frontiers | Frontiers in Human Neuroscience

TYPE Original Research
PUBLISHED 09 April 2025
DOI 10.3389/fnhum.2025.1560558

OPEN ACCESS

EDITED BY
Sascha Frühholz,
University of Zurich, Switzerland

REVIEWED BY
Lars Hausfeld,
Maastricht University, Netherlands
Sebastian Ocklenburg,
Medical School Hamburg, Germany

*CORRESPONDENCE
Mareike Daeglau
✉ mareike.daeglau@uni-oldenburg.de

RECEIVED 14 January 2025
ACCEPTED 27 March 2025
PUBLISHED 09 April 2025

CITATION
Daeglau M, Otten J, Grimm G, Mirkovic B, Hohmann V and Debener S (2025) Neural speech tracking in a virtual acoustic environment: audio-visual benefit for unscripted continuous speech. *Front. Hum. Neurosci.* 19:1560558. doi: 10.3389/fnhum.2025.1560558

Neural speech tracking in a virtual acoustic environment: audio-visual benefit for unscripted continuous speech

Mareike Daeglau[1]*, Jürgen Otten[2], Giso Grimm[2], Bojana Mirkovic[1], Volker Hohmann[2] and Stefan Debener[1]

[1]Neuropsychology Lab, Department of Psychology, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany, [2]Department of Medical Physics and Acoustics, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

The audio-visual benefit in speech perception—where congruent visual input enhances auditory processing—is well-documented across age groups, particularly in challenging listening conditions and among individuals with varying hearing abilities. However, most studies rely on highly controlled laboratory environments with scripted stimuli. Here, we examine the audio-visual benefit using unscripted, natural speech from untrained speakers within a virtual acoustic environment. Using electroencephalography (EEG) and cortical speech tracking, we assessed neural responses across audio-visual, audio-only, visual-only, and masked-lip conditions to isolate the role of lip movements. Additionally, we analysed individual differences in acoustic and visual features of the speakers, including pitch, jitter, and lip-openness, to explore their influence on the audio-visual speech tracking benefit. Results showed a significant audio-visual enhancement in speech tracking with background noise, with the masked-lip condition performing similarly to the audio-only condition, emphasizing the importance of lip movements in adverse listening situations. Our findings reveal the feasibility of cortical speech tracking with naturalistic stimuli and underscore the impact of individual speaker characteristics on audio-visual integration in real-world listening contexts.

KEYWORDS

neural speech tracking, EEG, virtual acoustic environment, continuous speech, unscripted conversation

1 Introduction

Auditory attention decoding (AAD) has traditionally aimed to distinguish between target and non-target speakers in environments with competing voices, capturing selective attention mechanisms in complex auditory scenes. Significant strides have been made in this field by decoding the speaker to whom a listener is attending, based on the brain's response to multiple simultaneous speakers (Ding and Simon, 2014; Luo and Poeppel, 2007; Mirkovic et al., 2015). AAD studies typically rely on controlled, multi-speaker environments, often using professional speakers and scripted speech for consistency and precision (Holtze et al., 2023; Jaeger et al., 2020; Mirkovic et al., 2016). Although these studies have been foundational, their reliance on controlled settings presents challenges for generalizing findings to more naturalistic auditory environments.

The AAD approach has been extended to explore the impact of visual cues—such as lip movements and facial expressions—on selective attention, especially in noisy settings (Chandrasekaran et al., 2009; Fu et al., 2019). Visual input can enhance speech

comprehension by providing congruent cues that aid auditory processing, particularly when auditory signals are degraded. Conversely, incongruent visual cues can create perceptual illusions, demonstrated by the McGurk effect, where mismatched audio and visual inputs can lead to the perception of a novel sound (Jiang and Bernstein, 2011; Mcgurk and Macdonald, 1976; Stropahl and Debener, 2017). This phenomenon underscores the intricate interplay between auditory and visual processing. However, audio-visual fusion seems to vary strongly between different speakers, different audio-visual stimulus combinations, and between participants (Mallick et al., 2015; Stropahl et al., 2017; Stropahl and Debener, 2017).

Most AAD studies are based on neural tracking procedures, which can be used to study how well brain activity captures continuous speech stream fluctuations (Crosse et al., 2015; Luo and Poeppel, 2007; Puschmann et al., 2019). Neural tracking is especially valuable for studying the neural dynamics of speech processing in naturalistic environments, where congruent multi-modal cues, such as lip movements, enhance speech comprehension without the complexity of competing voices. However, the ecological validity, i.e., the extent to which findings generalize to real-world communication, depends on how closely experimental conditions reflect real-world listening situations (Keidser et al., 2020). Traditional AAD and neural speech tracking studies often rely on highly controlled stimuli, limiting their ecological validity and leaving open the question of how neural tracking operates in more variable, real-life auditory settings.

Traditional context factors such as background noise, speaker position or varying hearing abilities have been thoroughly investigated in AAD and neural speech tracking studies (Geirnaert et al., 2021; Rosenkranz et al., 2021; Wang et al., 2023; Zion Golumbic et al., 2012). However, other factors, such as the likeability of the speaker or specific speech features of the speaking person, may contribute to how well a speech signal is followed by a listener (Wiedenmann et al., 2023). Research findings in the context of advertising or expert testimony suggest that likeability drives attention, meaning that more likeable people capture greater attention independent of their actions (Fam and Waller, 2006; Younan and Martire, 2021). Likeability, as a socio-emotional factor, may influence listener engagement and attention, potentially modulating speech tracking (Farley, 2008; Li et al., 2023).

Similarly, characteristics like articulation clarity, pitch range, or speech rhythm could contribute to individual differences in neural tracking efficacy. These rather unexplored context factors are particularly relevant for understanding real-world communication, where socio-emotional dynamics and individual speaker traits naturally interact with auditory processing (Bachmann et al., 2021; Etard and Reichenbach, 2019; Peelle and Davis, 2012). Ensuring ecological validity in speech tracking research requires considering such factors, as real-world listening situations rarely involve highly controlled, professional speech but rather a diverse range of speakers with varying vocal characteristics. While ecological validity is often cited as a justification for using more naturalistic stimuli, its definition and application in psychological research remain debated (Holleman et al., 2020). Beyond simply increasing external realism, ecological validity requires systematically considering the cognitive and environmental constraints that shape processing in real-world settings. In this study, we enhance ecological validity not only by incorporating spontaneous speech but also by systematically examining speaker-specific characteristics that influence neural tracking.

To date, it remains poorly understood which speaker characteristics contribute to effective neural tracking. Most studies in the field utilize professional speakers with precise articulation, creating a controlled foundation for understanding neural tracking mechanisms (Crosse et al., 2015; Jaeger et al., 2020). Real-world listening, however, typically involves understanding untrained speakers, whose articulation, pitch, and spontaneity can vary widely, thus limiting the ecological validity of such studies. The influence of individual differences in vocal characteristics on neural tracking efficacy may be of particular relevance when audio-visual cues come into play (Vanthornhout et al., 2018). For example, the emotional expressiveness, facial dynamics, and speech fluency of individual speakers may interact with neural tracking and comprehension in ways that are not yet fully understood (Scherer et al., 2019; Tomar et al., 2024).

Previous studies have integrated visual cues into neural speech tracking to determine how congruent visual information, like lip movements, enhances comprehension in dynamic, noisy contexts (Crosse et al., 2016b; Park et al., 2016). These results underscore the powerful role of visual–auditory integration in enhancing speech comprehension under challenging listening conditions. However, the interplay between speech content, speaker characteristics, and listener preferences or biases warrants further exploration.

In this study, we examined how speaker-specific characteristics, such as articulation, pitch, and visual expressiveness, influence neural speech tracking in single-speaker, naturalistic audio-visual scenarios. By incorporating diverse speaker profiles and realistic listening contexts, we aim to improve ecological validity and shed light on the interplay of individual speaker traits and contextual factors in shaping speech processing. We hypothesized that audio-visual (AV) conditions would yield a benefit in neural speech tracking, reflected by larger envelope tracking in AV compared to audio-only (A) stimuli across individual speakers. Additionally, we explored whether individual differences between speakers, characterized by various speech features, would influence the magnitude of A and AV speech tracking and the AV benefit. By linking these speaker-specific traits to neural responses, this study aims to address the gap in understanding how individual speaker characteristics modulate speech processing in naturalistic, single-speaker scenarios.

2 Methods

2.1 Participants

The sample size for this study was determined using a formal GPower analysis (Faul et al., 2007) for a repeated-measures ANOVA. Assuming a small effect size ($f = 0.2$; Cohen, 2013), an alpha level of 0.05, a power of 0.8, a correlation of 0.5 among repeated measures, and a nonsphericity correction of 0.8, the analysis indicated a required total sample size of 18. Thus, twenty normal hearing participants were recruited for the study. Data from two participants were incomplete owing to technical difficulties and were therefore excluded from further processing. Participants' ages ranged from 22 to 35 years (M:26 years; 13 f, 5 m). The inclusion criteria were self-reported normal hearing, normal or corrected-to-normal vision, no previous or current neurological or psychological disorders, and

native German skills. Participants completed questionnaires covering demographic information and general health assessments and gave written informed consent. The study protocol was approved by the Commission for Research Impact Assessment and Ethics of the University of Oldenburg.

