekaran et al., 2009). Results in this direction remain mixed, and the specific visual-acoustic conditions required to elicit a reliable AV benefit remain unknown.

1.2. Neurobiology of audiovisual integration

Neurophysiological and neuroimaging studies demonstrate that AV speech integration relies upon a distributed network involving classical multisensory regions and early sensory cortex (Kayser et al., 2010; Meredith and Stein, 1986). These findings suggest a multi-stage model involving both bottom-up convergence and top-down feedback (Choi et al., 2018; Stein and Stanford, 2008). Other authors have suggested a role of neuronal oscillations in multisensory integration (Hickok and Poeppel, 2007). For example, visual speech could entrain low-frequency oscillations — particularly in the delta (1–4 Hz) and theta (4–8 Hz) bands — which modulate auditory cortical excitability and align auditory processing with visual timing cues (Park et al., 2016; Schroeder et al., 2008).

Human electro-encephalography (EEG) and magneto-encephalography (MEG) studies increasingly use naturalistic stimuli, such as audiobooks paired with visual facial recordings, to investigate AV integration. Neural encoding and decoding models can be used to characterise the neural representation of continuous speech features such as the amplitude envelope (Holdgraf et al., 2017; Crosse et al., 2021). Indeed the amplitude envelope of auditory speech is robustly tracked in neural responses, and preliminary evidence suggests that subtle eye movements may align with this feature as well (Ding and Simon, 2014; Jin et al., 2018; Gehmacher et al., 2024). In one study, Crosse et al. (2016b) used decoding models to compare neural tracking of the speech envelope under AV versus unisensory stimulation, finding a superadditive increase in envelope tracking in the AV condition. Their neural measure of AV gain showed the strongest effects in the delta band (1–4 Hz), coinciding particularly with the timescale of words. Subsequent work from the same group revealed integrative effects for higher-level linguistic features as well O'Sullivan et al. (2021). These findings extend behavioural results, highlighting the role of both temporal and categorical cues in AV speech integration.

Some continuous-speech studies employ manipulated visual stimuli to examine the effects of particular features on AV integration. In an MEG study, Reisinger et al. (2025) found that neural tracking of mouth movements increased under noisy conditions, and that individuals with stronger visual tracking showed greater declines in comprehension when the oral region was occluded, thus highlighting the potential for neural data to provide an objective index of behaviour. Similarly, Wikman et al. (2024) applied digital blur to speakers' faces in a cocktail-party setting, finding texture degradation reduced both attentional enhancement of auditory tracking as well as a neural marker of semantic processing. These results suggest that visual cues support not only attention but also higher-level linguistic processing.

1.3. Motivations for this study

The present study examined the behavioural and neural consequences of simplifying visual speech signals, not through global manipulations like blurring, but by removing surface texture while preserving the full structural integrity of the face (Fig. 1). In addition to assessing behavioural outcomes, we investigated neural envelope tracking during AV speech perception with natural and simplified moving faces. To our knowledge, no previous neural tracking study has explored how simple geometric representations of the speech envelope influence audiovisual speech encoding. To this end, we introduced a highly simplified visual signal: an animated disk whose radius oscillated in synchrony with the amplitude envelope of the auditory speech.

2. Materials and methods

2.1. Experimental design and data analysis

We combined behavioural measures of speech-in-noise comprehension with EEG recordings of neural responses to audiovisual speech. In both experiments, participants were presented with the same set of visual stimuli, ranging from natural facial recordings to highly simplified geometric signals. The visual signals (shown in Fig. 1) varied monotonically in their degree of naturalism. The face-like stimuli preserved global facial structure and varied systematically in edge coarseness and fine textural detail. Speech was delivered in background noise at a fixed signal-to-noise ratio (SNR). In the behavioural experiment, participants were stimulated with audio-only and audiovisual sentences. In the EEG experiment, they also viewed visual-only stimuli to enable calculation of neural audiovisual gain by comparing AV responses to unimodal responses, following the approach of Crosse et al. (2016b). EEG data were analysed using linear models relating the speech envelope to EEG signals filtered in different frequency bands, and the models were then compared to the outcomes of the behavioural speech comprehension experiment.

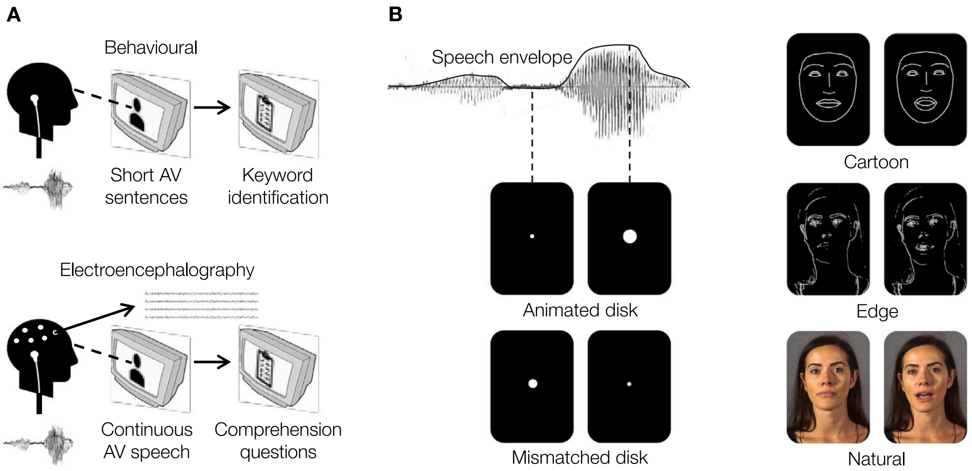

Fig. 1. Schematic of the experimental setup. **A**, We assessed the comprehension of audiovisual speech in noise by presenting subjects with syntactically correct but semantically unpredictable sentences. We then measured neural EEG responses to continuous audiovisual narratives with EEG. **B**, For the visual stimuli, we employed simplified signals as well as natural talking faces. The simplified visual stimuli were, in order of least-natural to most-natural, a disk with a temporally-varying radius that was proportional to the speech envelope ("Animated disk"), a cartoon that consisted of outlines of the natural face ("Cartoon"), an edge-filtered version of the natural face ("Edge"), and the original facial recordings ("Natural"). We also included a disk with a diameter that varied in relation to the envelope of an unrelated speech signal ("Mismatched disk") as a control signal. The facial signals preserved global facial structure whilst varying the level of fine textural detail and facial edge fidelity.

2.2. *Participants*

Behavioural data and EEG recordings were obtained from 24 participants. Volunteers were young adults between the ages of 18 and 29 with a mean age of 22 years. Eleven participants self-identified as female and thirteen as male in a written questionnaire. All participants also self-identified as right-handed native English speakers with no reported history of neurological disorders or hearing difficulties. The EEG data of two female and three male participants were unfortunately lost due to a faulty connection in the EEG recording equipment. For consistency, the behavioural data for these participants was left out of the analysis as well.

Participants were free to choose whether to have the behavioural data as well as the EEG data measured on the same day, or on two different days within two weeks. Of the 19 participants included in the final dataset, 9 completed both the behavioural and EEG sessions on the same day, and 11 completed them on separate days. To assess potential bias, we compared overall behavioural performance between these subgroups and found no significant differences (same: n = 9, mean = 60.48, sd = 8.62, diff: n = 10, mean = 58.35, sd = 6.00, Welch t-test: stat = 0.618, p = 0.5465). We also looked at performance on the questions regarding the context or specific details of the story which were asked after each passage of the EEG study (same: n = 9, mean = 1.90, sd = 0.38, diff: n = 10, mean = 1.94, sd = 0.21, Welch t-test: stat = –0.263, p = 0.7971) and found no significant difference.

The research was approved by the Imperial College Research Ethics Committee. All participants gave their written informed consent before the experiment and were compensated for their time.

2.3. *Audio stimuli*

To increase listening effort and prevent ceiling effects in the behavioural experiment, the target speech was presented in speech-shaped noise. The stationary noise stimuli were created by randomising the phase of the concatenated speech signals in the Fourier domain. Specifically, an FFT was applied to the full time series of each the concatenated speech waveform of a randomly selected subset of the corpus, the phase spectrum was randomised, and the inverse FFT was taken to generate the noise signal, which preserved the long-term spectral profile of the original utterance.

The signal-to-noise ratios (SNRs) of the noisy speech signals were −8.82 dB and −2 dB for the behavioural and EEG experiments, respectively. In our prior work, we found that these values resulted in mean comprehension levels of about 50% (under auditory stimulation only) for both the behavioural task as well as for the longer narratives that we employed in the EEG experiment (Varano et al., 2022, 2023). The higher SNR of the EEG recordings reflected the fact that comprehending the complex stories was more difficult than understanding the keywords in the single sentences of the behavioural experiment.

2.4. *Visual stimuli*

Five different visual conditions were employed in the behavioural and EEG experiments (Fig. 1). All visual signals were presented against a dark background. In the "Natural" condition, a video recording of the talker's face was presented. This was the most natural and highly textured visual signal that we employed.

The second-most-natural visual signal was the "Edge" signal, generated by applying the Canny edge detection algorithm (Canny, 1986) to the natural facial recordings (using the FFmpeg implementation: `ffmpeg -i input -filter:v edgedetect=low=0.1:high=0.4:mode=wires out`). This processing produced videos consisting of moving white edges that preserved global facial structure while removing most of the fine textural detail. The mouth opening area was clearly marked, but the internal articulators (tongue and teeth) were often difficult to discern.

The least natural face-like stimulus was the "Cartoon" signal, generated by processing the natural facial videos using the OpenFace 2.0 toolkit (Baltrušaitis et al., 2016). Facial landmarks corresponding to features such as the eyes and lips were connected with white lines to create a simplified, cartoon-like animation. Compared to the "Edge" signal, this representation conveyed facial structure through more coarse edges and omitted smaller features that survived Canny edge detection. As a result, the Cartoon stimulus was devoid of texture.

The simplest visual signal, the "Animated disk", consisted of a white disk with a radius proportional to the temporal envelope of the accompanying speech signal. The envelope was computed by low-pass filtering the absolute value of the analytic speech signal (cutoff frequency 12 Hz, 4th order Butterworth filter, zero-phase behaviour through forward–backward filtering).

In the “Mismatched disk” condition, we presented an animated disk which was derived from a speech signal unrelated to the actual auditory stimulus; both envelopes were derived from material narrated by the same talker, however. In the behavioural experiment, the visually-presented envelope was taken from a different, randomly-selected sentence in the GRID audiovisual corpus; in the EEG experiment, it corresponded to the speech envelope from a different chapter of the same audiobook in the AVbook corpus. The purpose of this sham stimulus was to distinguish which effects of the animated disk condition were due to the congruent nature of the audio and the visual signal, and which resulted merely from the fact that there was an animated disk presented visually.

