

RESEARCH REPORT OPEN ACCESS

The Cortical Contribution to the Speech-Frequency-Following Response Is Not Modulated by Visual Information

Jasmin Riegel¹  | Alina Schüller¹ | Mike Thornton¹ | Alexander Wißmann² | Steffen Zeiler² | Dorothea Kolossa²  | Tobias Reichenbach¹ 

¹Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany | ²Human-Centered Information Technologies, Technische Universität Berlin, Berlin, Germany

Correspondence: Tobias Reichenbach (tobias.j.reichenbach@fau.de)

Received: 3 February 2026 | **Revised:** 13 July 2026 | **Accepted:** 1 September 2026

Associate Editor: Edmund C. Lalor

Keywords: audiovisual speech perception | avatar | speech-FFR | temporal response function

ABSTRACT

Seeing a speaker's face can significantly aid understanding, particularly in challenging acoustic environments. An early neural response implicated in audiovisual speech processing is the frequency-following response (speech-FFR), which occurs at the fundamental frequency of the speech signal. This response arises from both subcortical areas and the auditory cortex. A previous study showed that subcortical responses to speech can be enhanced when a listener can see the talker's face. Here, we examined the cortical contribution to the speech-FFR and its potential modulation by visual information, motivated in part by the growing use of artificially generated talking-face avatars to support speech comprehension. We recorded MEG responses to four types of audiovisual signals: a still image, an artificially generated avatar, a degraded video, and a natural video. The audio stimuli were presented in a substantial level of background noise to make behavioral audiovisual effects stand out. Speech-in-noise comprehension increased significantly from the audio-only condition to the avatar and the degraded video, and further to the natural video. Moreover, we found that all types of audiovisual stimuli yielded robust speech-FFRs in the auditory cortex at an early latency of around 30 ms. However, the magnitude of this neural response was neither enhanced nor attenuated by the videos, nor could the cortical contribution of the speech-FFR explain a significant portion of the variance in the behavioral comprehension scores. Our results suggest that visual modulation of the speech-FFR in the auditory cortex is, if existent, too small to be measurable in scenarios where speech occurs in considerable background noise.

1 | Introduction

Understanding speech in acoustically challenging environments can be difficult, not only for individuals with normal hearing but also especially for those with hearing impairments

(Miles et al. 2022). Speech comprehension can be substantially enhanced when listeners see the speaker, since visual information, such as lip movements and facial cues, supports auditory speech processing (Thézé et al. 2020; Nirme et al. 2020). The improvement in speech comprehension is most notable in the

Abbreviations: ANOVA, analysis of variance; AV, audiovisual; BEM, boundary element model; ECoG, electrocorticography; EEG, electroencephalography; f0, fundamental frequency; FFR, frequency-following response; FIR, finite impulse response; Fps, frames per second; Hz, hertz; LCMV, linearly constrained minimum variance; LH, left hemisphere; MEG, magnetoencephalography; OLSA, Oldenburger Satztest; RH, right hemisphere; ROI, region of interest; SNR, signal-to-noise ratio; SRT50, speech reception threshold at 50%; STG, superior temporal gyrus; TRF, temporal response function.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *European Journal of Neuroscience* published by Federation of European Neuroscience Societies and John Wiley & Sons Ltd.

presence of background noise, a principle known as inverse effectiveness (Sumby and Pollack 1954; Grant and Seitz 2000; Ross et al. 2007).

Numerous studies have investigated the neural mechanisms underlying audiovisual speech integration. Most of this work has focused on event-related potentials (Van Wassenhove et al. 2005; Pilling 2009) or on cortical tracking of the low-frequency fluctuations in the speech envelope (Crosse et al. 2015, 2016; Varano et al. 2025). It identified, among others, a role of the delta frequency band in linking auditory and visual information on the word level (Varano et al. 2025; Riegel et al. 2026).

However, neural responses to speech also occur in the much higher gamma frequency range. In particular, neural activity can phase-lock to the fundamental frequency of speech, as well as to a lesser extent, to its higher harmonics. This frequency-following response (speech-FFR) has initially been assessed in response to single repeated syllables (Russo et al. 2004; Skoe and Kraus 2010).

Recent studies have shown how, through statistical modeling, the speech-FFR can also be detected in response to continuous, nonrepetitive speech in which the fundamental frequency varies over time (Forte et al. 2017; Etard et al. 2019). Because this continuous speech-FFR is estimated from naturalistic, running speech rather than from many repetitions of an isolated syllable, it differs from the classical, syllable-based FFR in important aspects. In particular, it is not computed by averaging the neural response over many repetitions, but instead through statistical estimation of the neural signal from the time-varying fundamental frequency, yielding estimates of both amplitude and latency. Comparisons between the two paradigms therefore have to be drawn with care.

The speech-FFR, as measured with electroencephalography (EEG), has largely subcortical origins, including in particular the inferior colliculus. The speech-FFR has therefore been conceptualized as reflecting predominantly feedforward, audio-stimulus-driven encoding in the auditory brainstem and midbrain (Coffey et al. 2019). However, there have also been reports of modulatory influences beyond purely feedforward processing. In particular, a series of studies suggested that long-term musical training can enhance the fidelity of subcortical FFRs to speech and nonspeech sounds (Musacchia et al. 2007; Musacchia et al. 2008). This interpretation has recently been challenged by a large-scale, preregistered study that failed to find evidence for such effects of musical training (Whiteford et al. 2025).

In contrast to experience-related effects such as musical training, the extent to which the subcortical speech-FFR is modulated by top-down attention remains debated (Lehmann and Schönwiesner 2014; Forte et al. 2017; Stoll et al. 2025). Of particular interest for audiovisual speech perception, audiovisual stimuli have been found to enhance the magnitude of the FFR (Musacchia et al. 2007). Of note, however, a previous study by the same authors found reduced onset responses to single repeated syllables in audiovisual speech, but no audiovisual effect on the FFR (Musacchia et al. 2006).

Coffey et al. (2016) used magnetoencephalography (MEG) to show that the speech-FFR elicited by repeated syllables contains a reliable cortical contribution, in addition to the responses from the brainstem and the midbrain. Kulasingham et al. (2020) added to this finding by showing that the cortical contribution to the speech-FFR could also be identified when elicited by continuous speech.