2.2 Apparatus

Participants were seated in the centre of a cylindrical projection screen, which had a radius of 1.74 m and a height of 2 m (Hohmann et al., 2020). A circular array of 16 active loudspeakers (Genelec 8020C) was positioned behind a screen. Behind the loudspeakers, which were positioned at ear level, was a heavy black curtain to reduce reflections and ambient light, and to provide acoustic treatment at mid and high frequencies. The video image was projected with a single ultra-short throw projector (NEC U321H) at a resolution of 1,920 × 1,080 pixels at 60 fps. The screen warping was processed in the graphics card (Nvidia Quadro M5000), and the field of view was 120 degrees. Due to the screen warping, the effective pixel density varied across the projection and was lowest in the centre, so the projected video was shifted to one side to achieve the highest possible pixel density. The Toolbox for Acoustic Scene Creation and Rendering (TASCAR) (Grimm et al., 2019) was used for audio playback, control of the virtual acoustic environment in the lab, data logging of all sensors, and experimental control. The videos were embedded in a simple 2D virtual visual environment rendered using the Blender game engine (version 2.79c). The content of the game engine (selection of videos, timing of video playback, position of virtual objects) was controlled by the acoustic engine TASCAR.

2.3 Stimuli

For this study, 18 videos, each comprising one of six different speakers (2 m; 1d; 3f) were taken from a set of pre-recorded audio-visual stimuli (Wiedenmann et al., 2023)[1] . Speakers sat in front of a dark grey background, showing their head and upper body up to their shoulders centered in frame. Speakers talked continuously at their natural pace in standard German about self-selected content right into the camera, but with natural movements and glances wandering occasionally. Each of the videos contained an enclosed story, e.g., about travel reports or daily life anecdotes about their student and work life. The duration of the videos varied between 180 s and 600 s, cut down from longer recording sessions. The videos were recorded using a Canon EOS 700D with a resolution of 1,920 × 1,080 pixels at 25 fps. The corresponding audio was recorded with a cardioid microphone (Neuman KM184) at approximately 0.7 m, using an RME Micstasy preamplifier and AD-converter with 48 kHz sampling rate. Speakers wore open earphones,[2] in which half of the recordings played babble background noise at a sound pressure level (SPL) of 65 dB unweighted. Video editing and audio and video synchronization were performed using DavinciResolve (Version17). Videos and audio were processed using FFMPEG.[3]

Each video was cut into consecutive 30-s segments, and the following conditions were prepared: audio-visual (AV), audio-only (A), visual-only (V) and masked-lips (ML). In the AV conditions the speaker was presented with the corresponding audio; in the audio-only conditions, the audio was presented alongside a video of a grey-background; in the V-conditions, only the video of the real speaker was shown while the audio was muted; for the ML conditions, the lips of the speaker were overlaid with a light blue horizontal bar, while the corresponding audio was played unaltered. The order of these conditions was pseudo-randomized within the 18 videos but kept constant across participants, changing every 30 s (Figure 1). While all participants experienced the same condition sequences, order of presentation of the 18 videos was randomized across participants. The design prevented immediate condition repetitions and minimized potential adaptation effects or prediction biases. Half of the conditions were block wise presented with background noise (65 dB SPL), while the other half was presented in quiet conditions. Participants were instructed to pay attention to the speakers and the content of the stories. After each experimental session, the participants were asked to rate the likeability of the speakers and how well they were able to follow their stories on a five-point Likert scale. Additionally, questions about the story content were asked in a multiple-choice format. Due to the total duration of 85 min of story content, the experiment was split into two sessions to avoid exhausting participants and potentially distorting neural speech tracking. The period between both sessions varied between one and 14 days.

2.4 EEG data acquisition and preprocessing

EEG data were acquired using a wireless, head-mounted 24-channel EEG system (SMARTING, mBrainTrain, Belgrade, Serbia). The system features a sampling rate of 500 Hz, a resolution of 24 bits, and a bandwidth from DC to 250 Hz. EEG data were collected from 24 scalp sites using sintered Ag/AgCl electrodes with FCz as the ground and AFz as the reference (Easycap, Herrsching, Germany). The electrode sites were prepared using 70% alcohol and an abrasive electrolyte gel (Abralyt HiCI, Easycap GmbH, Germany). The electrode impedances were maintained below 10 kΩ and tested before data acquisition. The EEG signal as well as other data (i.e., eye tracking, head movements; not investigated here) were wirelessly transmitted to a PC via Bluetooth and synchronized and recorded using the lab streaming layer protocol (Kothe et al., 2024) and saved into an .xdf file. Additional data recording was performed using TASCAR to .mat files (eye and head tracking, not analysed here).

For offline analysis, EEGLAB (Delorme and Makeig, 2004) and MATLAB (R2024a, MathWorks Inc., Marick, MA, United States) were used. Identification of improbable channels was conducted using the EEGLAB extension trimOutlier with an upper and lower boundary of two standard deviations of the mean standard deviation across all channels. Channels that exceeded this threshold were excluded. A copy of the EEG data was first low-pass filtered at 40 Hz (finite impulse

1 https://zenodo.org/records/8082844

2 https://batandcat.com/portable-hearing-laboratory-phl.html

3 http://www.ffmpeg.org

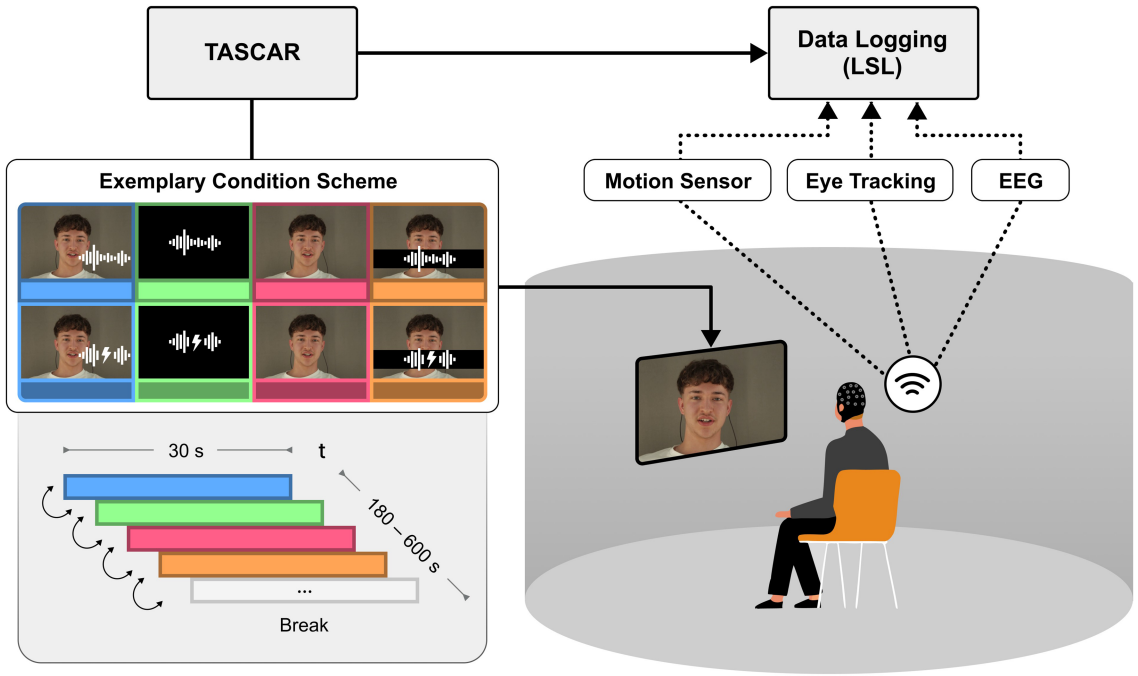

FIGURE 1
Experimental set-up including data logging. Using TASCAR VE, 18 participants were presented audio, visual or congruent combined audio-visual scenes comprising one out of six speakers at a time telling unscripted stories. Listening difficulty was manipulated by comparing speech with and without babble noise. Data logging was achieved wirelessly via Bluetooth and Wi-Fi. Written informed consent was obtained from all individuals for the publication of any potentially identifiable images or data included in this article.

response (FIR), Hamming window, filter order 166), downsampled to 250 Hz, and subsequently high-pass filtered at 1 Hz (FIR, Hamming window, filter order 414; filters integrated into EEGLAB, version 1.6.2). Afterwards, data were segmented into consecutive 1-s epochs and segments containing artifacts were removed (EEGLAB functions pop eegthresh.m, +-80uV; pop rejkurt.m, SD = 3). The remaining data were submitted to extended Infomax ICA. The unmixing matrix obtained from this procedure was applied to the original unfiltered EEG dataset to select and reject components representing stereotypical artifacts. Components reflecting eye, muscle, and heart activity were identified using ICLabel (Pion-Tonachini et al., 2019). Components flagged and identified as artifacts were removed from further analysis. Artifact-corrected EEG data were low-pass filtered with a FIR filter and a cut-off frequency of 30 Hz (hann window, filter order 220, Fs = 500 Hz), and subsequently high-pass filtered with a FIR filter and a cut-off frequency of 0.3 Hz (hann window, filter order 500, Fs = 500 Hz). After the data were re-referenced to the common average and corrupted channels were replaced by spherical interpolation, the data were resampled to 64 Hz (to reduce the computational demand for the envelope reconstruction) and cut into 30-s epochs (matching the presentation of conditions in the experiment). Pre-processed EEG data were further processed using the mTRF toolbox (Crosse et al., 2016a).