2.5. Presentation hardware and software

The audio stimuli were delivered diotically at a sound pressure level of 70 dB(A) using ER-3C insert earphones (Etymotic, USA) through a high-performance sound card (Xonar Essence STX II, Asus, USA). The sound level was calibrated with a Type 4157 ear simulator (Brüel & Kjær, Denmark).

The visual signals were delivered through a 144 Hz, 24-inch flat-screen monitor (24GM79G, LG, South Korea) set at a refresh rate of 120/1.001 Hz. The audiovisual latency was calibrated using two photodiodes (Photo Sensor, BrainProducts, Germany) and an acoustic adapter (StimTrak, BrainProducts, Germany) with which the timing of the continuous audiovisual speech material was recorded (Varano et al., 2023). A negligible audiovisual asynchrony of 8 ± 4 ms lag of the audio versus the visual stimuli was thus achieved.

2.6. Design of the behavioural experiment

Speech-in-noise comprehension can be measured with high accuracy when presenting subjects with semantically unpredictable sentences and scoring the percentage of words that they understood correctly. We therefore employed the meaningful, syntactically valid, but semantically unpredictable sentences from the GRID corpus to measure the audiovisual speech-in-noise benefit for the five different visual conditions (Cooke et al., 2006). GRID corpus sentences are six-word commands of the form: `verb COLOUR preposition LETTER/DIGIT adverb` (such as `place GREEN at P 2 now`), and are composed from a limited dictionary (Table 1). The participants were asked to report the colour, letter, and digit in each sentence, with the remaining words acting as syntactic cues. Of the 33 speakers available in the GRID corpus, four were selected for their lack of a strong accent (speakers 12, 19, 24, and 29, two of them male, two female). Speakers were rotated pseudorandomly selected for each trial such that participants could not predict the next speaker but such that the distribution was balanced across conditions. Each sentence was chosen randomly from a pool of all 1000 sentences from each of the four speakers, and presented in speech-shaped noise at an SNR of −8.82 dB.

The experiment consisted of 324 three-second GRID sentences (trials), which were presented in blocks of six sentences. All sentences within a single block were presented with the same visual condition; this was either one of the five conditions described in Section 2.4, or else the block was an audio-only block. The order of the six possible audiovisual conditions was randomised after every sixth block, and was different for every participant. Each participant therefore listened to 54 sentences in each of the six conditions. The participants took a brief rest for one minute between successive rounds of nine blocks.

Table 1
Structure of GRID corpus sentences employed in the behavioural experiment.

Command	Colour[a]	Preposition	Letter[a]	Digit[a]	Adverb
Bin	Blue	At	A–Z	0–9	Again
Lay	Green	By	excluding		Now
Place	Red	In	W		Soon
Set	White	With			Please

[a] The keywords on which participants were scored.

2.7. Behavioural data acquisition and analysis

A graphical interface was shown after the presentation of each sentence. Participants were presented with incomplete sentences, with the correct non-keyword terms displayed, and asked to complete the sentence by filling in the keyword terms that they had heard. This setup allowed for blind data collection and automatic computation of the comprehension scores. The setup was also blind with respect to the “Animated disk” and the “Mismatched disk” stimuli types: participants were not informed of the mismatched nature of half of these audiovisual stimuli before the start of each block.

The score for each trial was expressed as the fraction of keywords correctly identified. The final behavioural score s for each type of audiovisual stimulus and for each participant was then computed by averaging across the different trials and rounds. For each participant, the audiovisual gain $G_{av}^{\text{beh.}}$ for a certain visual condition v was computed as the difference between the corresponding audiovisual comprehension score s_{av} and the comprehension score for the audio-only condition, s_a, expressed as the percentage of the total possible improvement (Sommers et al., 2005):

$$G_{av}^{\text{beh.}} = 100 \times (s_{av} - s_a)/(1 - s_a). \tag{1}$$

Homoscedasticity and normality checks with subsequent two-sided paired-samples t-tests were conducted to identify statistically significant differences in the audiovisual gain elicited by the different types of visual stimuli. The tests were conducted on all combinations of conditions, as well as between each condition and zero, corresponding to no gain. A two-step Benjamini–Hochberg procedure was employed to control the false discovery rate (FDR) for multiple comparisons.

2.8. Design of the electroencephalography experiment

The presentation of the single sentences in the behavioural study resulted in about 2:40 min of speech presentation per audiovisual condition. Because the EEG measurement of neural speech tracking requires recording durations of about ten minutes or more for reliable results, this amount of speech data was not sufficient for this purpose. We therefore presented subjects with continuous audiovisual narratives, namely passages 3 through 57 of the AVBook corpus that we recorded recently (Varano et al., 2023). This corpus consists of audiovisual recordings of two actors reading the same 59 passages of a non-fiction historical account into a high-frame-rate video camera. The 55 selected passages had an average duration of 112 s. To reduce variance in the data, only the recordings of the female actress were used in this experiment, since she exhibited particularly clear articulation, as assessed through subjective ratings of clarity and naturalness during pilot testing. The first two passages of the corpus were presented to participants in the audio-only modality during EEG setup.

Each passage was presented in one of three modalities: audio-only (A), audiovisual (AV) or video-only (V). In the audio-only and audiovisual conditions, the speech signals were combined with speech-shaped noise at an SNR of −2 dB as described above. The five visual conditions described in Section 2.4 gave rise to five video-only conditions and to five audiovisual conditions (the “Animated disk” and the “Mismatched disk” stimulus types are equivalent in the visual-only modality – this was overlooked and both were recorded anyway.). The EEG recordings therefore utilised a total of eleven conditions across all three modalities, with a total of five passages (approximately 560 s) per condition.

For each participant, the order of conditions was pseudo-randomised in a way that avoided presenting more than two passages in the visual-only modality consecutively. In order to spread out each condition evenly across the duration of the experiment, we also ensured that passages of the same condition were separated by at least five and no more than fifteen other passages.

To encourage the participants to engage with the story presented in the EEG experiment, four multiple-choice questions regarding the context or specific details of the story were asked after each passage. These questions were not designed to be used as a behavioural measure of speech comprehension, and the answers were occasionally obvious or predictable. A short written summary, presented after the questions, allowed the participants to catch up with the story after each passage in case the background noise or the visual-only modality had impeded their understanding.

2.9. Electroencephalography data acquisition

The EEG measurements were obtained from a multichannel EEG amplifier (actiCHamp, BrainProducts, Germany) and a 64-channel active electrode cap (Easycap, BrainProducts, Germany). For the purpose of aligning the original stimuli and EEG recordings during offline processing, the audio presented to the participants was recorded through an auxiliary input to the amplifier using the StimTrak pass-through device (BrainProducts, Germany). A timing signal from the display was also recorded through an auxiliary channel using a photodiode (Photo Sensor, BrainProducts, Germany). The recordings were obtained at a sampling rate of 1 kHz with a frontal reference (Fz). The impedances of the electrodes were typically below 10 kΩ.

2.10. Data pre-processing

The speech envelope for each passage was extracted by applying a Hilbert transform to the clean speech signal and taking its absolute value. The resulting envelopes were then low-pass filtered below 12 Hz using a one-pass finite impulse response (FIR) filter (order 76,947, delay-compensated) and subsequently downsampled to 50 Hz. To ensure precise temporal alignment between the speech envelope and the EEG recordings, cross-correlation was performed between the recorded audio (captured via StimTrack, see Section 2.5 above) and the clean speech signal. Additionally, synchronisation between the auditory and visual stimuli was verified using the photodiode channel (Section 2.5).

The EEG data pre-processing was performed in MATLAB using EEGLab 2022.0 and ICLabel 1.3 (Delorme and Makeig, 2004; Pion-Tonachini et al., 2019). First, a band-pass filter (1–80 Hz) was applied before resampling the EEG data to 200 Hz. Next, bad channels were automatically detected and removed using the EEGLAB function `clean_artifacts()`, which flags channels with abnormal statistics (low correlation to other channels, poor predictability, or atypical spectral content), followed by interpolation of the removed channels with the EEGLAB function `eeg_interp()`, which performs a spherical spline interpolation based on the original 64-channel montage. The function `clean_artifacts()` was called with the default arguments of EEGLab (with the exception of the highpass filter options, which were disabled since the EEG measurements were already high-pass filtered). These default parameters correspond to conservative labelling of EEG channels as "bad". Finally, automatic rejection of artifact ICA components was performed using ICLabel: specifically, independent components classified as artifactual with a probability of at least 85% were rejected; artifacts labelled as "Brain" or "Other" were not rejected. This threshold corresponds to a conservative rejection of independent components. To improve computational efficiency when fitting the encoding models, both the EEG and speech envelope signals were subsequently resampled to 50 Hz using an anti-aliasing polyphase filter. The full pre-processing pipeline is available at https://github.com/enricovara/enricoICA.

We examined neural responses in two frequency bands relevant to neural envelope tracking: the delta band (1–4 Hz) and the theta band (4–8 Hz). Narrow-band EEG signals were extracted using MNE's default `mne.io.Raw.filter()` function, which applies a zero-phase, Hamming-window FIR filter. Zero-phase filtering was ensured by performing a forward and backward pass. All transition bandwidths were set to 2 Hz. For the delta band, only a low-pass filter with an upper passband edge at 4 Hz was applied, as the EEG data had already been high-pass filtered above 1 Hz in the band-pass filtering described above. Finally, EEG signals in each frequency band were average-referenced and normalised on a trial-by-trial basis by subtracting the time-averaged signal of each EEG channel and dividing by its standard deviation.

2.11. Forward model

We modelled the relationship between EEG recordings and the temporal envelope of speech using temporal response functions (TRFs) (Crosse et al., 2016a). The speech envelopes were derived from the audio signal that the participants heard. In the mismatched disk condition, two separate envelopes were extracted: one from the auditory stimulus and another corresponding to the mismatched video signal.

The TRFs represent the impulse response of EEG signals across a range of latencies relative to the temporally aligned speech envelope. They can be estimated as the parameters β_{ij} in the linear convolutive model defined in Eq. (2):

$$\hat{y}_i(t) = \sum_{j=1}^{L} \beta_{ij} x(t - \tau_j). \tag{2}$$

Here, $\hat{y}_i(t)$ denotes the predicted EEG signal at channel i and temporal sample t, while $x(t)$ represents the speech envelope at temporal sample t. The parameters τ_j correspond to the set of latencies (in samples) at which the impulse response function is explicitly modelled by the TRF, and L defines the order of the TRF, i.e. the total number of latencies considered. In our experiments, we modelled $L = 125$ TRF latencies ranging from -1 s to 1.5 s.