Since then, factors that potentially impact the subcortical speech-FFR have increasingly been investigated regarding their effect on the cortical contribution. In particular, we and others recently demonstrated sizable and robust attentional modulation of the speech-FFR in the auditory cortex (Schüller et al. 2023; Commuri et al. 2023). On the other hand, we were unable to find an effect of musical training on this response (Riegel et al. 2024). The demonstration that the cortical contribution to the speech-FFR is subject to top-down attentional control establishes it as a genuine target of higher level influence, and thereby a plausible candidate site at which visual speech, too, could shape early pitch-related encoding. Testing this directly is the aim of the present study.

Here, we therefore investigated whether visual speech modulates the cortical contribution to the speech-FFR. Establishing whether such a modulation exists, and bounding its magnitude, is relevant both for understanding at which stage of the processing hierarchy audiovisual integration acts. It may also guide practical efforts to enhance speech intelligibility with artificially generated visual speech. To investigate this, we measured the neural response to continuous audiovisual speech in 32 participants using MEG. Stimuli were presented in four conditions: a still image of the speaker (referred to as audio-only condition), an avatar, a degraded version of the speaker, and the natural video.

We included an artificially generated avatar in the experiment because visual access to a speaker's face is not always available, for example, in public settings such as train station announcements. In such cases, avatars may offer a promising alternative by providing visual speech cues to support comprehension. Previous behavioral studies have shown that avatars can partially improve speech understanding in noisy environments, although typically to a lesser extent than natural visual speech (Varano et al. 2022; Nirme et al. 2020; Shan et al. 2022).

The degraded video was included to determine the putative effect of simplified visual signals on the cortical speech-FFR, that is, of visual signals that contain sufficient information to enhance speech-in-noise comprehension, although to a lesser extent than natural videos. To maximize the potential audiovisual benefit—and, consequently, the expected effect size in the cortical speech-FFR—we followed the principle of inverse effectiveness and embedded the audiovisual speech stimuli in four-talker babble noise.

2 | Materials and Methods

We used the data from a previous MEG study on audiovisual speech perception (Riegel et al. 2026). Thirty-two participants listened to a single speaker uttering sentences in background noise.

After each sentence, the participants repeated what they understood. The audiovisual speech material was presented in four different visual conditions: (1) a still image of the speaker (referred to as audio-only), (2) a synthetically generated avatar, (3) a degraded video, and (4) the natural video. While our previous study focused on tracking low-frequency amplitude fluctuations of speech in the auditory cortex, in the current investigation, we determined the higher frequency tracking of the presented speech in the auditory cortex, namely, the speech-FFR. In Figure 1, an overview of the stimuli and data processing is presented.

2.1 | Audiovisual Stimuli

The speech material was based on recordings of a native German speaker uttering short sentences from the Oldenburger Satztest, a standardized matrix-sentence corpus designed to generate grammatically correct and phonetically balanced but semantically unpredictable sentences (Wagener et al. 1999). Each sentence follows the same syntactic structure—name, verb, number, adjective, and object. The lexical items are randomly drawn from sets of 10 alternatives per position, resulting in one million possible unique sentences.

Recordings were conducted at 29.97fps with accompanying audio sampled at 44.1 kHz. We collected 600 unique sentences and screened them to ensure consistent pronunciation and the absence of background noise. After this quality control step, 571 sentences remained and served as the basis for generating the four stimulus types.

To create the degraded audiovisual condition, we processed the original videos using the edgedetect function from Ffmpeg, which emphasizes prominent contours while suppressing surface detail.

The avatar-based condition was generated by combining a still image of the speaker with the corresponding audio file using software provided by D-ID (D-ID 2025). No further preprocessing of audio or image was necessary. The image was animated within the software by applying a convolutional neural network to encode the speaker's facial features and a generative adversarial model to animate the face in synchrony with the speech signal.

All sentences were presented in background noise to increase the potential effects of audiovisual integration (Ma et al. 2009).