2.5 Audio pre-processing and speech envelope reconstruction

A broadband audio envelope was extracted as follows: Each audio track was z-normalized and bandpass filtered into 128 logarithmically-spaced frequency bands between 100 and 6,500 Hz, using a gamma tone filter bank (Herzke and Hohmann, 2007; Hohmann, 2002). The 100–6,500 Hz range was chosen based on previous research suggesting a high temporal coherence between visual features and speech envelope within this frequency range (Chandrasekaran et al., 2009; Crosse et al., 2015). Hilbert transformation was used to compute the signal envelope within each of 128 frequency bands. The broadband envelope was then obtained by averaging the absolute Hilbert values across all bands. The broadband envelope was low-pass filtered at 30 Hz using a 3rd-order Butterworth filter and subsequently down-sampled to 64 Hz for further processing. The mTRF toolbox (Crosse et al., 2016a), was used to reconstruct the broadband envelope utilizing the presented speech signals and the EEG data. This approach is based on multivariate linear regression to obtain a linear mapping between the EEG sensor data and the broadband speech envelope. The determination of the ridge parameter λ was achieved through an optimization process involving a search grid and a leave-one-out cross-validation procedure to minimize the mean-squared error associated with the regression. The range of values within the search grid encompassed magnitudes such as 10^{-2}, 10^{-1},…, 10^{4}, 5×10^{4}, 10^{5}, …, 10^{9}. To ensure the generalizability of the relationship between speech input and neural response, we employed a leave-one-trial-out cross-validation strategy on subject level. For each trial, the speech envelope was reconstructed using the mean regression weights derived from all other trials for one subject within the same experimental condition and at the same temporal lag, excluding only the trial being reconstructed. This approach ensured that the reconstruction was based solely on independent data, preventing circularity and overfitting. The reliability of the reconstruction was quantified by computing Pearson's correlation coefficient between the reconstructed

and original speech envelopes. For statistical treatment, the correlation coefficients were subjected to Fisher's z-transformation to achieve normality and were subsequently averaged across trials. For an initial exploration of the temporal dynamics of speech envelope tracking, individual lag models, characterized by 24 regressors corresponding to each EEG channel, were computed for every trial across 33 discrete time lags spanning from stimulus presentation to EEG signal acquisition, covering a temporal range of 0 to 500 ms. This analysis yielded a time course, from which the time lag range of interest was discerned (200–325 ms). For further analyses of audio-visual enhancements, multi-lag models containing 24 x N(lags) regressors were computed for each of these time lag ranges and all trials (Puschmann et al., 2017, 2019). For further statistical evaluation, r values were normalized using MATLAB's atanh-function (r_z).

2.6 Exploratory analyses

2.6.1 Extraction of acoustic features

To analyze oscillatory components in the audio data, the frequency spectrum was divided into four bands: envelope-range (0.3–30 Hz), low-range (30–300 Hz), mid-range (300–1,000 Hz), and high-range (1,000–4,500 Hz). This division allowed for a detailed examination of low-frequency elements associated with prosody and high-frequency components characteristic of speech. MATLAB and the FieldTrip toolbox were employed to implement the multitaper method (mtmfft) for frequency-domain analysis, well-suited for the relatively short (30-s) audio segments in this study. Each audio segment was transformed into a power spectrum under specific configurations. Frequency smoothing was set at 0.5 Hz (cfg_tapsmofrq = 0.5), balancing resolution and noise reduction across frequency bands. The analysis was limited to a frequency range of 0.3 to 4,500 Hz (cfg.foilim) to exclude non-speech-relevant frequencies. To isolate oscillatory components in the data a division approach was employed. The original power spectrum was normalized by dividing it by the fractal component, reducing the influence of non-oscillatory noise (cfg.operation = 'x^2/x^1'). For each speaker, periodic power within each frequency band was summed and normalized by the segment duration, resulting in an average periodic power per band, which was stored for further analysis, respectively, FreqRsum<30 (envelope-range), FreqRsum<300 (low-range), FreqRsum<1 k (mid-range), and FreqRsum<4.5 k (high-range), indicating the amount of periodic proportions for each speaker. Additionally, a set of 16 acoustic features was extracted from each 30-s audio segment to capture essential elements of vocal dynamics and quality with Praat (Boersma and Weenink, 2024) using the in-build voice report metrics. These features included *Pitch Metrics* (meanPitch, medianPitch, sdPitch, minPitch, maxPitch), *Jitter Metrics* (jitter_loc, jitter_loc_abs, jitter_rap, jitter_ppq5), *Shimmer Metrics* (shimmer_loc, shimmer_loc_dB, shimmer_apq3, shimmer_apq5, shimmer_apq11), *Noise-to-Harmonic Ratio (NHR)* (mean_nhr) and *Intensity* (min_intensity). For each speaker, each feature was averaged across segments to reduce inter-segment variability, providing a robust profile for inter-speaker comparison.

2.6.2 Extraction of visual features

To consider the multimodal nature of our stimuli, two visual features were extracted from each video segment using a custom Python-based image processing script. The script specifically targeted *Lip Openness* (representing articulatory movements associated with speech) and *Lip Brightness* (capturing the visual clarity and lighting conditions of each video segment). Using OpenCV, the Python script processed video data to compute average values for each visual feature over the segment duration.

2.6.3 Feature processing

After feature extraction, including frequency-based, acoustic, and visual data, features were normalized from 0 to 1 using MATLAB's normalize function, facilitating comparability across features with different scales. To refine the feature set and to avoid multicollinearity, features that were correlated above an r-value of 0.8 were removed, reducing the number of features from 22 to 10. Each of the remaining feature's correlations with the average condition values obtained prior (see section 2.5) was assessed.

3 Results

3.1 Neural speech tracking across conditions

The AV condition exhibits the highest correlation in the presence of noise, peaking around 250 ms, whereas the ML and A conditions yield lower, thus comparable correlations. The V condition correlates lowest, but in most time lags above the chance level. The lower panels display topographic maps showing the decoder weight distribution of neural responses across the scalp for each condition in both no-noise and noise contexts. Each map represents the condition specific decoder weights, with color gradients indicating the strength and direction of the weighting. In the noise condition, AV and A show distinct patterns in frontal and temporal regions, suggesting enhanced neural tracking when both audio and visual cues are present. In the no-noise condition, the spatial response patterns are more evenly distributed, with AV and A conditions still demonstrating more pronounced activations than V or ML.

Figure 2 depicts the speech envelope reconstruction accuracy r_z for each listening condition as a function of the relative time lag between auditory input and EEG response. The time lag range of interest (i.e., 200–325 ms; indicated in grey) was defined based on the group-level peaks of envelope reconstruction accuracy in all conditions. To establish a data-driven chance level, we performed a permutation test across all conditions by randomly shuffling the trial labels 1,000 times. This approach provides a stable estimate of the null distribution while ensuring comparability across conditions. Since the classification framework remains identical for all conditions, we opted for a single permutation-derived chance level rather than condition-specific estimates.