TRFs were estimated using ridge-regularised least-squares regression. The solution to this optimisation problem is given in Eq. (3):

$$\hat{\vec{\beta}}_i = (X^T X + \lambda I)^{-1} X^T \vec{y}_i, \tag{3}$$

where $\hat{\vec{\beta}}_i$ is a vector of dimension $L \times 1$ which describes the estimated temporal response of the EEG measurements from channel i. In Eq. (3), $\vec{y}_i$ is a vector of EEG signals recorded at channel i, and X is a matrix whose entry $X_{lm} = x(l - \tau_m)$ represents a delayed observation of the speech envelope. The parameter λ is the ridge regularisation hyperparameter, which stabilises the solution when the covariance matrix $X^T X$ exhibits multi-collinearity (for example due to autocorrelation in the speech envelope signal from which X is derived), and I is the $L \times L$ identity matrix. It is common to normalise the regularisation hyperparameter by expressing it as $\lambda = \lambda_n.e_m$, where e_m is the mean eigenvalue of the covariance matrix $X^T X$, and λ_n is the normalised regularisation parameter (Biesmans et al., 2017). In this work, we used a fixed value of $\lambda_n = 1$.

2.12. Forward model evaluation

We fitted TRFs for all experimental conditions and participants, excluding those who were removed due to technical issues with the EEG apparatus. The estimated TRFs for individual participants were then averaged to compute grand-average TRFs for each condition. To assess the statistical significance of the grand-average TRFs, we employed permutation-based cluster tests, a non-parametric statistical approach commonly used to analyse evoked responses (Maris and Oostenveld, 2007).

Permutation-based cluster tests work by identifying prominent, spatio-temporally connected "clusters" in the TRFs. We computed a two-tailed, single-sample t-statistic for every grand-average TRF at each EEG channel and latency. Clusters were formed by grouping neighbouring t-statistics that exceeded a predefined threshold. The cluster test statistic is defined as the size (number of spatio-temporal samples) of the largest cluster.

To determine statistical significance, we generated a null distribution for the cluster statistic by randomly permuting the signs of participant-specific TRFs and repeating the procedure outlined above. By performing this process 5000 times, we obtained a robust estimate of the null distribution; a p-value for the cluster statistic was then obtained by computing its percentile relative to this distribution..

A threshold of $|t| = 2.878$ was applied to the t-statistics, corresponding to $p = 0.01$ for a single-sample two-tailed t-test with 18 degrees of freedom. The permutation-based cluster tests were carried out using the function `mne.stats.spatio_temporal_cluster_1samp_test` available in the MNE-Python package (Gramfort et al., 2014).

2.13. Neural correlate of behaviour

In our initial analysis plan, we aimed to first estimate unisensory TRFs by relating the acoustic amplitude envelope of speech to EEG signals acquired under auditory-only or visual-only stimulation. The summation of these unisensory response functions would then be compared to a third TRF obtained from EEG data collected under bi-modal (audiovisual) stimulation. This analysis plan reflects the approach put forward by Crosse et al. (2016b), who were able to derive a reliable measure of audiovisual gain from such a comparison. Crucially, those authors found that such an additive sum of unisensory responses could capture a high proportion of the variance of the full audiovisual TRF. However, for our data, we noticed that, whilst the audiovisual TRFs exhibited a strong likeness to the audio-only TRFs, the visual-only TRFs possessed spatio-temporal characteristics which were not reflected in the audiovisual TRFs.

To derive a more meaningful measure of AV gain, we instead compared AV TRFs directly to audio-only TRFs. With this approach, we observed AV enhancement in the delta band at a latency of 180 ms. From this, we developed a measure of neural audiovisual gain as described in the following.

We computed "difference TRFs" $\Delta\vec{\beta}_i$ by subtracting the audio-only TRF from the corresponding audiovisual TRF:

$$\Delta\hat{\vec{\beta}}_i = \hat{\vec{\beta}}_i^{av} - \hat{\vec{\beta}}_i^{a} \tag{4}$$

at each EEG channel i. We then extracted the mean activity of the difference TRF over two scalp regions. These were first the temporal regions, i.e. the channels from the set $\mathcal{T}$ = {F5, F7, F6, F8, FT9, FT7, FT8, FT10, FC5, FC6, T7, C5, C6, T8, TP9, TP7, CP5, CP6, TP8, TP10, P5, P7, P8}, and second the central region, i.e. channels from the set $\mathcal{C}$ = {F3, F1, F2, F4, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, CP4, P3, P1, Pz, P2, P4} (Fig. 7). The activities over these scalp regions were averaged across temporal delays ranging from τ_{min} = 120 ms to τ_{max} = 240 ms. We then defined the neural audiovisual gain as the difference between the averaged difference TRF in the temporal regions and in the central region:

$$G_{av}^{\text{neural}} = \sum_{\tau=\tau_{min}}^{\tau_{max}} \left(\frac{1}{|\mathcal{T}|} \sum_{i\in\mathcal{T}} \Delta\hat{\vec{\beta}}_i - \frac{1}{|\mathcal{C}|} \sum_{j\in\mathcal{C}} \Delta\hat{\vec{\beta}}_j \right) \tag{5}$$

The temporal delays from 120 ms to 240 ms were chosen based on the temporal extent of the clusters identified through permutation-based cluster tests applied to the difference TRFs, and the scalp topographies of the difference TRFs within this time window informed the choice of the temporal and central regions of interest.

Table 2
Behavioural audiovisual gain by video type.

Condition	AV Enhancement (%)		p-value	t statistic
	Mean	SEM	(FDR corrected)	
Animated disk	−1.866	2.581	4.91×10^{-1}	−0.704
Mismatched disk	−4.282	3.846	3.66×10^{-1}	−1.08
Cartoon	11.93	2.398	$\mathbf{2.17 \times 10^{-4}}$	4.84
Edge	29.98	3.476	$\mathbf{3.08 \times 10^{-7}}$	8.39
Natural	50.66	2.349	$\mathbf{2.08 \times 10^{-13}}$	21.0

Statistical analysis of the audiovisual gain in speech comprehension, for the different types of visual stimuli. The p-values are corrected for multiple comparisons through a two-step Benjamini–Hochberg procedure to control the false discovery rate.

Table 3
Pairwise comparisons in audiovisual gain across video types.

Comparison		p-value (FDR corrected)	t statistic
Animated disk	Mismatched disk	3.79×10^{-1}	0.902
Animated disk	Cartoon	$\mathbf{7.54 \times 10^{-4}}$	−4.15
Animated disk	Edge	$\mathbf{1.11 \times 10^{-7}}$	−8.75
Animated disk	Natural	$\mathbf{6.20 \times 10^{-11}}$	−15.7
Mismatched disk	Cartoon	$\mathbf{9.14 \times 10^{-4}}$	−4.01
Mismatched disk	Edge	$\mathbf{7.24 \times 10^{-8}}$	−9.26
Mismatched disk	Natural	$\mathbf{2.50 \times 10^{-10}}$	−13.7
Cartoon	Edge	$\mathbf{8.10 \times 10^{-5}}$	−5.23
Cartoon	Natural	$\mathbf{2.50 \times 10^{-10}}$	−13.5
Edge	Natural	$\mathbf{1.11 \times 10^{-7}}$	−8.78

Pairwise comparisons of the audiovisual gain in speech comprehension between the different types of visual stimuli. The p-values are corrected for multiple comparisons through a two-step Benjamini–Hochberg procedure to control the false discovery rate.

To examine the relationship between neural and behavioural audiovisual gain, we conducted a repeated-measures Pearson correlation test using the Pingouin package in Python (R., 2018). This test accounts for inter-participant variability and provides both a correlation coefficient and an associated p-value.

3. Results

3.1. Behavioural experiment

We first investigated the results of the behavioural experiment, which are shown in Fig. 2. We found that all three visual stimuli that were derived from the face-like visual signals, namely the "Cartoon", "Edge" and "Natural" type, significantly improved comprehension of the GRID sentences when compared to the comprehension of the audio signal alone (Table 2, two-tailed, one-sample t-test, FDR correction for multiple comparisons). On the other hand, neither the "Animated disk" nor the "Mismatched disk" type resulted in a significant speech-in-noise benefit.

To assess differences between conditions, we performed a repeated-measures ANOVA with condition as a within-subject factor, which showed a significant main effect of visual stimulus type on comprehension ($F(4, 72) = 93.22$, $p < 0.001$). Post-hoc two-tailed, paired-samples t-tests (FDR-corrected) revealed significant differences between almost all pairs of visual stimuli, aside from the comparison between the "Animated disk" and the "Mismatched disk" type (Table 3, two-tailed, paired samples t-test, FDR correction for multiple comparisons). In particular, speech comprehension increased significantly for the "Cartoon" type as compared to the simple disk-based stimuli, increased further for the "Edge" type, and was higher yet for the natural talking face recording. These results demonstrate that speech comprehension increased monotonically with the visual signal's level of naturalness, reflecting a graded benefit from simplified representations of facial structure to fully natural facial cues.

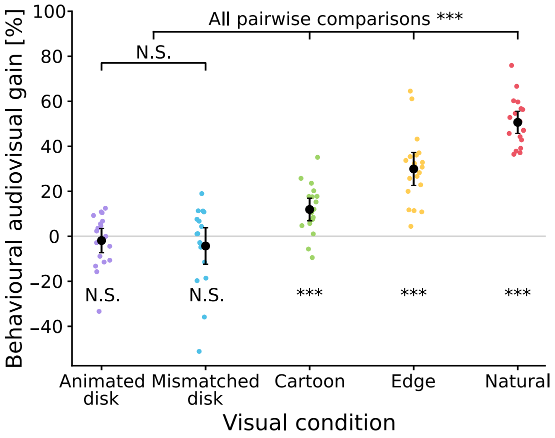

Fig. 2. Outcomes of the behavioural experiment. We measured the subjects' ability to identify words in noise using sentences from the GRID corpus (Cooke et al., 2006). The behavioural audiovisual gain was formed by comparing the comprehension scores in the audiovisual stimulation conditions to those in the audio-only stimulation condition. The results show a larger behavioural audiovisual gain for visual stimuli which appear more natural, whilst the simple geometric signals provide no audiovisual benefit.