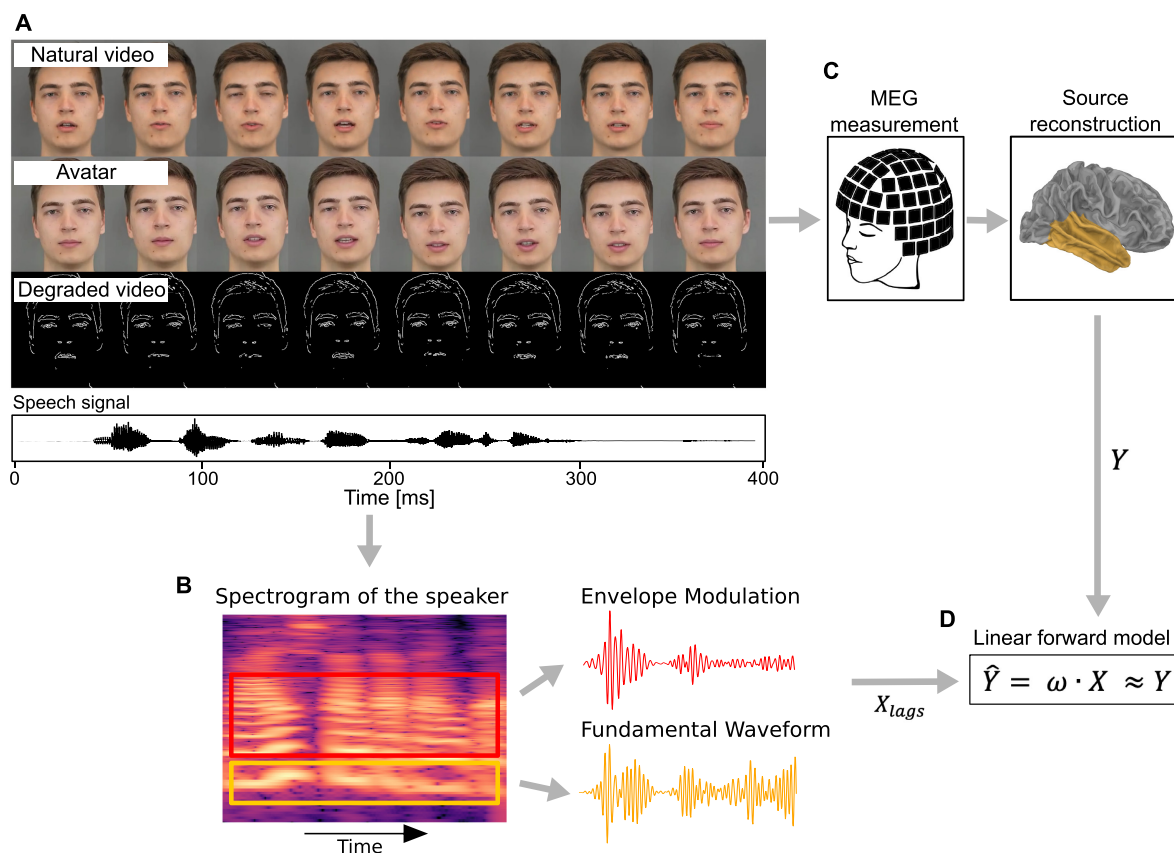


FIGURE 1 | Audiovisual stimuli and data processing pipeline. (A) Three types of audiovisual stimuli are generated for this study: the speaker's natural video, an avatar generated with DNNs from a still image and an audio file, and a degraded version of the natural video obtained by applying an edge detection algorithm. Below the frames, the corresponding speech signal is shown. (B) Two features (X_{lags}) are extracted from the spectrogram of the speech signal: a representation of the fundamental waveform that oscillates at the fundamental frequency (yellow) and the envelope modulation of its higher harmonics (red). (C) MEG measurements were conducted while participants watched the stimuli. The MEG data were source-reconstructed onto the auditory cortex (Y). (D) The speech-FFR is determined by estimating the source-reconstructed MEG data ($\hat{Y}$) from the two speech features (X_{lags}) using a linear forward model. The magnitude of the resulting coefficients (ω) over time yields the temporal response functions (TRFs).

We employed four-talker babble noise that began shortly before sentence onset. We determined the signal-to-noise ratio during a pilot study, with six participants, aiming for approximately 50% comprehension in the audio-only condition. The participants listened to 100 OLSA sentences and repeated every word they had understood. Via an adaptive procedure, the babble noise was increased to reach the desired level. This procedure resulted in an SNR of approximately -4 dB and an overall sound pressure level of 68 dB(A). We used this level uniformly for all participants.

2.2 | Experimental Procedure

The study was granted ethical permission by the ethics board of the University Hospital Erlangen (registration no. 22-361-S).

Thirty-three healthy, right-handed adults (17 female, 16 male; age range 19–31 years) were recruited for the experiment. Each participant completed two sessions of up to 60 min, scheduled 1–3 weeks apart. One female did not finish both sessions, leaving data from 32 participants for analysis.

To familiarize the participants with the restricted vocabulary and thereby minimize later learning effects, the first MEG session started with 80 audio-only sentences with adaptive background noise, estimating each participant's individual SRT50. Following the training, participants were exposed to the different types of audiovisual stimuli. The experiment was designed in blocks of sentences, with each block containing three randomly selected sentences from each condition. The order of all stimuli within them was fully randomized. Sentences were not grouped by condition. For every stimulus type, 60 sentences were presented. Individual sentences lasted approximately 3 s, yielding roughly 3 min of material per condition.

After each sentence, participants repeated all words they recognized while viewing the OLSA response matrix.

2.3 | Data Acquisition and Preprocessing

MEG recordings were obtained with participants lying supine and with eyes open using a 248-channel whole-head system (4D-Neuroimaging). The data were recorded at a sampling rate of 1017.25 Hz in a magnetically shielded room. During acquisition, an online analog band-pass filter from 1 to 200 Hz was applied. The system further incorporated a calibrated linear combination of 23 reference sensors from 4D Neuroimaging (San Diego, CA, USA), enabling effective suppression of environmental magnetic interference. Before each session, we digitized the head shape along with five fiducial points using a Polhemus tracking system.

Subsequent offline preprocessing was carried out in MNE-Python (Gramfort et al. 2014). We removed power-line artifacts with a 50-Hz notch filter (zero-phase FIR, transition bandwidth 0.5 Hz) and then resampled the data to 1000 Hz.

The auditory stimuli were presented via a custom setup designed by Schilling et al. (2020). To ensure precise timing, the audio signal was simultaneously routed to an MEG channel,

and its alignment with the presented sound was obtained by cross-correlation, yielding temporal precision on the order of 1 ms. Visual material was shown on a screen placed above the participant. A constant delay between audio and video of a presented sentence was introduced by the audiovisual presentation setup. This delay was measured and corrected in the stimuli following the procedure described in Varano et al. (2023).

2.4 | Neural Source Localization

We reconstructed neural sources using MNE-Python (Gramfort et al. 2014) together with the FreeSurfer “fsaverage” brain (Fischl 2012). To align this template to each participant, we co-registered it to the digitized head shape and fiducial points by applying rotations, translations, and uniform scaling based on the recorded head position data.

From the co-registered anatomy, we generated a volumetric grid with 5-mm spacing. Cortical regions of interest were then defined according to the *Aparc* atlas and encompassed bilateral temporal areas known to contribute to audiovisual speech processing (Van Attevelde et al. 2010; Macaluso et al. 2004). These ROIs included the inferior, middle, and superior temporal gyri, the transverse temporal gyri, the banks of the superior temporal sulci, and the fusiform, entorhinal, parahippocampal, and temporal pole regions. In total, this produced 862 source voxels (432 in the left hemisphere). Forward models were computed using a boundary element model (BEM).

To estimate source activity, we applied a linearly constrained minimum variance (LCMV) beamformer (Bourgeois and Minker 2009). The spatial filters were computed using the empirical covariance of the MEG data and the noise covariance derived from empty-room recordings. Filters were normalized to achieve unit noise gain, and their orientation was constrained to surface-normal directions.

2.5 | Behavioral Analysis

After each presented audiovisual stimulus, participants repeated every word they understood. For each stimulus type, the averaged speech comprehension of each participant was calculated.

2.6 | Linear Forward Models to Determine Neural Speech Tracking

The speech-FFRs were computed by estimating the neural response from two speech features, the fundamental waveform and the envelope modulation of the higher harmonics. The fundamental frequency was computed from the speech signal using the probabilistic YIN-algorithm (Mauch and Dixon 2014). The range of the estimated f_0 (70–120 Hz) was used to define the passband edge frequencies of a band-pass filter applied to the speech stimuli in order to extract the fundamental waveform. The envelope modulation of the higher harmonics, first described by Hertrich et al. (2012), was obtained by first applying a model of the auditory periphery and subsequently

filtering the resulting signal in the same frequency range as the fundamental frequency (Chi et al. 2005). In detail, the model begins by applying a set of constant-Q band-pass filters, which are designed to provide frequency-dependent resolution across the auditory spectrum. The filter outputs are subsequently processed using nonlinear compression and multiscale derivative operations, leading to enhanced spectral selectivity. Next, 86 frequency channels corresponding to harmonics above 300 Hz were selected. Within each channel, the amplitude envelope was extracted and band-pass filtered around the fundamental frequency f_0 (70–120 Hz). Finally, these filtered envelopes were averaged across all channels to produce a single high-frequency envelope modulation feature.