To investigate expected differences in r_z between conditions, we performed a 2×4 repeated measures ANOVA with r_z as the dependent variable and two within-subject factors: background noise (two levels: noise, no-noise) and audio-visual effect (four levels: congruent audio-visual, visual-only, audio-only and masked-lips). To ensure that our data met the assumptions for parametric statistical tests, we conducted normality tests for each

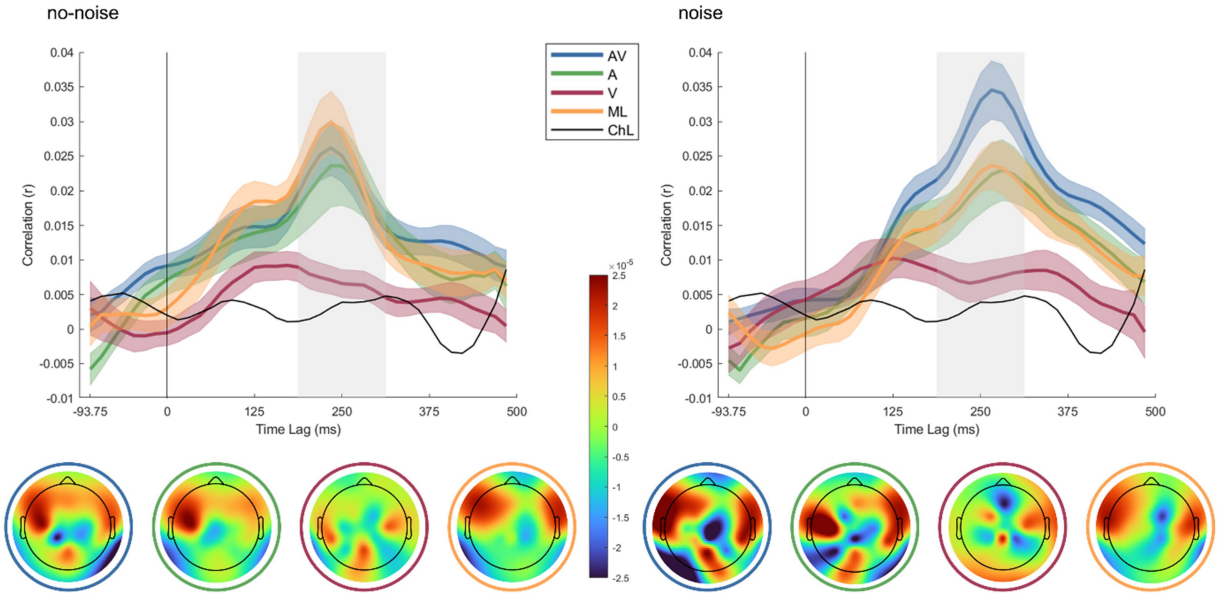

FIGURE 2
Temporal dynamics of cortical speech tracking across different audio-visual conditions in noise and no-noise environments and decoder weight topographies. In the upper panels, line plots display the time-lagged correlation (R) between the neural response and the speech envelope for four conditions: audio-visual (AV) in blue, audio-only (A) in green, visual-only (V) in red, and masked-lips in orange, in both noise (right) and no-noise (left) conditions. The shaded areas around each line indicate the standard error of the mean (SEM) across participants. The chance level (ChL), computed via random permutation between EEG and audio data, is depicted in black. The x-axis represents the time lag (in ms) relative to the trial onset, with positive values indicating the delayed neural response, while the y-axis shows the correlation coefficient (r). The lower panels display topographic maps of decoder weights averaged across participants, corresponding to the highlighted grey time lag range in the upper plots.

condition using the Kolmogorov–Smirnov. The results revealed that the assumption of normal distribution was not violated ($p > 0.05$ for all conditions). Greenhouse–Geisser correction was applied when Mauchly's test indicated a violation of sphericity. ANOVA results indicated a significant main effect for audio-visual effect ($F_{1.72,29.3} = 16.95$, $p = < 0.001$, $\eta^2 = 0.36$), and a significant interaction effect for background noise x audio-visual effect ($F_{3,51} = 3.2$, $p = 0.03$, $\eta^2 = 0.03$) but no significant main effect for background noise ($F_{1,17} = 0.06$, $p = 0.8$, $\eta^2 = 0.003$).

Planned *post hoc* paired t-tests revealed significant differences between AV and A in noise ($t_{(17)} = 3.91$, $p = 0.001$, $d = 0.87$) and AV and ML in noise ($t_{(17)} = 3.71$, $p = 0.002$, $d = 0.92$), with AV being more pronounced then A or ML, but not between A and ML in noise ($t_{(17)} = -0.37$, $p = 0.72$, $d = 0.09$). Further, for the no-noise conditions, no significant differences between AV and A ($t_{(17)} = 0.38$, $p = 0.071$, $d = 0.09$), AV and ML ($t_{(17)} = -1.01$, $p = 0.33$, $d = -0.24$), and A and ML ($t_{(17)} = -1.11$, $p = 0.28$, $d = -0.26$). Additionally, AV in noise was significantly higher than AV no-noise ($t_{(17)} = 2.99$, $p = 0.008$, $d = 0.7$). *p*-values were corrected for multiple comparisons over seven tests using Holm-Bonferroni (Holm, 1979). Figure 3 displays boxplots of the obtained contrasts.

To confirm that possible differences in pause durations between speakers and conditions did not systematically influence our results, we conducted a Bayesian repeated-measures ANOVA (default prior) with background (noise/no-noise) and condition (AV, A, V, ML) as factors. The analysis revealed no substantial interaction effect ($BF_{10} = 0.105$, error = 2.89%), indicating that pause durations were comparable across conditions and unlikely to confound the decoding results.

3.2 Exploratory results

3.2.1 Cortical speech tracking for conditions AV, a and ML for each speaker

Unlike in the previous section, conditions were not separated for noise and no-noise conditions due to the limited number of trials for each individual speaker. Therefore, all analyses in this section were conducted across both noise and no-noise conditions combined. Pictures of all speakers as well as their individual neural time courses averaged over conditions AV, A, and ML are depicted in Figure 4 along with the respective time course averaged across speakers. Time courses show similar patterns across speakers over later time lags (234–296 ms) but are more diverging in earlier time lags (140–187 ms).

3.2.2 Individual speaker's auditory and visual features

Highly intercorrelated features (correlated above an r-value of 0.8; see section 2.6.3) were removed from the initial set of features (medianPitch, sdPitch, jitter_loc, jitter_rap, jitter_ppq5, all shimmer metrics, NHR and min_Intensity), resulting in 10 features for further investigation (cf. Figure 5).

Summed power within predefined frequency bands (freqRsum<30, freqRsum<300, freqRsum<1 k, freqRsum<4.5 k) demonstrated variability across speakers. Notably, speaker2 showed the highest summed power in the lower frequencies, speaker4 in the low-range, speaker5 in the mid-range and speaker6 in the high-range, respectively. MaxPitch is highest for speaker1, whereas speaker3 and speaker5 share lowest minPitch. MeanPitch is comparably high for speakers 2, 4, and 6. JitterLocAbs (reflecing the absolute average cycle-to-cycle variation in

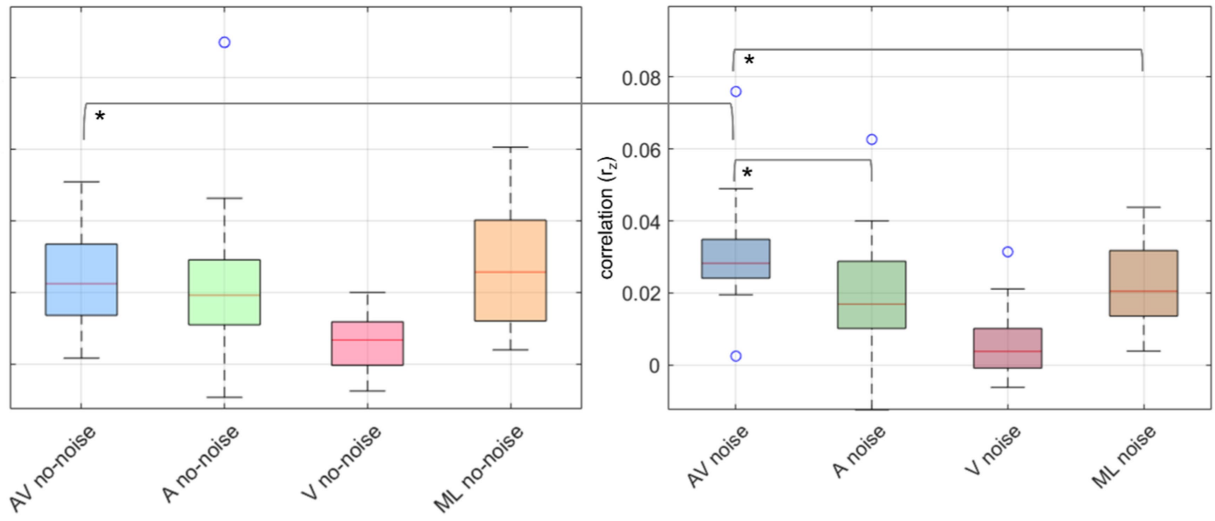

FIGURE 3
Boxplots of neural speech tracking correlations (r_z) across different conditions in noise and no-noise environments. The left panel represents no-noise conditions, while the right panel represents noise conditions. Each boxplot displays the distribution of correlation values across participants for four conditions: audio-visual (AV) in blue, audio-only (A) in green, visual-only (V) in red, and masked-lips (ML) in orange. The central line in each box represents the median, the box edges represent the interquartile range (IQR), and the whiskers extend to 1.5 times the IQR. Outliers are shown as individual circles. Asterisks (*) indicate statistically significant differences between conditions.