3.2. EEG encoding models for the disk conditions

To investigate how neural activity tracks the speech envelope for the disk-based visual stimuli, we computed temporal response functions (TRFs) from EEG recordings in both the delta (1–4 Hz) and theta (4–8 Hz) frequency bands. The stimulus conditions included an audio-only condition, as well as an audio-visual condition in which the visual stimulus was a dynamically changing disk whose radius was proportional to the speech envelope (the "Animated disk" condition). There was also a "Mismatched disk" condition, in which the radius of the disk corresponded to a different auditory stimulus than the one actually presented to the participant.

For each of the audio-only and the "Animated disk" conditions, TRFs were estimated by relating the band-passed EEG signals to the amplitude envelope of the speech that the participants listened to. Similarly, for the visual-only condition, TRFs were estimated to relate the EEG signals to the speech envelope that was driving the disk animations. For the "Mismatched disk", which served as a control condition, the speech that the participants listened to was different to the speech which drove the visual signal. Accordingly, two TRFs were estimated in this condition: an "auditory TRF", which related the EEG signals to the amplitude envelope of the speech to which the participants listened; and a "visual TRF", which related the EEG signals to the amplitude envelope which was driving the disk animation. Comparisons were subsequently made between the audio-only TRF and the auditory TRF derived from the "Mismatched disk" condition; the visual-only TRF and the visual TRF derived from the "Mismatched disk" condition; the "Animated disk" TRF and the additive sum of the two TRFs derived from the "Mismatched disk" condition. These comparisons serve to highlight any differences between differential and integrated responses to congruent and incongruent stimuli. Differences between TRFs were tested for statistical significance via permutation-based cluster tests. The outcomes of these analyses for the delta- and theta bands are presented in Figs. 3 and 4, respectively.

First, to assess the influence of the incongruent visual signal on auditory envelope tracking, we assessed the difference between the audio-only TRFs and the auditory TRFs in the "Mismatched Disk" condition. Despite the presence of mismatched visual signals in the latter stimulus, we found no significant difference between these TRFs, neither in the delta nor in the theta band (permutation-based cluster tests. The *p*-values were Bonferroni-corrected for multiple comparisons. Delta-band: $p = 0.6462$; cluster statistic = 38. Theta-band: $p = 0.6606$; cluster statistic = 21).

However, the same comparison between the TRF for the visual-only "Animated disk" condition and the visual TRF in the "Mismatched Disk" audiovisual condition did yield a significant difference in both frequency bands (Delta-band: $p = 0.0108$; cluster statistic = 78. Theta-band: $p = 0.0162$; cluster statistic = 41). Both TRFs as well as their difference showed similar patterns and topographies, with the TRFs in the visual-only condition having smaller amplitudes than those for the audiovisual stimulus and being dominated by occipital channels. This suggests that both TRFs reflect largely neural responses to the dynamic changes in the visual stimuli, with larger responses occurring when only the visual stimulus was presented unimodally.

We also compared the full audiovisual TRF in the "Animated Disk" condition to an additive model formed from the auditory and visual TRFs from the "Mismatched Disk" condition. We found no significant differences in either frequency band (permutation-based cluster tests. The *p*-values were Bonferroni-corrected for multiple comparisons. Delta-band: $p = 1.0$; cluster statistic = 20. Theta-band: $p = 1.0$; cluster statistic = 12). This suggests that the TRFs in the audiovisual "Animated disk" condition can be approximated by the linear superposition of the auditory and visual TRFs of the "Mismatched Disk" stimulus.

3.3. EEG encoding models for the face-like conditions

We also computed delta-band (1–4 Hz) and theta-band (4–8 Hz) TRFs for the three face-like visual conditions: The "Natural" signal, which consisted of the original recording of the talker; the "Edge" signal, consisting of a low-texture, edge-filtered version of the natural signal; and an even more heavily simplified "Cartoon" signal which was devoid of texture and comprised only simplified, key facial structures.

As in Section 3.2, TRFs were estimated to relate the acoustic amplitude envelope to EEG measurements obtained under audio-only stimulation. The envelopes of the auditory speech were also related to the EEG measurements obtained under audiovisual stimulation (note that all audiovisual conditions were congruent for the face-like visual signals). For the visual-only conditions, the acoustic amplitude envelopes of the speech (which was seen, but not heard) were related to the EEG measurements. Statistical comparisons were made between the audiovisual TRFs and the audio-only TRFs via permutation-based cluster tests, as shown in Figs. 5 and 6 for the delta- and theta bands, respectively. The lack of correspondence between the visual-only TRFs and the "difference TRFs" (obtained by subtracting the audio-only TRFs from the audiovisual TRFs) can be visually assessed in these figures.

For the delta-band TRFs (Fig. 5), we observed that the AV TRFs from all three face-like conditions were highly similar to the audio-only TRFs. This motivated us to look at AV-A "difference TRFs", to see if we could identify any differences between the audiovisual and audio-only responses. We assessed the difference TRFs for statistical significance using permutation based cluster tests, which returned a positive result for the "Natural" condition only ($p = 0.0018$, cluster statistic = 87; *p*-value Bonferroni-corrected for multiple comparisons). This cluster test highlighted a prominent cluster near a latency of 180 ms. Unlike for the "Animated disk" condition, the difference TRFs did not resemble the visual-only TRFs, which exhibited markedly different morphologies and predominantly occipital scalp topographies.

For the theta-band TRFs (Fig. 6), we again observed that the audiovisual TRFs resembled the audio-only TRFs. As with the delta-band TRFs, we analysed the $AV - A$ difference TRFs to see if there was a difference between the theta-band TRFs during audio-visual and audio-only stimulation. The only significant $AV - A$ difference TRF was obtained in the "Cartoon" condition (permutation-based cluster test; $p = 0.0252$; cluster statistic = 37; *p*-value Bonferroni corrected for multiple comparisons). None of the visual-only TRFs exhibited statistical significance in the theta band after Bonferroni-correction for multiple

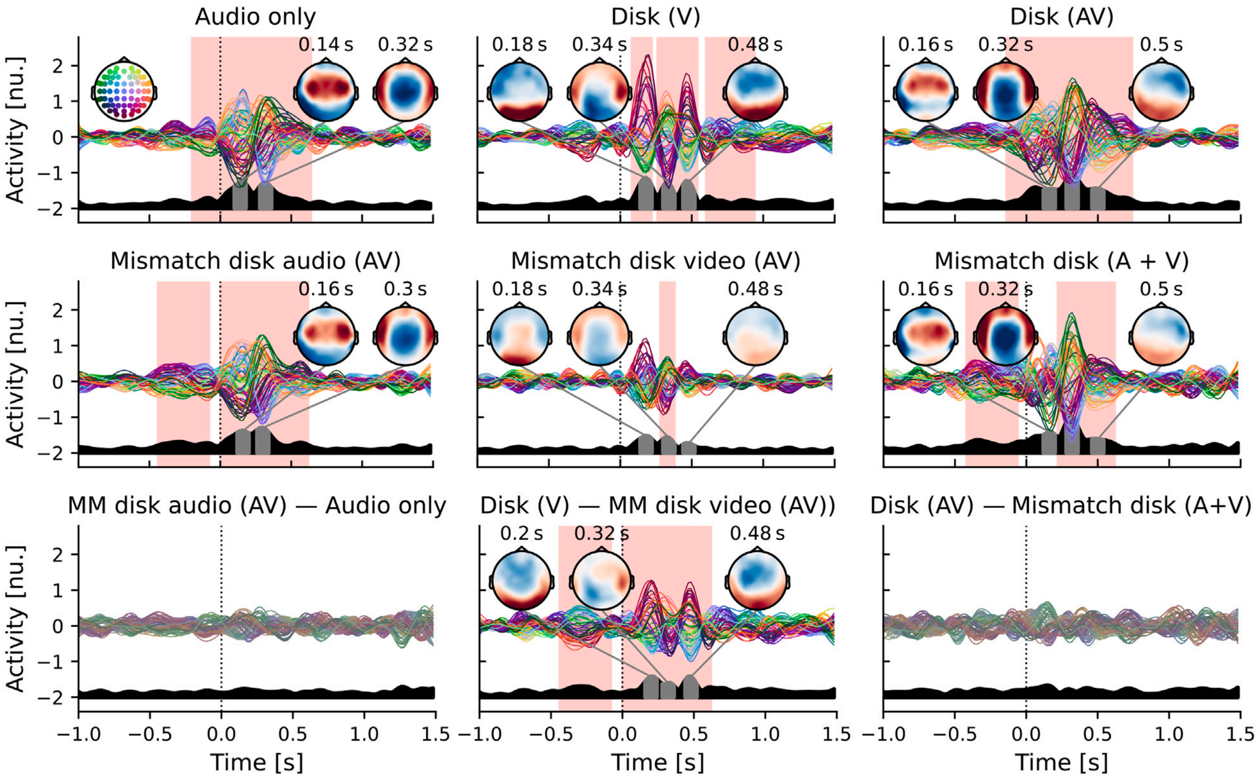

Fig. 3. Delta-band (1–4 Hz) temporal response functions (TRFs) for the “Audio-only”, “Animated Disk”, and “Mismatched Disk” conditions. Top row (left to right): TRFs for the audio-only condition; the visual-only “Animated Disk” condition; and the full audiovisual “Animated Disk” condition. Middle row (left to right): auditory TRF for the “Mismatched Disk” audiovisual condition; visual TRF for this stimulus; and the additive TRF formed by summing these partial unimodal responses. Bottom row: Pairwise difference TRFs, obtained by subtracting each middle-row TRF from its corresponding top-row TRF. Global field potentials (GFPs) are depicted in black. The TRFs were averaged over prominent peaks in the GFP profiles (indicated in grey) to obtain the inset scalp topographies. The temporal extent of statistically significant spatio-temporal clusters is indicated by the region shaded in red. Any TRFs which did not achieve statistical significance are illustrated with greyed colours. The TRFs are given in normalised units (nu.).

comparisons; only the “Cartoon” condition yielded a significant pre-Bonferroni p-value (permutation-based cluster test; $p = 0.0134$; cluster statistic = 30).

We wondered if the statistically significant $AV - A$ TRF for the “Cartoon” condition could be explained by the visual-only TRF (i.e., whether an additive model formed from TRFs obtained under unimodal stimulation could better fit the data). If not, these results would suggest an effect of audio-visual integration. We therefore performed another statistical test on the difference between the visual-only TRF, and the $AV - A$ difference TRF. We found no significant difference (permutation-based cluster test; $p = 0.190$; cluster statistic = 17).