The temporal response functions (TRFs) were derived using the stimuli brain mapping toolbox sPyEEG (Guilleminot et al. 2026). The two speech features were then used as two predictors $x_i(t)$, with $i \in \{1, 2\}$, to estimate the MEG response $y^{(j)}(t)$ at voxel j and time t through a linear model with temporal delays:

$$y^{(j)}(t) = \sum_{\tau, i} \omega_i^{(j)}(\tau) x_i(t - \tau) + \epsilon^{(j)}(t),$$

where $y^{(j)}(t)$ denotes the measured MEG response, $x_i(t - \tau)$ is the i th stimulus feature at time lag τ , the weights $\omega_i^{(j)}(\tau)$ are the TRFs for the i th predictor variable and the neural activity at voxel j at time lag τ , and $\epsilon^{(j)}(t)$ is the residual error term.

For training the model weights, ridge regression was applied with a normalized regularization parameter, such that the cost function reads:

$$\min_{\omega} \|Y^{(j)} - X\omega^{(j)}\|^2 + \frac{\lambda}{\langle \text{eig}(X^T X) \rangle} \|\omega^{(j)}\|^2,$$

where X is the matrix of predictor variables at different time lags, $Y^{(j)}$ is the source-reconstructed measurement data at voxel j , λ is the defined regularization parameter, and $\langle \text{eig}(X^T X) \rangle$ is the mean eigenvalue of $X^T X$ used to normalize the regularization strength.

A range of regularization parameters was applied, $\lambda = [10^{-3}, \dots, 10^3]$. The regularization parameter yielding the highest reconstruction score was estimated using fivefold cross-validation and Pearson correlation for each of the four audiovisual conditions. This yielded $\lambda = 10$. A slightly higher value of $\lambda = 100$ was chosen as the optimal parameter to additionally ensure the avoidance of overfitting. Before computing the linear forward model, both stimulus and MEG data were z-scored and concatenated for each participant and condition.

The TRFs were calculated for each of the 862 voxels between -50 and 170 ms. The absolute values of the resulting TRFs were averaged across voxels to obtain an individual TRF for each participant, condition, and predictor variable. For the population-level TRFs, the average across participants was computed. The time lag of the peak in this response was identified. For the statistical analysis, the peaks in the TRFs of the individual participants were determined within a ± 50 -ms range around the latency of the peak in the population average.

2.7 | Statistical Analysis

We first tested whether the TRFs at the population level differed significantly from noise. We therefore generated a noise distribution using all data from the TRF at negative delays from -40 to 0 ms and calculated empirical p -values for each time lag at the population level.

An ANOVA was conducted to assess differences in speech comprehension across the different audiovisual stimulus conditions. Following significant ANOVA results, pairwise comparisons between conditions were performed using paired t -tests.

An ANOVA was also used to test for differences in the amplitudes and latencies of the TRF peaks between the four audiovisual conditions. Additionally, repeated-measures correlations were employed to assess the putative relationship between the neural data and each participant's speech comprehension score.

The neural response was also analyzed for lateralization. Therefore, the signal strength at the population-level peak latency was extracted for each voxel separately. The data were normally distributed according to the Shapiro–Wilk test. Consequently, the signal strengths between the source points in the left hemisphere and in the right hemisphere were compared with each other through a t -test followed by Bonferroni correction for multiple comparisons.

3 | Results

3.1 | Speech Comprehension

We determined the average comprehension scores of the participants on the population level, separately for each of the four types of audiovisual stimuli. We reported the resulting speech comprehension scores already in our previous work on this dataset (Riegel et al. 2026). A repeated-measures ANOVA across all audiovisual conditions revealed a significant main effect of condition on speech comprehension ($p < 0.001$).

In the audio-only condition, participants understood, on average, 66% of the words correctly (Figure 2). Their understanding increased significantly ($p < 0.001$) to an average of 78% when they watched an avatar of the speaker while listening to the spoken sentences. The same increase ($p < 0.001$) was observed between the audio-only condition and the degraded video, in which only strong edges in the speaker's face were visible. This condition was designed to serve as a control version of the avatar. This worked well, as the participants correctly understood 76% of the words, which did not differ significantly from the avatar condition. Comprehension increased significantly again ($p < 0.001$) when watching the natural video, with which the participants correctly understood, on average, 84% of the words.

3.2 | Speech-FFR

We determined the speech-FFR by relating two speech features to the MEG data. The first feature, the fundamental waveform,

captured neural activity phase-locked to the fundamental frequency of the speech. The second feature, the envelope modulation of the higher harmonics, captured the power fluctuations

in the higher harmonics at the fundamental frequency and the related neural response.

We found that both features produced significant peaks in their TRFs (Figure 3). The TRFs exhibited comparable behavior across the four types of audiovisual conditions. Neural responses to the fundamental waveform peaked between 30 and 36 ms (Figure 3A), whereas responses to the envelope modulation occurred slightly earlier between 26 and 30 ms (Figure 3B). These latencies spanned only a narrow range and did not differ systematically across the four audiovisual conditions, for either speech feature (ANOVA, $p = 0.76$).

To assess whether audiovisual integration was apparent in the speech-FFR, we compared the response strength across the different audiovisual conditions. Therefore, ANOVA was used to test for significance in the distribution of the peak magnitudes of the individual participants across all conditions. Neither for the fundamental waveform (Figure 3C, left) nor for the envelope modulation (Figure 3D, left) did these yield significant results (fundamental waveform: $p = 0.17$; envelope modulation: $p = 0.74$).

Although a direct comparison of the magnitude of the speech-FFR across the four conditions did not produce significant results, we further examined whether the trend in neural signal strength correlated with the speech comprehension scores. We therefore computed repeated-measures correlations. We again

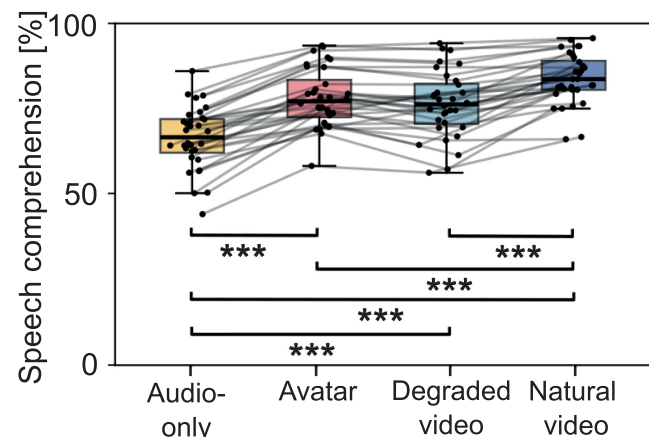


FIGURE 2 | Speech comprehension scores for each audiovisual condition across participants. Comprehension was lowest for the audio-only condition, higher for the avatar and degraded video, and most words per sentence were understood for the natural video. All increases in comprehension were strongly significant ($p < 0.001$) except between the avatar and its control condition, the degraded video. The lines connect speech comprehension scores of individual participants between neighboring conditions.