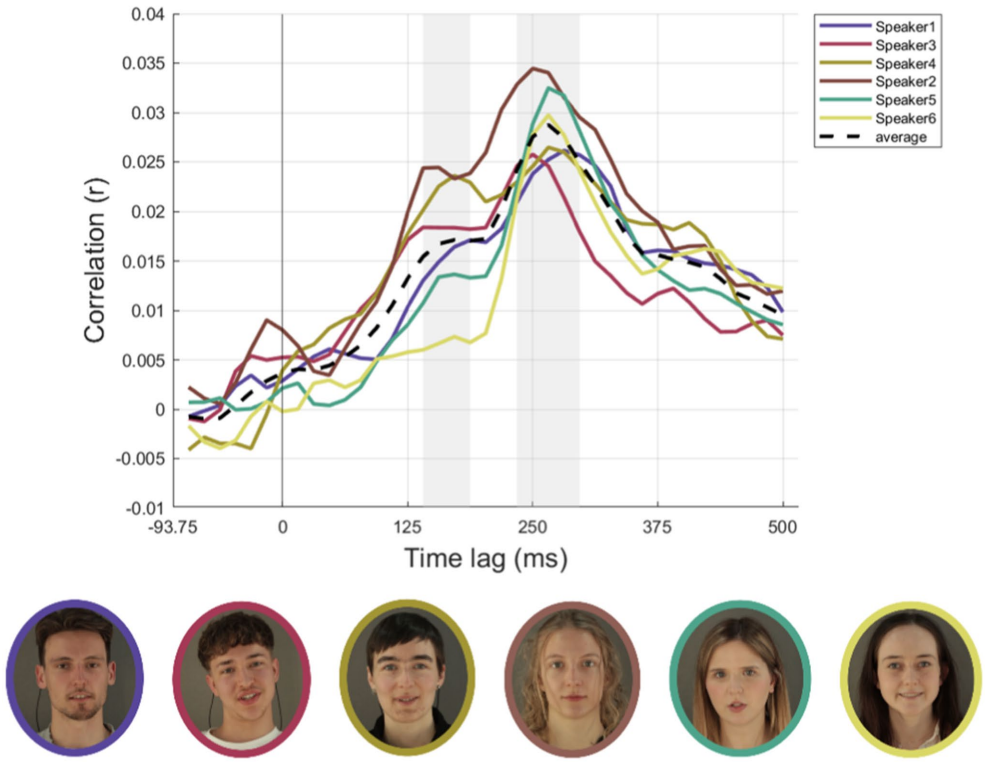

FIGURE 4
Cortical speech tracking for AV, A and ML conditions for all six speakers. The top row illustrates the time-lagged correlations (r_z) between audio envelopes and neural responses for each of the six speakers (speaker1, speaker3, speaker4, speaker2, speaker5, speaker6). Conditions include audio-visual (AV) on the left, auditory-only (A) in the centre, and masked-lips (ML) on the right. Individual speaker data are represented as colored solid lines, with colors corresponding to each speaker (speaker1: purple, speaker3: red, speaker4: yellow-green, speaker2: brown, speaker5: teal, speaker6: yellow). The black dashed line represents the average envelope across speakers. Grey areas represent an early (140–187 ms) and a later (234–296 ms) time lag range of interest. The bottom row displays photographs of each speaker, bordered in colors corresponding to their respective line plots in the top row. Conditions are not separated for noise and no-noise conditions due to the limited number of trials for each individual speaker. Written informed consent was obtained from the individuals for the publication of any potentially identifiable images or data included in this article.

10.3389/fnhum.2025.1560558

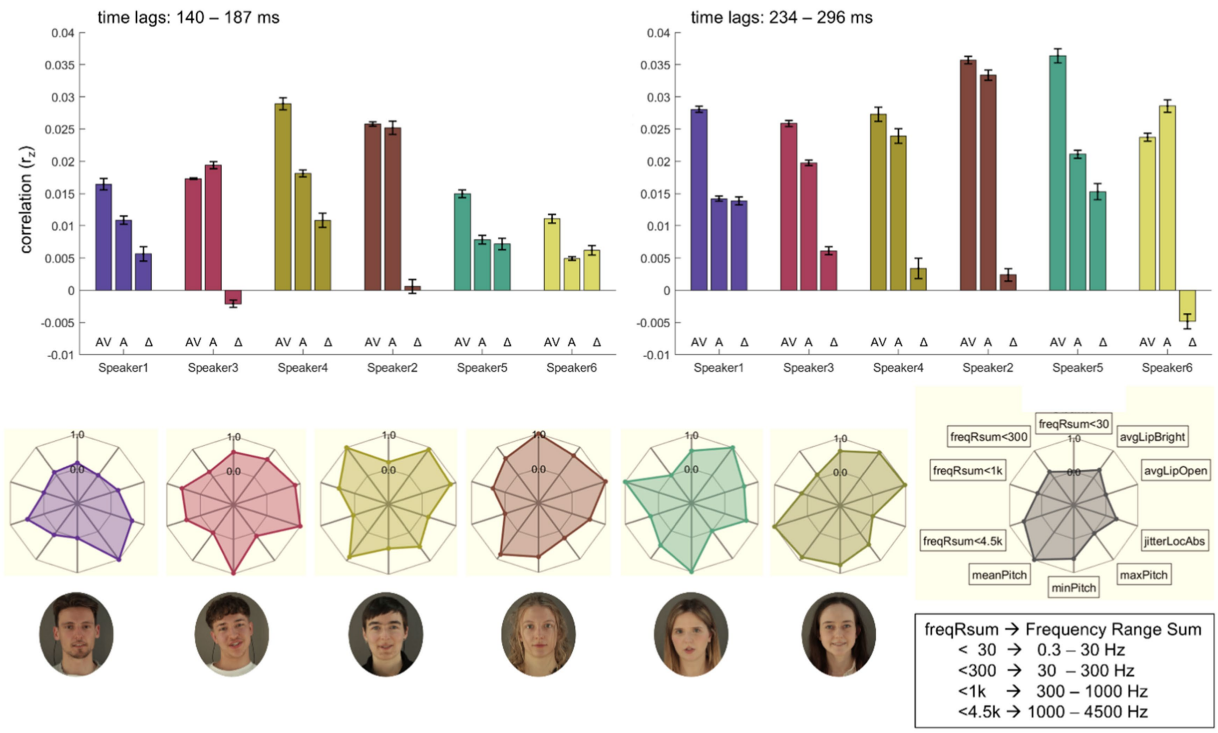

FIGURE 5
Illustration of individual speaker's auditory and visual features. (Top Row): The bar plots illustrate the mean correlations (r_z) across two different time lags (left: 140–187 ms; right: 234–296 ms) for audiovisual (AV), audio-only (A), and audiovisual benefit (AV-A; Δ) conditions for each speaker (speaker1, speaker3, speaker4, speaker2, speaker5, speaker6). Error bars represent the standard error of the mean. (Middle Row): Radar plots depict acoustic feature distributions per speaker, including frequency range sums (freqRsum <30 Hz, <300 Hz, <1 k Hz, <4.5 k Hz), pitch features (mean, min, max), jitter (jitterLocAbs), and lip-based brightness and openness averages (avgLipBright, avgLipOpen). Each radar plot highlights inter-speaker variability across the selected features. (Bottom Row): Portraits of the six speakers visually align with their corresponding radar plots and bar plots, facilitating a direct comparison of individual acoustic feature profiles. Written informed consent was obtained from the individuals for the publication of any potentially identifiable images or data included in this article.

fundamental frequency (F_0) in seconds as a measure of voice stability) is highest for speaker3. Visual features derived from lip brightness (avgLipBright) and openness (avgLipOpen) show further notable differences. All speakers, except for speaker1, exhibited higher values for these visual features. Radar plot representations further illustrate the unique multimodal profiles, capturing variability across frequency sums, pitch, jitter measures, and visual parameters (cf. Figure 5). Following the individual speaker's time-lagged correlations (r_z) between audio envelopes and neural responses, the bar plots in Figure 5 display respective averages across these time lags for conditions AV and A and the audio-visual benefit (AV-A; Δ). In early time lags (140–187 ms) four out of six speakers show the expected pattern, while in later time lags (234–296 ms) five out of six speakers show higher correlations in AV compared to A conditions.

3.2.3 Influence of likeability ratings on individual speaker's cortical speech tracking

On average, participants rated all six speakers comparably high in likeability on a scale from 0–5 (Mean: 3.79 ± 0.02, range: 3.47–3.94).

To investigate the relationship between likeability ratings and neural speech tracking, we fitted a Generalized Linear Mixed Model (GLMM) with speech tracking correlations as the dependent variable and likeability, time lags (early, late), and their interaction as fixed effects. Subject and speaker were included as random effects to account for repeated measures. The model included 204 observations, with 4 fixed effects coefficients and 119 random effects coefficients. The model was fitted using the Laplace approximation, assuming a normal distribution and an identity link function. The results showed a significant interaction between likeability and time lags ($\beta = 0.003$, $t_{(200)} = 2.13$, $p = 0.034$), indicating that the relationship between likeability and speech tracking differed across time lags. The main effect of likeability was not significant ($\beta = -0.002$, $t_{(200)} = -1.43$, $p = 0.156$), nor was the main effect of time lags ($\beta = -0.002$, $t_{(200)} = -0.36$, $p = 0.72$). The intercept was significant ($\beta = 0.0233$, $t_{(200)} = 4.13$, $p < 0.001$), suggesting a positive baseline correlation between speech tracking and the presented stimuli.