3.4. Neural correlate of behavioural audiovisual speech enhancement

In Section 3.1, we demonstrated that behavioural speech-in-noise comprehension was systematically modulated by varying the visual stimuli from the most simple signals (the animated disk, followed by the simplified moving faces) to the natural facial recordings. In Section 3.3 we found an enhancement in the neural response to audiovisual speech in the “Natural” condition relative to the audio-only condition. We wondered whether this measure of neural audiovisual gain, which occurred at a latency of around 180 ms relative to the stimulus presentation, was linked to the behavioural performance of the participants.

To this end, we computed a measure of neural audiovisual gain from the $AV - A$ difference TRFs across the latencies from 120–240 ms. This range of latencies was based on the broad temporal extent of the significant cluster in the “Natural” difference TRF. This measure, which we refer to as the neural audiovisual gain G_{av}^{neural} and which is defined by Eq. (5), was obtained as the difference between activities at temporal and central electrode sites, since these two locations reflect opposite polarities in the scalp distribution of the auditory response. This pattern is consistent with the topographical pattern of auditory cortical activation observed in this study and previous literature, and captures the magnitude of the response gain.

We compared our neural measure G_{av}^{neural} to the behavioural results G_{av}^{beh} for each participant (Fig. 7). We observed a strong correlation between the neural and behavioural audiovisual gain (repeated-measures correlation test: Pearson’s $R = 0.59$; 56 degrees of freedom; $p = 1.08 \times 10^{-6}$). Since the “Mismatched disk” condition served as an artificial control condition, we left it out of the analysis when we computed these results. However, if these measurements are included, the result remains highly significant (repeated-measures correlation test: Pearson’s $R = 0.51$; 75 degrees of freedom; $p = 1.80 \times 10^{-6}$).

4. Discussion

4.1. Summary

In this study, we used a range of visual stimuli that varied systematically in fine structural detail to investigate how visual signal quality influences audiovisual speech-in-noise comprehension and its neural correlates. Although a simple visual analogue of the speech temporal envelope did not enhance comprehension, natural and simplified face-like visual signals did, and the size of this AV enhancement was related to the degree to which the visual signal resembled a natural talking face. By quantifying how neural envelope tracking is enhanced in the delta band during AV speech perception, we were able to develop a neural marker of audiovisual integration which was highly correlated with behavioural outcomes.

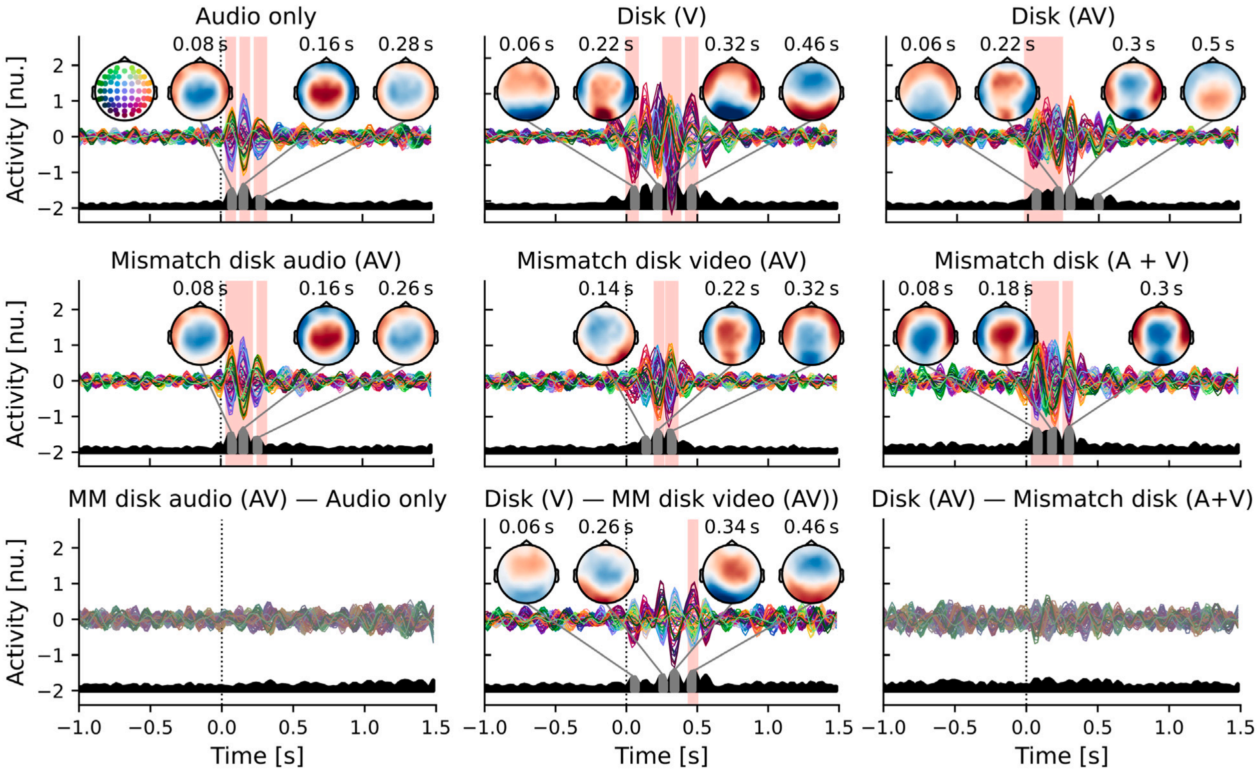

Fig. 4. Theta-band (4–8 Hz) temporal response functions (TRFs) for the "Audio-only", "Animated Disk", and "Mismatched Disk" conditions. Top row (left to right): TRFs for the audio-only condition; the visual-only "Animated Disk" condition; and the full audiovisual "Animated Disk" condition. Middle row (left to right): auditory TRF for the "Mismatched Disk" audiovisual condition; visual TRF for this stimulus; and the additive TRF formed by summing these partial unimodal responses. Bottom row: Pairwise difference TRFs, obtained by subtracting each middle-row TRF from its corresponding top-row TRF. Global field potentials (GFPs) are depicted in black. The TRFs were averaged over prominent peaks in the GFP profiles (indicated in grey) to obtain the inset scalp topographies. The temporal extent of statistically significant spatio-temporal clusters is indicated by the region shaded in red. Any TRFs which did not achieve statistical significance are illustrated with greyed colours. The TRFs are given in normalised units (nu.).

4.2. Behaviourally-measured benefits associated with simplified facial signals

Even the highly simplified "Cartoon" stimulus — constructed from facial landmarks extracted via automated tracking — yielded a modest but reliable audiovisual benefit. This signal preserved a clear lip contour, likely providing salient temporal cues and coarse place-of-articulation information, particularly for bilabial phonemes—both of which are thought to enable AV speech-in-noise benefit (Peelle and Sommers, 2015). Despite lacking texture, the Cartoon signal demonstrated robust expression, maintaining a coherent global facial configuration that included simplified facial structures (such as the eyebrows, as can be observed in Fig. 1). These configural elements may have contributed to the observed AV benefit, consistent with previous findings that global facial structure supports AV speech perception in noise (Rosenblum et al., 2000).

The "Edge" condition produced a significantly higher AV benefit than the "Cartoon" condition. This enhancement may be attributed to the availability of additional articulatory cues, such as partial visibility of tongue movements and internal mouth structures, which promote clearer distinctions in place-of-articulation. Additionally, the natural edges — which were not formed by linking facial landmarks, as was the case in the Cartoon condition — and the presence of less prominent facial contours may have improved configural processing, allowing for a more coherent percept of the moving face as a whole.

The greatest behavioural benefit was observed for the fully natural facial videos. These stimuli conveyed the richest set of visual cues: the tongue and teeth were visible with natural fidelity, and expressive facial structures which are primarily delineated through surface-level features were perceivable. Notably, the cheeks — identified in previous studies as important visual landmarks (Rosenblum et al., 1996) — are defined primarily by luminance gradients, which were absent in both the "Cartoon" and "Edge" conditions. This raises an important question: could a shaded, landmark-based cartoon that includes visible cheek structures and internal oral articulators (e.g., tongue, teeth) facilitate an AV benefit that is close to that of natural moving faces? Such questions are not only theoretically interesting but practically relevant. As augmented/virtual reality technologies and virtual avatars become more integrated into communication environments, understanding which visual features are critical for supporting speech perception could help guide the design of effective visual hearing aids or expressive virtual agents.

One limitation of using relatively naturalistic facial stimuli is that it becomes difficult to isolate which specific features are responsible for driving improvements in AV benefit. Ablation-style approaches such as that of Rosenblum et al. (1996), in which individual features are systematically removed or introduced in isolation, may offer a path forward in identifying the visual components most essential for audiovisual speech integration, and are increasingly more convenient to conduct thanks to advances in digital image processing technology.

4.3. Behavioural findings for the animated disk conditions

In contrast to the facial signals, the simplest visual signals, namely the "Animated disk" and the "Mismatched disk", did not elicit any measurable behavioural benefit. While the "Mismatched disk" served as a control condition, the lack of effect for the "Animated disk" adds to the existing literature which reports mixed results regarding speech comprehension improvements when visualising the speech envelope (Benz et al., 2025b; Summerfield, 1979; Yuan et al., 2020).