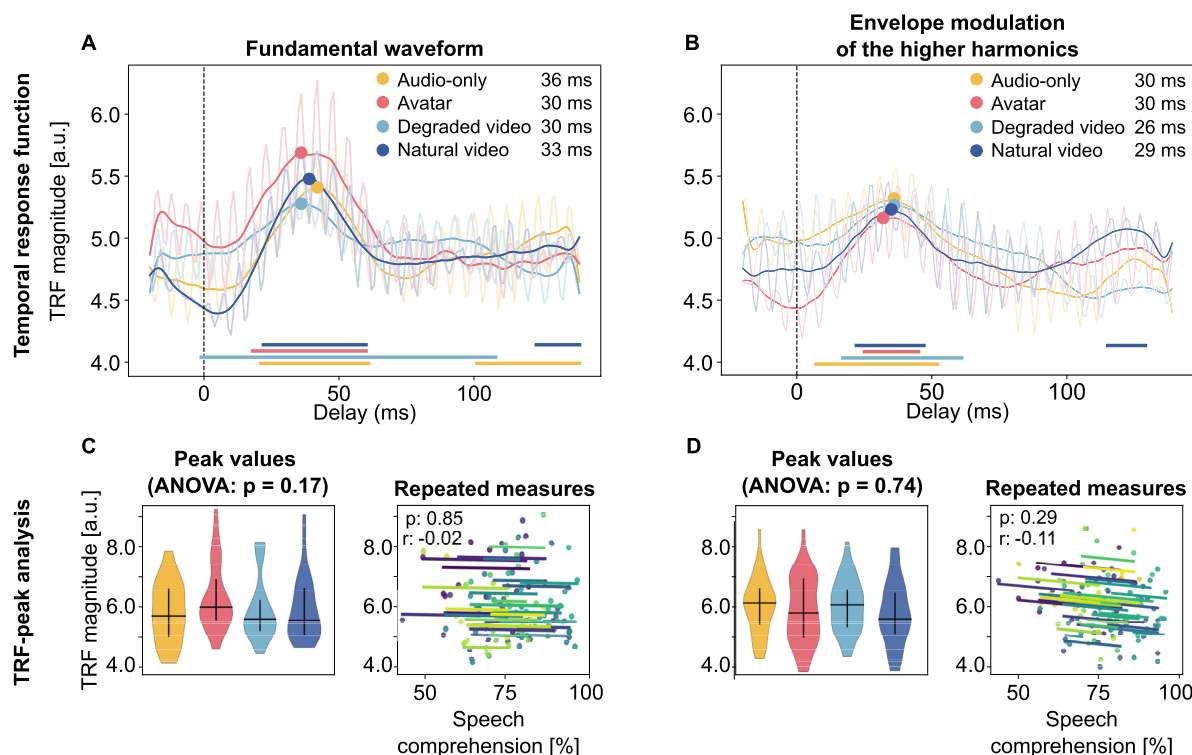


FIGURE 3 | Temporal response functions (TRFs) in the four different audiovisual conditions. (A, B) TRFs for the fundamental waveform (A) and for the envelope modulation of the higher harmonics (B) for all four audiovisual conditions. The transparent lines show the amplitude of the TRFs, and the opaque lines are the low-pass filtered version of the amplitudes. The dots in the plots represent the peaks of the low-pass filtered TRFs. The timing of the peaks is given in the legend. The bars at the bottom indicate at which time lags the TRFs are significant. (C, D) The TRF-peak analysis for both the fundamental waveform (C) and the envelope modulation of the higher harmonics (D). The left violin plot in both (C) and (D) shows the magnitudes of the TRF peaks across participants for each audiovisual condition. The right plot shows the repeated-measures correlation between the magnitude of the TRF peaks and the behavioral scores across the four conditions. Each color represents one participant.

did not obtain a statistically significant result (fundamental waveform: $p=0.85$, Figure 3C, right; envelope modulation: $p=0.29$, Figure 3D, right).

3.3 | Lateralization

In addition to analyzing the source-reconstructed MEG data averaged across the whole auditory cortex, we also investigated lateralization between the left hemisphere (LH) and the right hemisphere (RH). We therefore compared the neural activity at the population-level peak latencies, averaged over all participants. We found a stronger neural response in the right hemisphere, both to the fundamental frequency and to the envelope modulation of the higher harmonics (Figure 4).

Although the difference is visible as a trend across all four audiovisual conditions, it is significant only in the three conditions that include visual speech information (Bonferroni-corrected). The neural response to the fundamental waveform is significantly right lateralized ($p<0.001$) in these three conditions (Figure 4A) (audio-only: $p=1.0$).

The lateralization of the response to the envelope modulation is also significant in all AV conditions (avatar: $p=0.006$, degraded: $p=0.01$, and natural: $p=0.04$). In contrast, the lateralization of the audio-only condition ($p=1.0$) is not significant.

To check whether the strength of lateralization per participant was significantly correlated with participants' speech comprehension, we calculated a lateralization score L for each participant as $L = \frac{M_{rh} - M_{lh}}{M_{rh} + M_{lh}}$, with M_{rh} denoting the magnitude

of the TRF peak in the right hemisphere and M_{lh} that in the left hemisphere. We then computed the repeated-measures correlation between this lateralization score L and the participants' speech comprehension score in each of the four audiovisual conditions. Again, this did not yield a significant effect (fundamental waveform: $p=0.75$; envelope modulation: $p=0.21$).

4 | Discussion

We investigated whether the cortical contribution to the speech-FFR is influenced by seeing a speaker's face. We not only presented the speaker's natural video but also a degraded version of the video and an avatar that resembles the speaker. Background noise was employed to reduce speech comprehension and to make the audiovisual integration apparent. This strategy worked as desired: The participants understood significantly more of the spoken words with the visual signals than without. Moreover, speech comprehension for the natural videos exceeded that for the avatar and the degraded video.