The variance in speech tracking was accounted for by random effects at the Subject level ($\sigma = 0.01$) and at the Speaker level nested within Subject ($\sigma = 0.003$), indicating individual differences in speech tracking ability and speaker variability (see Figure 6).

4 Discussion

This study investigated whether young, normal-hearing individuals benefit from congruent facial cues of speakers, when listening to unscripted, natural speech in both quiet and noisy environments. Our results demonstrate that congruent audio-visual input enhances neural speech tracking in noise, with significantly higher correlations (r_z) in the audio-visual condition compared to audio-only and masked-lips

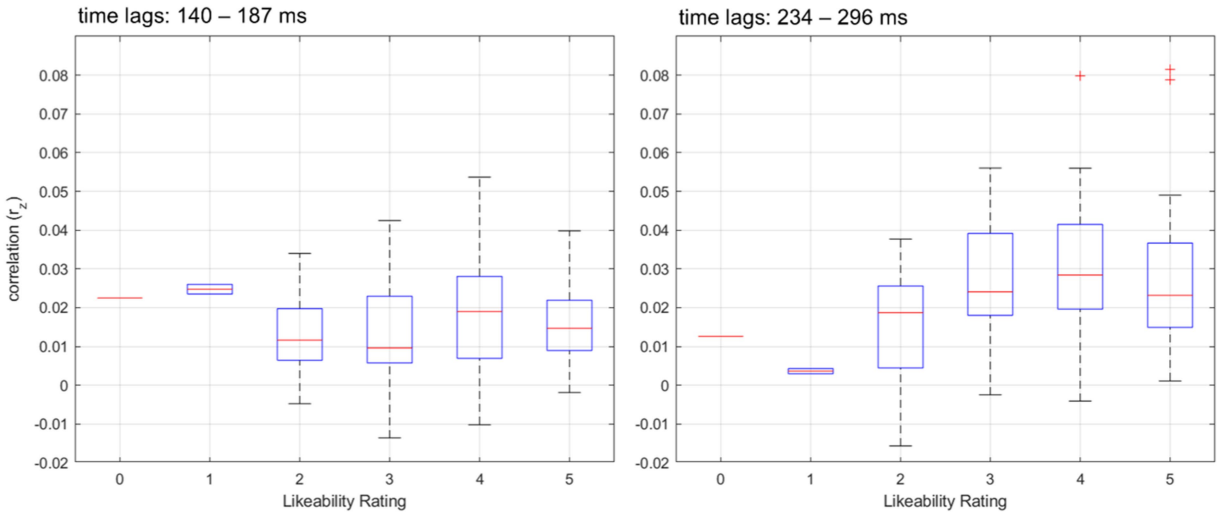

FIGURE 6
Distribution of speech tracking correlations across likeability ratings and time lags. Boxplots showing the distribution of speech tracking correlations (r_z) across different likeability ratings for two distinct time lag ranges. The left panel represents correlations within the early time lags (140–187 ms), while the right panel depicts correlations within the later time lags (234–296 ms). The blue boxes represent the interquartile range (IQR), with the central red line indicating the median correlation value for each likeability rating. The whiskers extend to 1.5 times the IQR, and outliers beyond this range are marked as red crosses.

conditions. These findings support the well-established notion that visual input facilitates speech perception, particularly in challenging listening environments (Peelle and Sommers, 2015; Sumby and Pollack, 2005), even in participants with (self-reported) normal hearing abilities.

4.1 Audio-visual benefit for neural speech tracking with unscripted speech stimuli

Our findings reveal a clear audio-visual benefit for speech envelope tracking, particularly in noisy environments. This aligns with previous studies showing that visual cues enhance auditory processing when the speech signal is degraded (Peelle and Sommers, 2015; Zion Golumbic et al., 2013). In our results, the AV condition consistently induced higher correlation values compared to both the audio-only and masked-lips conditions in noise. Importantly, AV tracking peaked at around 250 ms, representing time lags consistent with cortical auditory–visual integration processes. This supports the idea that visible lip movements help align auditory cortical oscillations with the speech envelope (Arnal and Giraud, 2012; Schroeder et al., 2008). The lack of significant differences between AV and A in the no-noise condition suggests that visual cues primarily become beneficial when the auditory input is compromised, as noted in earlier work (Sumby and Pollack, 2005). In contrast, the absence of a significant difference between ML and A in noise highlights the specific role of visible lip movements in driving the AV benefit. This finding underscores that visual articulation cues are central to the AV advantage in neural speech tracking models.

Interestingly, in the no-noise condition, the masked-lips condition showed, on a descriptive level, even higher neural tracking than the audio-visual condition. Several studies (Rahne et al., 2021; Sönnichsen et al., 2022) conclude that face masks reduce speech perception and increase listening effort in different noise signals even in normal hearing participants. A relevant contribution to this effect, masked-induced auditory deterioration, was not included in our study. We speculate that listeners may have adapted to auditory-only communication during the COVID-19 pandemic, when face masks frequently obscured visual cues like lip movements and caused auditory degradation. Research suggests that prolonged exposure to masked faces can lead to increased reliance on auditory processing and reduced dependence on visual input (Saunders et al., 2021). In the no-noise condition, where the auditory signal was clear and unaltered, participants may have defaulted to auditory-only strategies, ignoring the incongruent or incomplete visual cues in the ML condition. This could have reduced cognitive load, allowing for more efficient speech envelope tracking compared to AV, where lip movements might introduce redundant or misaligned visual information (Yi et al., 2021).

Our study's focus on unscripted, naturally told stories adds ecological validity by resembling real-world listening conditions, where continuous speech provides contextual richness. This approach more effectively enhances neural speech tracking compared to isolated words or sentences, as previously demonstrated (Gross et al., 2013). By contrast, studies relying on highly controlled stimuli may miss the natural dynamics of conversation. While our findings broadly align with prior research, they diverge from a study reporting no AV benefit in single-speaker contexts (O'Sullivan et al., 2013). This discrepancy could arise from differences in stimulus duration or the specific envelope tracking methods used.

4.2 Speaker-specific differences in neural speech tracking

We observed considerable variability in neural speech tracking across speakers, especially in the A and AV conditions. For example, speaker2 exhibited the highest overall tracking correlations in both conditions but had the smallest audio-visual benefit. In contrast,

speakers like speaker4 and speaker5 demonstrated pronounced AV benefits, suggesting that individual speaker characteristics influence how effectively visual and auditory inputs integrate. Specific acoustic traits, such as speaker3's stronger mid-range spectral power or speaker6's higher-range power, may influence their ability to engage neural tracking. Pitch variability also may play a role: speaker6's higher mean pitch and speaker3's wider pitch range likely contributed to distinct neural representations that aid speaker differentiation (Bidelman and Howell, 2016). Brodbeck and Simon (2022) demonstrated that voice pitch variability significantly modulates cortical neural tracking, particularly under conditions requiring selective attention to distinct speakers. These results highlight that speaker-specific traits—including frequency content and articulation variability—shape the dynamics of neural speech tracking.

Crucially, our results indicate that likeability ratings were not uniform across speakers but instead modulated neural speech tracking, particularly in the later time lags (234–296 ms). This suggests that subjective evaluations of a speaker's voice and articulation may influence how effectively their speech is processed. Prior research has linked speaker likeability to perceptual and cognitive factors such as voice clarity, prosody, and familiarity (Zuckerman and Driver, 1989). The observed speaker-specific differences emphasize the importance of accounting for individual multimodal profiles when studying neural speech tracking. This variability is particularly relevant for real-world scenarios where listeners engage with speakers of diverse expressiveness and acoustic profiles. Future studies should further examine how subjective factors such as speaker preference and familiarity dynamically shape audiovisual speech integration.

4.3 Speech features, neural speech tracking and likeability ratings

On a descriptive level, low-frequency spectral power (freqRsum<30) was highest in speakers who also showed greater audiovisual benefit, aligning with previous research emphasizing the importance of low-frequency energy for neural speech tracking (Ding and Simon, 2014; Luo and Poeppel, 2007). Jitter measures, such as jitter_loc_abs, varied across speakers, with some exhibiting higher levels of vocal irregularity than others. While we did not directly assess the perceptual impact of these variations, previous research suggests that increased jitter may enhance speech salience in noisy conditions by introducing subtle acoustic cues that aid in distinguishing the speech signal (Eadie and Doyle, 2005; Oganian et al., 2023). In our data, speakers with greater jitter values did not consistently show higher audiovisual benefit, but given the known role of vocal perturbations like jitter and shimmer in modulating speech clarity (Smiljanic and Gilbert, 2017), it is possible that these micro-level irregularities interact with other acoustic and visual features in shaping neural speech tracking responses. Future work could explore whether specific jitter characteristics contribute to enhanced auditory–visual integration under degraded listening conditions.