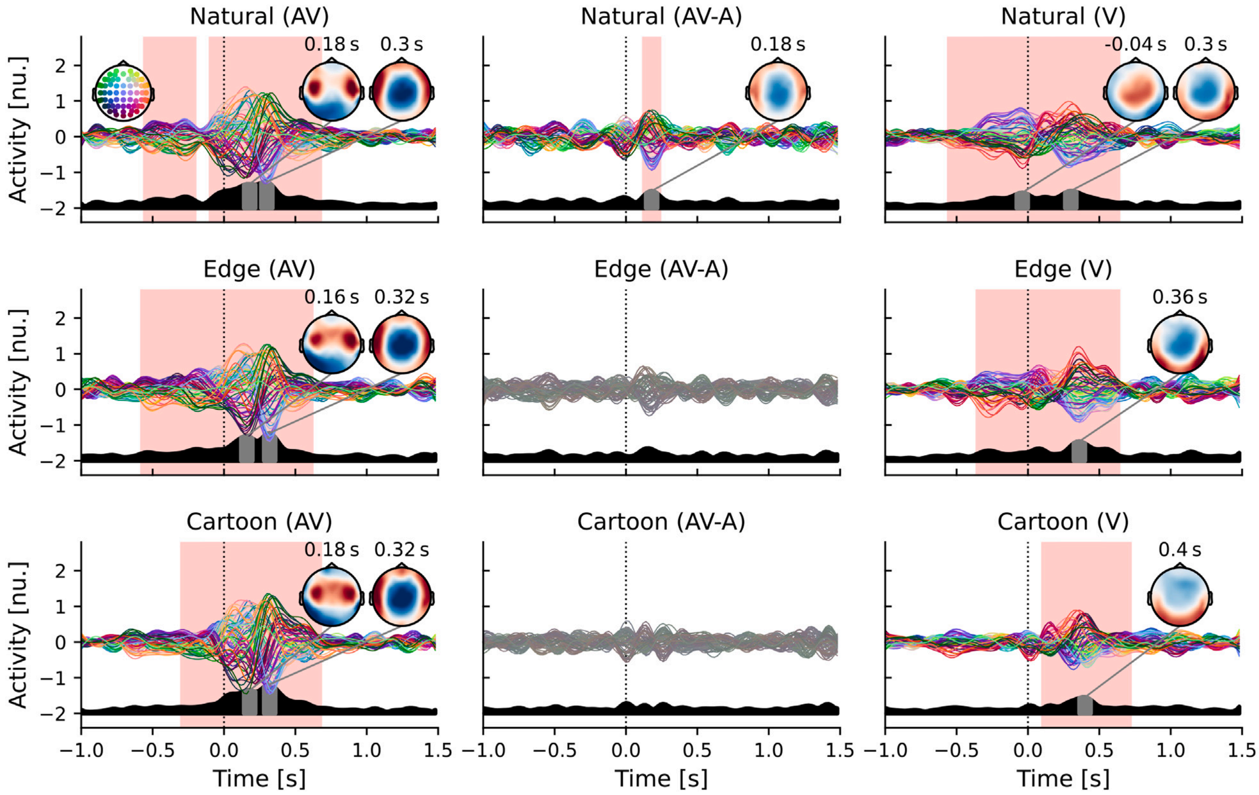

Fig. 5. Delta-band (1–4 Hz) temporal response functions (TRFs) for face-like audiovisual and visual-only conditions. The first column displays the TRFs obtained under audiovisual stimulation. The second column depicts the differences between the audiovisual TRFs and the audio-only TRFs, and the third displays the TRFs obtained under visual-only stimulation. Global field potentials (GFPs) are depicted in black. The TRFs were averaged over prominent peaks in the GFP profiles (indicated in grey) to obtain the inset scalp topographies. The temporal extent of statistically-significant clusters is indicated by the red shaded region. Any TRFs which did not achieve statistical significance are illustrated with greyed colours. The TRFs are given in normalised units (nu.).

Previous studies by Yuan and colleagues have reported modest AV benefits from abstract visualisations of the speech envelope (extracted via the Hilbert transform of the broadband speech signal), such as coloured oscillating spheres (Yuan et al., 2020, 2021b,a). Notably, these effects were small and highly sensitive to specific parameters, including the signal-to-noise ratio (SNR) of the auditory speech and the modulation depth of the visual signal. Curiously, in one study, a benefit emerged only when the radius of the disk was negatively correlated with the speech envelope (Yuan et al., 2020). On the one hand, this result appears hard to reconcile with the subsequent studies which employed a congruent oscillating sphere to elicit a behavioural benefit; on the other, the authors argued that perhaps the auditory-visual system might not prefer a particular direction of correlation, since the temporal dynamics of the speech envelope are encoded in the visual stimulus either way. Our results neither contradict nor affirm those of Yuan et al. due to the subtlety of their reported effect size, and the differences between their coloured sphere signal and our solid animated disk. Indeed, one key difference between the two approaches is that the "Animated Disk" produced strong brightness fluctuations, which — in a worst-case scenario — could even have been distracting to the participants.

Yuan et al. treated the Hilbert envelope of broadband speech as a proxy for mouth motion. While this approximation holds to some extent, it is important to point out its shortfalls. Particularly, the amplitude envelope is a feature of the acoustic signal that contains speech, and as such contains noise which is irrelevant for speech processing; particularly when it is compared to features of visual speech which are actually available in the visual modality. In fact, more accurate visual correlates of articulatory dynamics are available. For instance, the speech envelope filtered within specific spectral bands — particularly near the first or second-to-third formants — has been shown to better predict mouth opening area (Chandrasekaran et al., 2009; Grant and Seitz, 2000). Alternatively, the mouth opening area itself could be used as the modulating signal for a simplified visual stimulus. Recent MEG findings (Bourguignon et al., 2019; Benz et al., 2025a) employing visual-only speech suggest that neural activity in visual regions more closely tracks the area of mouth opening than the acoustic amplitude envelope (although the opposite result is reported for auditory cortex). Accordingly, a highly abstract stimulus such as our "Animated Disk", but based on the dynamics of the mouth-opening area instead of the envelope, could have been investigated as well.

As an alternative view to that of Yuan et al. visual motion could entrain neural oscillations which in turn modulate auditory excitability (Hickok and Poeppel, 2007; Park et al., 2016; Schroeder et al., 2008). In this case, the visual stimulus would be required to possess the correct rhythmicity so as to entrain oscillations at the correct phase to enhance auditory processing. However, if entrainment depends on the natural timing of articulatory gestures — such as mouth openings or syllable boundaries — then a stimulus driven by an abstract signal like the broadband envelope may lack the temporal specificity needed to optimise phase alignment. The absent behavioural benefit of the animated disk might then reflect its failure to provide sufficiently naturalistic or behaviourally relevant visual timing cues.

As a final remark on the design of the "Animated Disk" visual signal, we point out the considerable conceptual and perceptual gap between the isolated moving lips used in early work (Summerfield, 1979) and the abstract, envelope-driven stimuli explored in that same study and in more recent work by Yuan and colleagues and Benz et al. (2025b). Although results of our study, as well as the work of Benz et al. and Sommerfield, raise doubts over the potential for such highly abstract, envelope-driven stimuli to facilitate an AV benefit, one can imagine a broad design space of stimuli that occupy the middle ground

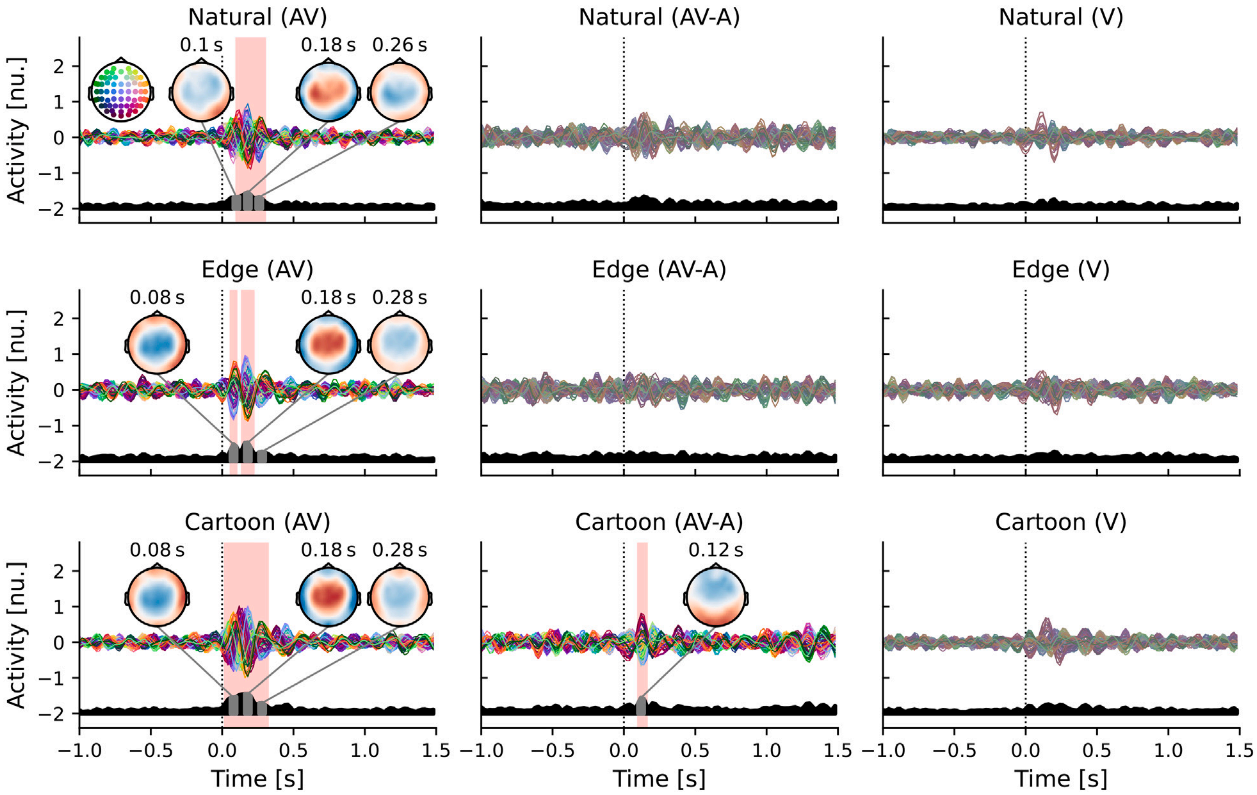

Fig. 6. Theta-band (4–8 Hz) temporal response functions (TRFs) for face-like audiovisual and visual-only conditions. The first column displays the TRFs obtained under audiovisual stimulation. The second column depicts the differences between the audiovisual TRFs and the audio-only TRFs, and the third displays the TRFs obtained under visual-only stimulation. Global field potentials (GFPs) are depicted in black. The TRFs were averaged over prominent peaks in the GFP profiles (indicated in grey) to obtain the inset scalp topographies. The temporal extent of statistically-significant clusters is indicated by the red shaded region. Any TRFs which did not achieve statistical significance are illustrated with greyed colours. The TRFs are given in normalised units (nu.).