Several recent studies investigated how well avatars can support speech understanding (Munhall et al. 2004; Varano et al. 2022; Shan et al. 2022). The findings reported here are in line with these previous studies, showing that such avatars yield speech comprehension scores that lie between those for the audio-only condition and for the natural video. Although many participants did not notice a difference between the avatar and the natural video (undocumented comments by the participants), the behavioral difference in speech comprehension between the avatar and the natural video was highly significant. This is probably

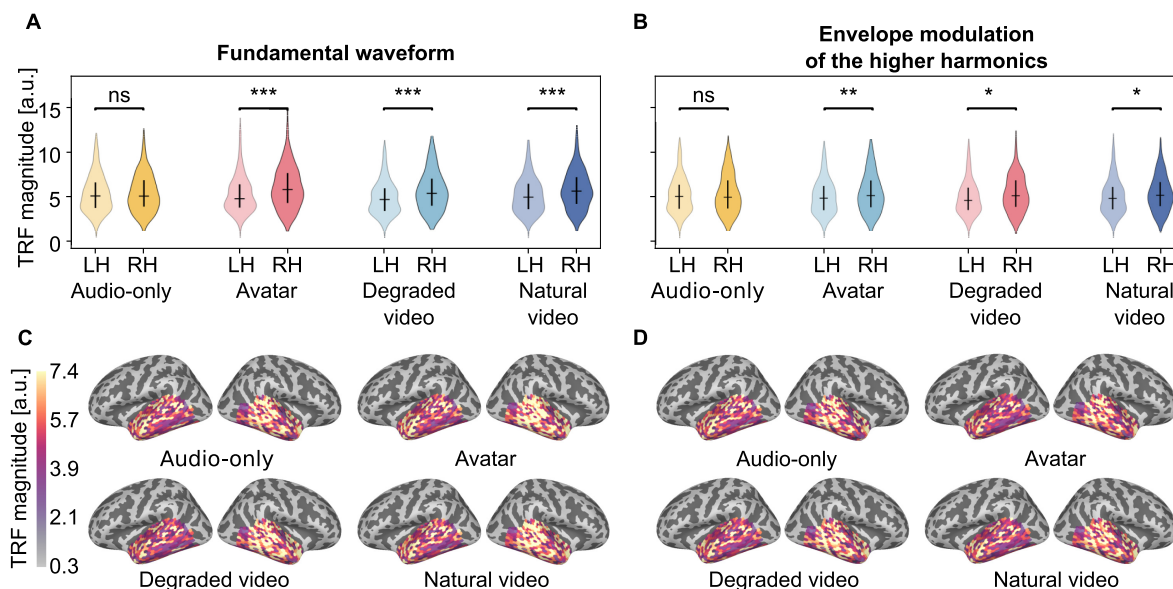


FIGURE 4 | Lateralization of the speech-FFR in the four audiovisual conditions at the lag of the population TRF peak. (A, B) Per condition, the lighter, left violin plot shows the neural response to the predictor variable at the population level in the left hemisphere across voxels. The darker violin plot on the right represents the corresponding activation in the right hemisphere. (A) The response to the fundamental waveform in the audio-only condition shows a nonsignificant tendency of right lateralization. All conditions with video show a strong, significant right lateralization. (B) The neural response to the envelope modulation of the higher harmonics is not significantly lateralized for the audio-only condition. In all conditions containing a video, the neural response exhibits significant right lateralization. (C, D) Neural activation in the auditory cortex in lateral and medial views is presented in each condition. Both the response to the fundamental waveform (C) and the envelope modulation of the higher harmonics (D) appear right lateralized on visual inspection.

not only due to the lips of the avatar not articulating perfectly but also due to the timing between lip movement and speech not being fully accurate. In contrast, the degraded video resulted in the same level of speech comprehension as the avatar. Although the degraded video had, by construction, perfect temporal alignment between lip movement and audio signal, it lacked a realistic appearance.

Regarding neural responses, we found significant peaks around the expected latency of 30 ms, as identified in previous studies (Schüller et al. 2024, 2025). This timing confirmed that we could reliably detect the cortical contribution to the speech-FFR, despite the comparably high background noise level (SNR of -4 dB) (Schüller et al. 2024). However, when comparing the peak magnitudes across the different audio-visual conditions, no statistically significant differences were observed.

To the best of our knowledge, ours is the first study to investigate the cortical contribution to speech-FFR elicited by audiovisual speech, as well as the first to measure this neural response when elicited by an artificially generated avatar. Existing evidence for audiovisual influences on high-frequency auditory processing largely comes from studies using simplified, speech-like stimuli. In particular, Musacchia et al. (2007) examined the effect of audiovisual presentation of the syllable/da/on subcortical responses measured with EEG and reported enhancement of the speech-FFR through the visual signal. At the cortical level, an ECoG study by Karthik et al. (2021) investigated responses to the same type of syllabic stimulus and observed larger neural activity in the high-gamma band in the presence of the visual signal. However, these effects were restricted to a very narrow temporal window closely aligned with stimulus onset and were spatially very limited.

Importantly, both Musacchia et al. (2007) and Karthik et al. (2021) employed audio-only stimuli that did not include a realistic visual depiction of the speaker's face. In contrast, our paradigm presents a static image of the speaker in the audio-only condition, which may reduce the visual contrast between conditions with and without video, and thereby limit the magnitude of the observable audiovisual effects. Furthermore, our experimental setup more closely resembles real-world listening situations, as it involves continuous speech and relatively strong background babble noise. Given that high-gamma audiovisual effects were only marginally detectable in very precise areas within the STG (Karthik et al. 2021), even under highly controlled conditions measured with electrocorticography, it is plausible that such effects are not sufficiently robust to be observed in our more challenging and naturalistic MEG paradigm. In particular, we note that the relatively high level of background noise that we employed reduces the magnitude of the speech-FFR compared to studies with less background noise, which may make putative audiovisual effects harder to detect. Notably, Karthik et al. (2021) reported more robust audiovisual modulations in lower frequency bands, effects that we have also reported separately for our data (Riegel et al. 2026).

Although our principal neural findings are null, they are informative rather than merely inconclusive. First, our design was constructed to maximize sensitivity to any audiovisual

modulation: We followed the principle of inverse effectiveness by embedding speech in strong background noise, we contrasted the still image against three distinct types of visual speech, and we analyzed the response with a well-powered continuous-speech paradigm that has previously resolved robust attentional effects on the same cortical response (Schüller et al. 2023). That no modulation of the peak magnitude emerged under these conditions places an upper bound on the size of any such effect and indicates that, if visual speech shapes the cortical speech-FFR at all, its influence is small relative to the attentional effects previously reported for this response.