In the visual domain, speakers differed in articulatory expressiveness, as measured by lip openness and brightness. Speakers with more pronounced articulation also exhibited higher audiovisual benefit. This is in line with previous work showing that clear visual articulation can aid in integrating auditory and visual speech cues. Overall, these observations highlight individual differences in both acoustic and visual speech features, suggesting that audiovisual benefit may emerge from a combination of speaker-specific characteristics and perceptual integration processes (Campbell, 2007; Munhall et al., 2004).

Beyond acoustic and articulatory properties, our GLMM analysis further revealed that likeability may influence neural speech tracking, but this effect is time-dependent. While there was no significant main effect of likeability, we observed a significant likeability × time. It is indicated that likeability-related differences in speech tracking emerge at later latencies (234–296 ms). This suggests that while early speech tracking might primarily reflect basic auditory encoding processes (e.g., envelope tracking in auditory cortex; Ding and Simon, 2014), later time lags may be more sensitive to higher-order social or cognitive influences (e.g., speaker familiarity, attention allocation, or affective salience). One possible explanation for the delayed effect of likeability on neural tracking is that social and affective processing mechanisms require additional integration time. Previous studies suggest that listener expectations and speaker attributes can modulate cortical speech tracking, particularly when top-down mechanisms (e.g., attention, predictive coding) come into play (Liao et al., 2023; Vanthornhout et al., 2018). If likeability reflects a socially relevant signal, it could shape attention allocation and thus enhance neural tracking at later processing stages. Alternatively, the observed effect might reflect differences in speech comprehension, as previous work has shown that more engaging or socially preferred voices tend to facilitate speech perception (Schmälzle et al., 2015). Crucially, these findings underscore that neural speech tracking is not purely an acoustic-driven process but is modulated by social factors. While classic models of speech tracking emphasize the role of low-frequency auditory information, our results suggest that social and cognitive factors—such as likeability—may influence speech tracking in later, more integrative processing stages. Future research should explore whether these effects generalize to real-world conversational settings, where speaker identity, emotional prosody, and interaction dynamics further shape neural tracking responses.

4.4 Implications for multimodal speech processing

Our findings have significant implications for understanding how the brain integrates auditory and visual cues during natural, unscripted speech. In addition to studies using controlled or scripted stimuli, we show that neural speech tracking is also robust in more ecologically valid listening conditions. The enhanced tracking observed in noisy AV conditions highlights the critical role of visible lip movements in compensating for degraded auditory signals, emphasizing the importance of cross-modal integration in real-world communication. From an application perspective, these results can inform technologies like hearing aids and brain-computer interfaces. Incorporating speaker-specific acoustic and visual profiles could improve auditory attention decoding models, optimizing neural tracking performance in naturalistic settings (Geirnaert et al., 2021). Understanding how individual speaker traits influence audiovisual integration—and attention—is crucial for developing personalized solutions to enhance real-world speech perception.

5 Limitations

Several limitations should be noted. First, we included only a the relatively small number of speakers ($N = 6$), which limits the informative value of our exploratory analyses. While our investigations provide valuable insights, larger datasets including more speaker variability are needed to further explore the role of specific acoustic and visual features in audiovisual benefit. Second, we were not able to fully investigate differences across time lags and conditions due to a limited amount of data. In the main analyses, we focused on one distinct time lag range because no consistent effects were observed across multiple lags and conditions. Similarly, at the speaker level, differences between silent and noise conditions could not be explored due to trial constraints. This limits our ability to determine how speaker traits interact with background noise in shaping neural speech tracking. Future studies should consider expanding the dataset to allow for a more fine-grained analysis of temporal and condition-dependent effects, as well as incorporating subjective biases and emotional expressiveness as additional covariates.

6 Summary

In summary, this study highlights the interplay between speaker-specific acoustic and visual attributes and their effect on audio-visual integration and neural speech tracking. These insights have implications for personalized auditory attention models and assistive technologies, emphasizing the need to account for individual variability in natural, unscripted multi-speaker environments. Future research should extend these findings by exploring multimodal integration in diverse populations, including those with hearing impairments, to further enhance predictive models of auditory attention.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Kommission für Forschungsfolgenabschätzung und Ethik, University of Oldenburg, Oldenburg, Germany. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

MD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. JO: Methodology, Software, Writing – review & editing. GG: Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – review & editing. BM: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. VH: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing. SD: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the Cluster of Excellence "Hearing4all", funded by the German Research Foundation (Deutsche Forschungsgemeinschaft; under Germany's Excellence Strategy – EXC 2177/1 - Project ID 390895286). Mareike Daeglau and Jürgen Otten were supported by a research grant from the German Research Foundation (Deutsche Forschungsgemeinschaft; SPP 2236 project 444761144). Partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project-ID 352015383 – SFB 1330 project B1).

Acknowledgments

The authors thank the speakers (Aaron Möllhof, Anna Dorina Klaus, Janto Klunder, Mara Wendt-Thorne, Laura Knipper & Jupiter Dunkelgut), Focke Schröder and Ina Cera (technical assistance, video editing) and Kevin Brumme (valuable advice on visual designs). The EEG data was acquired with the help of Jennifer Decker and Emma Wiedenmann.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

Arnal, L. H., and Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Bachmann, F. L., MacDonald, E. N., and Hjortkjær, J. (2021). Neural measures of pitch processing in EEG responses to running speech. *Front. Neurosci.* 15:738408. doi: 10.3389/fnins.2021.738408

Bidelman, G. M., and Howell, M. (2016). Functional changes in inter- and intra-hemispheric cortical processing underlying degraded speech perception. *NeuroImage* 124, 581–590. doi: 10.1016/j.neuroimage.2015.09.020

Boersma, P., and Weenink, D. (2024). Praat: Doing phonetics by computer (version 6.4.25) [computer software]. Available online at: http://www.praat.org/ (Accessed August 26, 2024).

Brodbeck, C., and Simon, J. Z. (2022). Cortical tracking of voice pitch in the presence of multiple speakers depends on selective attention. *Front. Neurosci.* 16:828546. doi: 10.3389/fnins.2022.828546

Campbell, R. (2007). The processing of audio-visual speech: empirical and neural bases. *Philos. Trans. Soc B Biol. Sci.* 363, 1001–1010. doi: 10.1098/rstb.2007.2155

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5:e1000436. doi: 10.1371/journal.pcbi.1000436

Cohen, J. (2013). Statistical Power Analysis for the Behavioral Sciences *(2nd ed.)*. *Routledge*. doi: 10.4324/9780203771587

Crosse, M. J., Butler, J. S., and Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J. Neurosci.* 35, 14195–14204. doi: 10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016a). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* 10:604. doi: 10.3389/fnhum.2016.00604

Crosse, M. J., Liberto, G. M. D., and 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. doi: 10.1523/JNEUROSCI.1396-16.2016

Daeglau, M., Otten, J., Mirkovic, B., Grimm, G., Debener, S., and Hohmann, V., (2023). Audiovisual recordings of unscripted monologues [video recording]. https://zenodo.org/records/8082844

Delorme, A., and 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. doi: 10.1016/j.jneumeth.2003.10.009

Ding, N., and Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Front. Hum. Neurosci.* 8:311. doi: 10.3389/fnhum.2014.00311

Eadie, T. L., and Doyle, P. C. (2005). Classification of dysphonic voice: acoustic and auditory-perceptual measures. *J. Voice* 19, 1–14. doi: 10.1016/j.jvoice.2004.02.002

Etard, O., and 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. doi: 10.1523/JNEUROSCI.1828-18.2019

Fam, K., and Waller, D. S. (2006). Identifying likeable attributes: a qualitative study of television advertisements in Asia. *Qual. Mark. Res. Int. J.* 9, 38–50. doi: 10.1108/13522750610640549

Farley, S. D. (2008). Attaining status at the expense of likeability: pilfering Power through conversational interruption. *J. Nonverbal Behav.* 32, 241–260. doi: 10.1007/s10919-008-0054-x

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191. doi: 10.3758/BF03193146

Fu, Z., Wu, X., and Chen, J. (2019). Congruent audiovisual speech enhances auditory attention decoding with EEG. *J. Neural Eng.* 16:066033. doi: 10.1088/1741-2552/ab4340

Geirnaert, S., Vandecappelle, S., Alickovic, E., De Cheveigne, A., Lalor, E., Meyer, B. T., et al. (2021). Electroencephalography-based auditory attention decoding: toward Neurosteered hearing devices. *IEEE Signal Process. Mag.* 38, 89–102. doi: 10.1109/MSP.2021.3075932

Grimm, G., Luberadzka, J., and Hohmann, V. (2019). A toolbox for rendering virtual acoustic environments in the context of audiology. *Acta Acust. United Acust* 105, 566–578. doi: 10.3813/AAA.919337

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., et al. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol.* 11:e1001752. doi: 10.1371/journal.pbio.1001752

Herzke, T., and Hohmann, V. (2007). Improved numerical methods for Gammatone Filterbank analysis and synthesis. *Acta Acust United Acust* 93, 498–500.