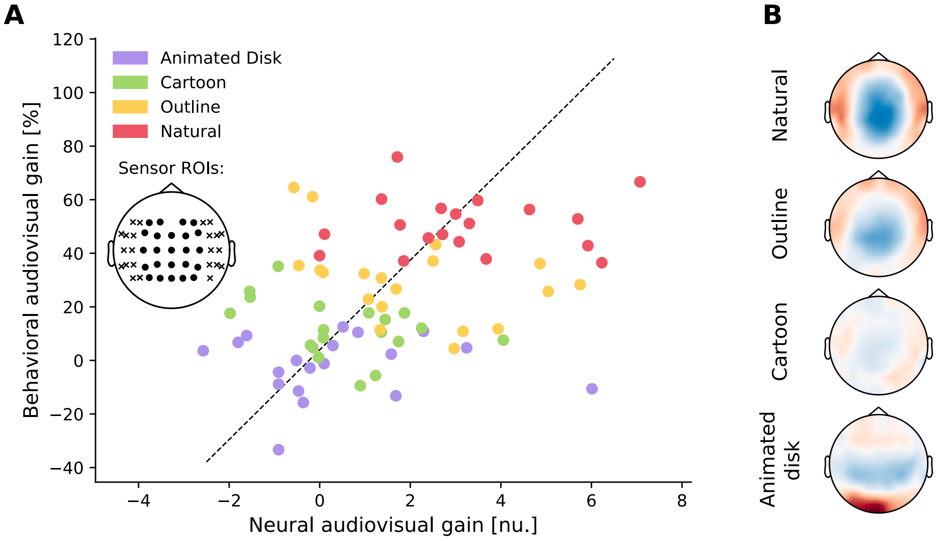

Fig. 7. We developed a measure of neural audiovisual gain using the delta-band $AV - A$ TRFs and compared it to the audiovisual gain which was measured during the behavioural experiment. **A**, Scatter plot of behavioural audiovisual gain against neural audiovisual gain. Each dot represents a pair of neural and audiovisual gains for each participant and each visual condition. The different conditions are indicated by the colours of the dots. A regression line between the two variables (neural and behavioural audiovisual gain) is represented by the black dashed line. The neural gain was computed from the difference TRFs across latencies of 120–240 ms. Specifically, we took the difference between temporal (marked with crosses in the sensor legend) and central (marked with dots) EEG activity as our neural measure, which is given in normalised units (nu.). **B**, Scalp topographies of the $AV - A$ TRFs between 120–240 ms for each visual condition.

between fully natural facial motion and simple geometric motion. Exploring this space could result in clarifications as to which visual features are necessary and sufficient for producing an AV benefit. For instance, ellipses whose major and minor axes track the vertical and horizontal lip aperture (i.e., inter-lip distance and mouth corner spread); stylised mouths constructed from simple elements (e.g. lines or curves) that dynamically reflect lip motion; or simplified cartoon faces with parametrically controlled visibility of internal articulators such as the tongue or teeth.

4.4. Neural encoding analysis: animated disks

We computed TRFs in the delta and theta bands to examine how auditory and visual envelope information was encoded in cortical responses during the EEG experiment. Specifically, we sought to quantify differences in neural envelope tracking across conditions using an approach based on Crosse et al. (2016b), in which full audiovisual TRFs are compared to a linear summation of auditory and visual responses.

For the disk conditions, we constructed such an additive model by summing the partial auditory and visual TRFs obtained from the incongruent “Mismatched Disk” condition. The resulting model was not statistically different from the full audiovisual TRF obtained from the “Animated Disk” condition. Within the framework proposed by Crosse et al. (2016b), this lack of difference suggests an absence of audiovisual integration in these conditions—at least in terms of delta- and theta-band neural tracking of the speech envelope.

The use of the partial auditory and visual TRFs from the “Mismatched Disk” condition to form our additive model is a notable departure from the approach of Crosse et al. (2016b), who instead used truly unimodal stimulation to estimate isolated auditory and visual components of the response. Whilst there was no statistical difference between the auditory-only TRF and the partial auditory TRF of the “Mismatched Disk” condition — reflecting no impact of an irrelevant visual stimulus on neural envelope tracking within these frequency bands — there was a large difference in magnitude between the visual-only “Animated Disk” TRFs and the partial visual TRFs of the “Mismatched Disk” condition. This could suggest the influence of visual attention, which is well-known to modulate visual responses such as the steady-state visually-evoked potential (SSVEP) (Joo. Kim et al., 2006; Morgan et al., 1996). As for why the face-like visual stimuli did not produce similar TRFs to the disks during visual-only stimulation, it is possible that the SSVEP-like response was a product of the strong brightness fluctuations induced by the filled white circle oscillating against a dark background; by contrast, the brightness of the facial visual signals was much more stable over time.

4.5. Neural encoding analysis: facial signals

For the face-like visual stimuli, the audio-only and audiovisual TRFs were highly similar. Meanwhile, in the delta band, the visual-only TRFs differed markedly from the audiovisual TRFs in both latency and scalp topography. These observations suggest that the additive modelling framework of Crosse et al. (2016b) was not well-suited to our data. The reason for this discrepancy remains unclear: both studies used high-quality facial recordings featuring trained speakers with clear articulation. Moreover, behavioural results in both cases confirmed that participants made use of the visual information—most notably, our “Natural” condition yielded strong audiovisual benefits in comprehension.

While Crosse et al. (2016b) reported that audiovisual enhancement compared to an additive model tracked behavioural benefit, we found something quite different: in our study, the absence of visual components in the full audiovisual TRFs coincided with the presence of behavioural AV gain, with the natural audiovisual TRF exhibiting a significant, auditory enhancement over only the audio-only TRF. This raises important questions about how integration manifests neurally—whether through super-additive enhancement over summed unisensory components, or through a different mechanism that suppresses (or at least does not require) visual envelope processing in the delta band. The approach of Crosse et al. (2016b) could be reconciled with our findings if their visual stimuli contained features that continued to elicit visual envelope responses even under audiovisual stimulation, while the neural integration itself manifested primarily as an auditory enhancement. Potentially, subtle variations in stimulus properties could give rise to such differences. For example, our facial recordings were more tightly cropped on the speaker’s face and did not include visible articulatory hand gestures, which could influence the prominence of visual envelope tracking. Testing such hypotheses might be feasible through experiments that employ a wider range of audiovisual stimuli.

Indeed, a limitation of the use of naturalistic AV stories is the reduced control over the stimuli, complicating comparisons between different studies. The use of a single talker, which is common practice in neural speech tracking studies, also leaves open the question of generalisability to other talkers.

In the delta band, the AV–A difference TRF revealed a significant enhancement at a latency of approximately 180 ms for the “Natural” condition. The “Edge” condition also showed a similar trend, although it did not reach significance. This does align with the finding of Crosse et al. (2016b) that delta-band neural envelope tracking is a key marker of audiovisual integration, and is consistent with the perspective outlined by Bourguignon et al. (2019) that early visual processing of speech facilitates an enhanced neural representation of the speech envelope in the auditory cortex. Since the delta band has been associated with word-level processing and higher-level cognitive function (Weissbart et al., 2020; Broderick et al., 2018; Etard and Reichenbach, 2019), this result provides evidence for an aspect of audiovisual integration that occurs at a higher level of auditory cognition.

Although the visual-only TRFs did not reach statistical significance in the theta band, they exhibited occipitally-dominant topographies which are suggestive of visual-driven responses. It is plausible that with a larger sample size, these visual-only TRFs would have reached significance. In such a case, an additive model might have provided a more appropriate baseline for assessing audiovisual integration effects in the theta band. Certainly in the “Cartoon” condition, the significant and visually-dominant AV-A difference TRF supports this view. But what would drive such a strong visual response in the theta band during audiovisual stimulation in the “Cartoon” condition? The limited sample size makes it difficult to answer this; one tentative suggestion could be that, due to their lack of texture, the “Edge” videos and in particular the “Cartoon” videos exhibit stronger local luminance contrasts near the bright edges presented on a dark background, which could drive a visual contrast response (i.e. an SSVEP) (Gardner et al., 2005). Our finding that this was only observable in the theta frequency band might be explained by the fact that facial dynamics such as the area of mouth opening are more highly correlated with the speech envelope in this frequency band (Chandrasekaran et al., 2009).

Beyond the speech envelope, future studies could explore alternative features that fall within the delta and theta frequency bands, the encoding of which may be more directly relevant to AV speech integration. For example, in a comparable study of audio-tactile speech integration, Guilleminot and Reichenbach (2022) demonstrated that tactile pulses aligned with the perceptual centres of syllables were shown to improve speech comprehension. Presumably, one could also use such a feature to more directly estimate the neural encoding of such events, rather than relying on an indirect proxy such as the theta-band speech envelope. Similarly, word-level features such as onsets, lexical frequency, and surprisal — already shown to modulate auditory neural tracking in continuous speech paradigms (Weissbart et al., 2020) — may also play a role in AV integration and warrant further investigation.

4.6. A neural correlate of behaviour

In the delta band, we observed a significant enhancement of the audiovisual TRF compared to the audio-only TRF for the "Natural" condition. The effect occurred at latencies of around 180 ms, and was particularly prominent in central and temporal areas. We note here that this latency is compatible with the findings of Crosse et al. (2016b), who found a slightly earlier gain (around 100 ms) in noise-free conditions, and a slightly later gain (around 230 ms) in extremely challenging levels of background noise. Additionally, their neural gain was also particularly prominant in the delta frequency band.

The neural enhancement that we observed exhibited a typical auditory-like topography, likely reflecting audiovisual integration within the auditory cortex. While the $AV - A$ topographies for the other, simplified face-like visual signals exhibited similar patterns, the effects were weaker. Future work will establish whether significant effects can be observed for these visual signals when a larger cohort of participants is recruited. Nonetheless, the stability of the topographies across the visual conditions led us to investigate whether the identified audiovisual enhancement in neural activity was related to the behavioural audiovisual gain. To this end, we developed a measure of neural audiovisual gain based on the $AV - A$ difference TRFs in the delta band, which was subsequently correlated against the behavioural audiovisual gain using repeated measures correlation. We found a highly significant correlation between the proposed neural measure and behaviour ($p = 0.0008$; $R = 0.52$).

When relating behaviour to electrophysiology, we worked across data from two separate experiments. Indeed, to obtain high-accuracy measurements of speech-in-noise comprehension, the latter was determined by presenting subjects with a range of single, semantically unpredictable sentences. However, because this audio material was too short to provide reliable quantification of neural speech tracking, the EEG recordings utilised longer narratives of audiovisual speech. We recently employed this type of dual measurement design for disentangling the contribution of the delta and theta frequency band to clarity and comprehension during speech-in-noise listening (Etard and Reichenbach, 2019).

Despite the advantages of high accuracies in the individual measurements, this approach meant that the continuous speech segments used for the EEG recordings were semantically predictable, which may have caused slightly higher levels of comprehension than the semantically unpredictable sentences employed for the behavioural assessment. However, since this effect is expected to affect all the speech narratives and all subjects equally, we do not expect this to alter the observed relation between behaviour and neural speech tracking.