Second, our results dissociate the behavioral and the neural levels: The visual signal produced a large, graded improvement in comprehension, yet left the cortical speech-FFR unchanged and unable to account for the between-condition variance in comprehension. This dissociation suggests that the comprehension benefit conferred by seeing a talker's face is not mediated by a strengthening of early, pitch-related cortical encoding but arises at other stages or in other frequency bands, consistent with the low-frequency audiovisual effects that we observe in the same dataset (Riegel et al. 2026).

Third, the result carries a practical implication: The avatar improved comprehension to a level comparable with the degraded video without measurably engaging the cortical speech-FFR, indicating that avatar-based intelligibility enhancement need not rely on faithfully driving early pitch-related encoding.

In addition to analyzing the magnitude of the speech-FFR, we also examined its hemispheric lateralization. Across all conditions, responses showed a clear right-hemispheric bias. Previous audio-only studies have similarly reported right-lateralized cortical responses to the fundamental waveform and the envelope modulation of the higher harmonics of continuous speech (Schüller et al. 2023). Interestingly, in our study, lateralization was significant only for stimuli with visual information. We consider three, not mutually exclusive, accounts of this pattern. First, the right-hemispheric bias was present as a numerical trend in every condition, including the still-image condition, and reached significance only in the three conditions containing visual speech. The still-image response was also the weakest overall, so the simplest account is one of statistical power: The same underlying right lateralization may be present throughout but crosses the significance threshold only where the response, and hence the effect size, is largest. Consistent with this, we would expect the trend in the still-image condition to also reach significance in a paradigm with less background noise.

Second, and more speculatively, a genuinely stronger right lateralization for stimuli with visual input would hint at a minor audiovisual influence on high-gamma encoding that is expressed in the spatial distribution of the response rather than in its overall magnitude. Third, the difference may reflect not audiovisual integration per se, but a difference in engagement: Participants may attend more to conditions containing visual speech than to the static image. This would be in line with Hartmann and Weisz (2019), who reported that attentional enhancement of the high-gamma response occurs specifically in the right auditory cortex, and it would predict exactly the condition-dependent right lateralization we observe.

Because this lateralization effect emerged from an exploratory analysis across several conditions and did not correlate with comprehension, we interpret it cautiously and regard it as a target for confirmatory, adequately powered replication rather than as firm evidence of audiovisual integration.

In summary, we examined the cortical contribution to the speech-FFR elicited by audiovisual speech. The magnitude of this response did not depend on the type of visual stimulus and was also identical when participants watched an artificially-created avatar. Despite substantial variation in the speech-in-noise comprehension scores between the participants and the audiovisual conditions, we found that these could not be explained by the speech-FFR. Our results suggest that audiovisual integration occurs at later cortical stages. Future investigations may further clarify the latencies at which audiovisual integration regarding speech stimuli appears, and whether pitch processing may be involved.

Author Contributions

Jasmin Riegel: conceptualization, investigation, writing – original draft, methodology, validation, visualization, formal analysis, data curation. **Alina Schüller:** methodology, data curation, writing – review and editing. **Mike Thornton:** writing – review and editing, methodology, formal analysis. **Alexander Wißmann:** methodology, formal analysis, writing – review and editing. **Steffen Zeiler:** writing – review and editing, supervision, methodology. **Dorothea Kolossa:** methodology, supervision, writing – review and editing. **Tobias Reichenbach:** conceptualization, funding acquisition, writing – review and editing, project administration, supervision, methodology.

Acknowledgments

This project was supported by the German Science Foundation (DFG, project number 523344822). Open Access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data used in this study are available on Zenodo: <https://zenodo.org/records/18258659>.

Peer Review

For transparency, the peer review documents associated with this article are available at <https://doi.org/10.1111/ejn.70680>.

References

Bourgeois, J., and W. Minker. 2009. "Linearly Constrained Minimum Variance Beamforming." In *Time-Domain Beamforming and Blind Source Separation: Speech Input in the Car Environment*, 27–38. Springer.

Chi, T., P. Ru, and S. A. Shamma. 2005. "Multiresolution Spectrotemporal Analysis of Complex Sounds." *Journal of the Acoustical Society of America* 118: 887–906.

Coffey, S. C., A. M. P. Herholz, S. B. Chepesiuk, S. Baillet, and R. J. Zatorre. 2016. "Cortical Contributions to the Auditory Frequency-Following Response Revealed by MEG." *Nature Communications* 7: 11070.

Coffey, T. N., T. Nicol, T. White-Schwoch, et al. 2019. "Evolving Perspectives on the Sources of the Frequency-Following Response." *Nature Communications* 10: 5036.

Commuri, V., J. P. Kulasingham, and J. Z. Simon. 2023. "Cortical Responses Time-Locked to Continuous Speech in the High-Gamma Band Depend on Selective Attention." *Frontiers in Neuroscience* 17: 1264453.

Crosse, M. J., J. S. Butler, and E. C. Lalor. 2015. "Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions." *Journal of Neuroscience* 35, no. 42: 14195–14204.

Crosse, M. J., G. M. Di Liberto, and E. C. Lalor. 2016. "Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration." *Journal of Neuroscience* 36, no. 38: 9888–9895.

D-ID. 2025. "D-ID: Creative Reality Studio." <https://www.d-id.com/>. Software Used for Creating AI-Generated Avatars.

Etard, O., M. Kegler, C. Braiman, A. E. Forte, and T. Reichenbach. 2019. "Decoding of Selective Attention to Continuous Speech From the Human Auditory Brainstem Response." *NeuroImage* 200: 1–11.

Fischl, B. 2012. "FreeSurfer." *NeuroImage* 62, no. 2: 774–781.

Forte, A. E., O. Etard, and T. Reichenbach. 2017. "The Human Auditory Brainstem Response to Running Speech Reveals a Subcortical Mechanism for Selective Attention." *eLife* 6: e27203.

Gramfort, A., M. Luessi, E. Larson, et al. 2014. "MNE Software for Processing MEG and EEG Data." *NeuroImage* 86: 446–460.

Grant, K. W., and P.-F. Seitz. 2000. "The Use of Visible Speech Cues for Improving Auditory Detection of Spoken Sentences." *Journal of the Acoustical Society of America* 108: 1197–1208.

Guilleminot, P., M. Thornton, E. Varano, and M. Kegler. 2026. "sPy-EEG: A Stimuli-Brain Mapping Toolbox for Electrophysiological Data Analysis." HAL HAL-05453437, <https://hal.science/hal-05453437v1>.