Hohmann, V. (2002). Frequency analysis and synthesis using a Gammatone filterbank. *Acta Acustica United with Acustica* 88, 433–442.

Hohmann, V., Paluch, R., Krueger, M., Meis, M., and Grimm, G. (2020). The virtual reality lab: realization and application of virtual sound environments. *Ear Hear.* 41, 31S–38S. doi: 10.1097/AUD.0000000000000945

Holleman, G. A., Hooge, I. T. C., Kemner, C., and Hessels, R. S. (2020). The 'real-world approach' and its problems: a critique of the term ecological validity. *Front. Psychol.* 11:721. doi: 10.3389/fpsyg.2020.00721

Holm, S. (1979). A simple sequentially Rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

Holtze, B., Rosenkranz, M., Bleichner, M., Jaeger, M., and Debener, S. (2023). Eye-blink patterns reflect attention to continuous speech. *Adv. Cogn. Psychol.* 19, 177–200. doi: 10.5709/acp-0387-6

Jaeger, M., Mirkovic, B., Bleichner, M. G., and Debener, S. (2020). Decoding the attended speaker from EEG using adaptive evaluation intervals captures fluctuations in attentional listening. *Front. Neurosci.* 14:603. doi: 10.3389/fnins.2020.00603

Jiang, J., and Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1193–1209. doi: 10.1037/a0023100

Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., et al. (2020). The quest for ecological validity in hearing science: what it is, why it matters, and how to advance it. *Ear Hear.* 41, 5S–19S. doi: 10.1097/AUD.0000000000000944

Kothe, C., Shirazi, S. Y., Stenner, T., Medine, D., Boulay, C., Grivich, M. I., et al. (2024). The lab streaming layer for synchronized multimodal recording. *Biorxiv*. doi: 10.1101/2024.02.13.580071

Li, J., Hong, B., Nolte, G., Engel, A. K., and Zhang, D. (2023). EEG-based speaker–listener neural coupling reflects speech-selective attentional mechanisms beyond the speech stimulus. *Cereb. Cortex* 33, 11080–11091. doi: 10.1093/cercor/bhad347

Liao, W., Oh, Y. J., Zhang, J., and Feng, B. (2023). Conversational dynamics of joint attention and shared emotion predict outcomes in interpersonal influence situations: An interaction ritual perspective. *Journal of Communication*, 73, 342–355. doi: 10.1093/joc/jqad003

Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001–1010. doi: 10.1016/j.neuron.2007.06.004

Mallick, D. B., Magnotti, J. F., and Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychon. Bull. Rev.* 22, 1299–1307. doi: 10.3758/s13423-015-0817-4

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

Mirkovic, B., Bleichner, M. G., De Vos, M., and Debener, S. (2016). Target speaker detection with concealed EEG around the ear. *Front. Neurosci.* 10:349. doi: 10.3389/fnins.2016.00349

Mirkovic, B., Debener, S., Jaeger, M., and Vos, M. D. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *J. Neural Eng.* 12:046007. doi: 10.1088/1741-2560/12/4/046007

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychol. Sci.* 15, 133–137. doi: 10.1111/j.0963-7214.2004.01502010.x

O'Sullivan, J. A., Crosse, M. J., Power, A. J., and Lalor, E. C. (2013). The effects of attention and visual input on the representation of natural speech in EEG. 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC), 2800–2803.

Oganian, Y., Bhaya-Grossman, I., Johnson, K., and Chang, E. F. (2023). Vowel and formant representation in the human auditory speech cortex. *Neuron* 111, 2105–2118.e4. doi: 10.1016/j.neuron.2023.04.004

Park, H., Kayser, C., Thut, G., and Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *Elife* 5:e14521. doi: 10.7554/eLife.14521

Peelle, J. E., and Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Front. Psychol.* 3:320. doi: 10.3389/fpsyg.2012.00320

Peelle, J. E., and Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex* 68, 169–181. doi: 10.1016/j.cortex.2015.03.006

Pion-Tonachini, L., Kreutz-Delgado, K., and Makeig, S. (2019). ICLabel: an automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* 198, 181–197. doi: 10.1016/j.neuroimage.2019.05.026

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. doi: 10.1016/j.neuroimage.2019.04.017

Puschmann, S., Steinkamp, S., Gillich, I., Mirkovic, B., Debener, S., and Thiel, C. M. (2017). The right Temporoparietal junction supports speech tracking during selective listening: evidence from concurrent EEG-fMRI. *J. Neurosci.* 37, 11505–11516. doi: 10.1523/JNEUROSCI.1007-17.2017

Rahne, T., Fröhlich, L., Plontke, S., and Wagner, L. (2021). Influence of surgical and N95 face masks on speech perception and listening effort in noise. *PLoS One* 16:e0253874. doi: 10.1371/journal.pone.0253874

Rosenkranz, M., Holtze, B., Jaeger, M., and Debener, S. (2021). EEG-based Intersubject correlations reflect selective attention in a competing speaker scenario. *Front. Neurosci.* 15:685774. doi: 10.3389/fnins.2021.685774

Saunders, G. H., Jackson, I. R., and Visram, A. S. (2021). Impacts of face coverings on communication: an indirect impact of COVID-19. *Int. J. Audiol.* 60, 495–506. doi: 10.1080/14992027.2020.1851401

Scherer, K. R., Ellgring, H., Dieckmann, A., Unfried, M., and Mortillaro, M. (2019). Dynamic facial expression of emotion and observer inference. *Front. Psychol.* 10:508. doi: 10.3389/fpsyg.2019.00508

Schmälzle, R., Häcker, F. E. K., Honey, C. J., and Hasson, U. (2015). Engaged listeners: Shared neural processing of powerful political speeches. *Social Cognitive and Affective Neuroscience*, 10, 1137–1143. doi: 10.1093/scan/nsu168

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.* 12, 106–113. doi: 10.1016/j.tics.2008.01.002

Smiljanic, R., and Gilbert, R. C. (2017). Acoustics of clear and noise-adapted speech in children, young, and older adults. *J. Speech Lang. Hear. Research: JSLHR* 60, 3081–3096. doi: 10.1044/2017_JSLHR-S-16-0130

Sönnichsen, R., Llorach Tó, G., Hochmuth, S., Hohmann, V., and Radeloff, A. (2022). How face masks interfere with speech understanding of Normal-hearing individuals: vision makes the difference. *Otol. Neurotol.* 43, 282–288. doi: 10.1097/MAO.0000000000003458

Stropahl, M., and Debener, S. (2017). Auditory cross-modal reorganization in cochlear implant users indicates audio-visual integration. *NeuroImage* 16, 514–523. doi: 10.1016/j.nicl.2017.09.001

Stropahl, M., Schellhardt, S., and Debener, S. (2017). McGurk stimuli for the investigation of multisensory integration in cochlear implant users: the Oldenburg audio visual speech stimuli (OLAVS). *Psychon. Bull. Rev.* 24, 863–872. doi: 10.3758/s13423-016-1148-9

Sumby, W. H., and Pollack, I. (2005). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Tomar, P. S., Mathur, K., and Suman, U. (2024). Fusing facial and speech cues for enhanced multimodal emotion recognition. *Int. J. Inf. Technol.* 16, 1397–1405. doi: 10.1007/s41870-023-01697-7

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., and Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *J. Assoc. Res. Otolaryngol.* 19, 181–191. doi: 10.1007/s10162-018-0654-z

Wang, B., Xu, X., Niu, Y., Wu, C., Wu, X., and Chen, J. (2023). EEG-based auditory attention decoding with audiovisual speech for hearing-impaired listeners. *Cereb. Cortex* 33, 10972–10983. doi: 10.1093/cercor/bhad325

Wiedenmann, E., Daeglau, M., Otten, J., Mirkovic, B., Grimm, G., Hohmann, V., et al. (2023). The influence of likeability ratings of audio-visual stimuli on cortical speech tracking with Mobile EEG in virtual environments. In DAGA 2023 49. Jahrestagung für Akustik. 1439–1442.

Yi, H., Pingsterhaus, A., and Song, W. (2021). Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic. *Front. Psychol.* 12:682677. doi: 10.3389/fpsyg.2021.682677

Younan, M., and Martire, K. A. (2021). Likeability and expert persuasion: Dislikeability reduces the perceived persuasiveness of expert evidence. *Front. Psychol.* 12:785677. doi: 10.3389/fpsyg.2021.785677

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party'. *Neuron* 77, 980–991. doi: 10.1016/j.neuron.2012.12.037

Zion Golumbic, E. M., Poeppel, D., and Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. *Brain Lang.* 122, 151–161. doi: 10.1016/j.bandl.2011.12.010

Zuckerman, M., and Driver, R. E. (1989). What sounds beautiful is good: the vocal attractiveness stereotype. *J. Nonverbal Behav.* 13, 67–82. doi: 10.1007/BF00990791