4.7. Conclusion

This study shows that facial structural detail in visual speech modulates both behavioural and neural indices of audiovisual speech integration. While abstract envelope-based animations failed to enhance comprehension, even highly simplified facial signals yielded measurable benefits, suggesting that visual speech needs to retain some resemblance to natural articulation to support understanding in noise. At the neural level, audiovisual integration manifested not through superadditive sensory responses but as enhanced auditory encoding in the delta band, with visual tracking of the speech envelope suppressed in both the delta and theta bands. In the delta band, this effect was behaviourally predictive, providing a highly significant neural marker of audiovisual benefit. Overall, our results support the view that AV speech integration enhances auditory cortical encoding of word-rate dynamics, and that this enhancement scales with the visual signal's fidelity to natural facial speech.

CRediT authorship contribution statement

Enrico Varano: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mike Thornton:** Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis. **Dorothea Kolossa:** Writing – review & editing, Conceptualization. **Steffen Zeiler:** Writing – review & editing, Conceptualization. **Tobias Reichenbach:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no competing financial interests.

Acknowledgments

This research was supported by the Royal British Legion Centre for Blast Injury Studies, the EPSRC (Grant No. EP/R032602/1), the UKRI Centre for Doctoral Training in AI for Healthcare (AI4Health, Grant No. P/S023283/1), and the German Science Foundation (DFG, project number 523344822).

Data availability

The raw data and pre-processing code are publicly available and can be accessed at https://zenodo.org/record/6855794. The Python analysis code is available at https://github.com/Mike-boop/varano-av-analysis.

References

Baltrušaitis, T., Robinson, P., Morency, L.-P., 2016. Openface: An open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–10.

Benz, K.R., Hauswald, A., Suess, N., Gehmacher, Q., Demarchi, G., Schmidt, F., Herzog, G., Ŕosch, S., Weisz, N., 2025a. Eye movements in silent visual speech track unheard acoustic signals and relate to hearing experience eneuro 12(4). ENEURO 005, 5–25.

Benz, K.R., Hauswald, A., Weisz, N., 2025b. Influence of visual analogue of speech envelope, formants, and word onsets on word recognition is not pronounced. Hear. Res. 460, 109237.

Biesmans, W., Das, N., Francart, T., Bertrand, A., 2017. Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. IEEE Trans. Neural Syst. Rehabil. Eng. 25 (5), 402–412.

Bourguignon, M., Baart, M., Kapnoula, E.C., Molinaro, N., 2019. Lip-reading enables the brain to synthesize auditory features of unknown silent speech. J. Neurosci. 40 (5), 1053–1065.

Broderick, M.P., Anderson, A.J., Liberto, G.M.D., Crosse, M.J., Lalor, E.C., 2018. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr. Biology 28, 803–809.

Canny, J., 1986. A computational approach to edge detection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. (6), PAMI-8, pp. 679–698.

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

Choi, I., Lee, J.-Y., Lee, S.-H., 2018. Bottom-up and top-down modulation of multisensory integration. Curr. Opin. Neurobiol. 52, 115–122.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. The GRID audio-visual speech corpus (1.0) [data set]. Zenodo.

Crosse, M.J., D. Liberto, G.M., Bednar, A., Lalor, E.C., 2016a. The multivariate temporal response function (mTRF) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. Front. Hum. Neurosci. 10.

Crosse, M.J., Liberto, G.M.D., Lalor, E.C., 2016b. Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. J. Neurosci. 36, 9888–9895.

Crosse, M.J., Zuk, N.J., D. Liberto, G.M., Nidiffer, A.R., Molholm, S., Lalor, E.C., 2021. Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. Front. Neurosci. 15.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134, 9–21.

Ding, N., Simon, J.Z., 2014. Cortical entrainment to continuous speech: Functional roles and interpretations. Front. Hum. Neurosci. 8, 311.

Erber, N.P., 1979. Auditory-visual perception of speech with reduced optical clarity journal of speech. Lang. Hear. Res. 22 (2), 212–223.

Etard, O., Reichenbach, T., 2019. Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. J. Neurosci. 39, 5750–5759.

Gardner, J.L., Sun, P., Waggoner, R.A., Ueno, K., Tanaka, K., Cheng, K., 2005. Contrast adaptation and representation in human early visual cortex. Neuron 47 (4), 607–620.

Gehmacher, Q., Schubert, J., Schmidt, F., Hartmann, T., Reisinger, P., R´osch, K., Popov, T., Chait, M., Weisz, N., 2024. Eye movements track prioritized auditory features in selective attention to natural speech. Nat. Commun. 15 (1).

Gibson, J.J., 1950. The perception of visual surfaces. Am. J. Psychol. 63 (3), 367.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M.S., 2014. MNE software for processing MEG and EEG data. NeuroImage 86, 446.

Grant, K.W., Seitz, P.-F., 2000. The use of visible speech cues for improving auditory detection of spoken sentences. J. Acoust. Soc. Am. 108, 1197.

Guilleminot, P., Reichenbach, T., 2022. Enhancement of speech-in-noise comprehension through vibrotactile stimulation at the syllabic rate. Proc. Natl. Acad. Sci. 119 (13), e2117000119.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Rev. Neurosci. 8, 393–402.

Holdgraf, C.R., Rieger, J.W., Micheli, C., Martin, S., Knight, R.T., Theunissen, F.E., 2017. Encoding and decoding models in cognitive electrophysiology. Front. Syst. Neurosci. 11.

Jin, P., Zou, J., Zhou, T., Ding, N., 2018. Eye activity tracks task-relevant structures during speech and auditory sequence perception. Nat. Commun. 9 (1).

Joo. Kim, Y., Grabowecky, M., Paller, K.A., Muthu, K., Suzuki, S., 2006. Attention induces synchronization-based response gain in steady-state visual evoked potentials. Nature Neurosci. 10 (1), 117–125.

Jordan, T.R., McCotter, M.V., Thomas, S.M., 2000. Visual and audiovisual speech perception with color and gray-scale facial images. Percept. Psychophys. 62 (4–5), 1394–1404.

Kayser, C., Logothetis, N.K., Panzeri, S., 2010. Visual enhancement of the information representation in auditory cortex. Curr. Biology 20, 19–24.

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of eeg- and meg-data. J. Neurosci. Methods 164 (1), 177–190.

McCotter, M.V., Jordan, T.R., 2003. The role of facial colour and luminance in visual and audiovisual speech perception. Perception 32 (8), 921–936.

Meredith, M.A., Stein, B.E., 1986. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration.. J. Neurophysiol. 56, 640–662.

Morgan, S.T., Hansen, J.C., Hillyard, S.A., 1996. Selective attention to stimulus location modulates the steady-state visual evoked potential. Proc. Natl. Acad. Sci. 93 (10), 4770–4774.

O'Sullivan, A.E., Crosse, M.J., Liberto, G.M.D., d. Cheveigné, A., Lalor, E.C., 2021. Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects. J. Neurosci. 41, 4991–5003.

Park, H., Kayser, C., Thut, G., Gross, J., 2016. Lip Movements Entrain. Observers' Low-Frequency Brain Oscil. Facil. Speech Intelligibility ELife 5.

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

Pion-Tonachini, L., Kreutz-Delgado, K., Makeig, S., 2019. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. NeuroImage 198, 181–197.

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

R., Vallat, 2018. Pingouin: statistics in python. J. Open Source Softw. 3 (31), 1026.

Reisinger, P., Gillis, M., Suess, N., Vanthornhout, J., Haider, C.L., Hartmann, T., Hauswald, A., Schwarz, K., Francart, T., Weisz, N., 2025. Neural speech tracking contribution of lip movements predicts behavioral deterioration when the speakermouth is occluded. ENeuro 12 (2).

Rosenblum, L.D., Johnson, J.A., Saldana, H.M., 1996. Point-light facial displays enhance comprehension of speech in noise. J. Speech Hear. Res. 39 (6), 1159–1170.

Rosenblum, L.D., Yakel, D.A., Green, K.P., 2000. Face and mouth inversion effects on visual and audiovisual speech perception. J. Exp. Psychol. [Hum. Percept.] 26 (2), 806–819.

Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C., Foxe, J.J., 2007. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex 17, 1147–1153.

Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., Puce, A., 2008. Neuronal oscillations and visual amplification of speech. Trends Cogn. Sci. 3, 106–113.

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

Stein, B.E., Stanford, T.R., 2008. Multisensory integration: current issues from the perspective of the single neuron. Nature Rev. Neurosci. 9 (4), 255–266.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212.

Summerfield, Q., 1979. Use of visual information for phonetic perception. Phonetica 36 (4–5), 314–331.

Summerfield, Q., 1992. Lipreading and audio-visual speech perception. In: Processing the Facial Image. New York, NY, US), Clarendon Press/Oxford University Press, pp. 71–78.

Thomas, S.M., Jordan, T.R., 2004. Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. J. Exp. Psychol. 30 (5), 873–888.

Troscianko, T., Benton, C.P., Lovell, P.G., Tolhurst, D.J., Pizlo, Z., 2008. Camouflage and visual perception. Phil. Trans. R. Soc. B 364 (1516), 449–461.

Varano, E., Guilleminot, P., Reichenbach, T., 2023. AVbook, a high-frame-rate corpus of narrative audiovisual speech for investigating multimodal speech perception. J. Acoust. Soc. Am. 153 (5), 3130.

Varano, E., Vougioukas, K., Ma, P., Petridis, S., Pantic, M., Reichenbach, T., 2022. Speech-driven facial animations improve speech-in-noise comprehension of humans. Front. Neurosci. 15, 1747.

Weissbart, H., Kandylaki, K.D., Reichenbach, T., 2020. Cortical tracking of surprisal during continuous speech comprehension. J. Cogn. Neurosci. 32, 155–166.

Wikman, P., Salmela, V., Sj´oblom, M., Laine, M., Alho, K., 2024. Attention to audiovisual speech shapes neural processing through feedback-feedforward loops between different nodes of the speech network. PLOS Biology 22 (3), e3002534.

Yuan, Y., Lleo, Y., Daniel, R., White, A., Oh, Y., 2021a. The impact of temporally coherent visual cues on speech perception in complex auditory environments. Front. Neurosci. 15.

Yuan, Y., Meyers, K., Borges, K., Lleo, Y., Fiorentino, K.A., Oh, Y., 2021b. Effects of visual speech envelope on audiovisual speech perception in multitalker listening environments. J. Speech Lang. Hear. Res. 64 (7), 2845–2853.

Yuan, Y., Wayland, R., Oh, Y., 2020. Visual analog of the acoustic amplitude envelope benefits speech perception in noise. J. Acoust. Soc. Am. 147, EL246.