Hartmann, T., and N. Weisz. 2019. "Auditory Cortical Generators of the Frequency Following Response Are Modulated by Intermodal Attention." *NeuroImage* 203: 116185.

Hertrich, I., S. Dietrich, J. Trouvain, A. Moos, and H. Ackermann. 2012. "Magnetic Brain Activity Phase-Locked to the Envelope, the Syllable Onsets, and the Fundamental Frequency of a Perceived Speech Signal." *Psychophysiology* 49: 322–334.

Karthik, G., J. Plass, Z. L. Beltz, et al. 2021. "Visual Speech Differentially Modulates Beta, Theta, and High Gamma Bands in Auditory Cortex." *European Journal of Neuroscience* 54: 7301–7317.

Kulasingham, J. P., C. Brodbeck, A. Presacco, S. E. Kuchinsky, S. Anderson, and J. Z. Simon. 2020. "High Gamma Cortical Processing of Continuous Speech in Younger and Older Listeners." *NeuroImage* 222: 117291.

Lehmann, A., and M. Schönwiesner. 2014. "Selective Attention Modulates Human Auditory Brainstem Responses: Relative Contributions of Frequency and Spatial Cues." *PLoS ONE* 9, no. 1: e85442.

Ma, W. J., X. Zhou, J. J. Ross, a. L. C. Foxe, and L. C. Parra. 2009. "Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space." *PLoS ONE* 4: e4638.

Macaluso, E., N. George, C. S. Dolan, C. Spence, and J. Driver. 2004. "Spatial and Temporal Factors During Processing of Audiovisual Speech: A PET Study." *NeuroImage* 21: 725–732.

Mauch, M., and S. Dixon. 2014. "PYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions." IEEE International Conference on Acoustics, Speech and Signal Processing.

- Miles, K., T. Beechey, V. Best, and J. Buchholz. 2022. "Measuring Speech Intelligibility and Hearing-Aid Benefit Using Everyday Conversational Sentences in Real-World Environments." *Frontiers in Neuroscience* 16: 789565.
- Munhall, K. G., J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson. 2004. "Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception." *Psychological Science* 15, no. 2: 133–137.
- Musacchia, G., M. Sams, T. Nicol, and N. Kraus. 2006. "Seeing Speech Affects Acoustic Information Processing in the Human Brainstem." *Experimental Brain Research* 168: 1–10.
- Musacchia, G., M. Sams, E. Skoe, and N. Kraus. 2007. "Musicians Have Enhanced Subcortical Auditory and Audiovisual Processing of Speech and Music." *Proceedings of the National Academy of Sciences of the United States of America* 104: 15894–15898.
- Musacchia, G., D. Strait, and N. Kraus. 2008. "Relationships Between Behavior, Brainstem and Cortical Encoding of Seen and Heard Speech in Musicians and Non-Musicians." *Hearing Research* 241, no. 1–2: 34–42.
- Nirme, J., B. Sahle'n, V. Lyberg Åhlander, J. Brännström, and M. Haake. 2020. "Audio-Visual Speech Comprehension in Noise With Real and Virtual Speakers." *Speech Communication* 116: 44–55.
- Pilling, M. 2009. "Auditory Event-Related Potentials (ERPs) in Audiovisual Speech Perception." *Journal of Speech, Language, and Hearing Research* 52, no. 4: 1073–1081.
- Riegel, J., A. Schüller, C. Jehn, A. Wißmann, S. Zeiler, and T. Reichenbach. 2026. "Talking Avatars Can Differentially Modulate Cortical Speech Tracking in the High and in the Low Delta Band." bioRxiv. <https://doi.org/10.64898/2026.01.07.695461>.
- Riegel, J., A. Schüller, and T. Reichenbach. 2024. "No Evidence of Musical Training Influencing the Cortical Contribution to the Speech-Frequency-Following Response and Its Modulation Through Selective Attention." *eNeuro* 11: ENEURO.0127-24.2024.
- Ross, L. A., D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe. 2007. "Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments." *Cerebral Cortex* 17, no. 5: 1147–1153.
- Russo, N., T. Nicol, G. Musacchia, and N. Kraus. 2004. "Brainstem Responses to Speech Syllables." *Clinical Neurophysiology* 115, no. 9: 2021–2030.
- Schilling, A., R. Tomasello, A. Z. Henningsen-Schomers, et al. 2020. "Analysis of Continuous Neuronal Activity Evoked by Natural Speech With Computational Corpus Linguistics Methods." *Language, Cognition and Neuroscience* 36, no. 2: 167–186.
- Schüller, A., A. Mücke, J. Riegel, and T. Reichenbach. 2025. "The Impact of Selective Attention and Musical Training on the Cortical Speech Tracking in the Delta and Theta Frequency Bands." *Journal of Cognitive Neuroscience* 37, no. 2: 464–481.
- Schüller, A., A. Schilling, P. Krauss, S. Rampp, and T. Reichenbach. 2023. "Attentional Modulation of the Cortical Contribution to the Frequency-Following Response Evoked by Continuous Speech." *Journal of Neuroscience* 43, no. 44: 7429–7440.
- Schüller, A., A. Schilling, P. Krauss, and T. Reichenbach. 2024. "The Early Subcortical Response at the Fundamental Frequency of Speech Is Temporally Separated From Later Cortical Contributions." *Journal of Cognitive Neuroscience* 36, no. 3: 475–491.
- Shan, T., C. X. Wenner, C. Xu, Z. Duan, and R. K. Maddox. 2022. "Speech-in-Noise Comprehension Is Improved When Viewing a Deep-Neural-Network-Generated Talking Face." *Trends in Hearing* 26: 23312165221136934.
- Skoe, E., and N. Kraus. 2010. "Auditory Brainstem Response to Complex Sounds: A Tutorial." *Ear and Hearing* 31, no. 3: 302–324.
- Stoll, T. J., N. D. Vandjelovic, M. J. Polonenko, N. R. S. Li, A. K. C. Lee, and R. K. Maddox. 2025. "The Auditory Brainstem Response to Natural Speech Is Not Affected by Selective Attention." *PLoS Biology* 23: e3003407.
- Sumby, W. H., and I. Pollack. 1954. "Visual Contribution to Speech Intelligibility in Noise." *Journal of the Acoustical Society of America* 26, no. 6: 212.
- Thézé, R., M. A. Gadiri, L. Albert, A. Provost, A.-L. Giraud, and P. Mégevand. 2020. "Animated Virtual Characters to Explore Audio-Visual Speech in Controlled and Naturalistic Environments." *Scientific Reports* 10, no. 1: 15540.
- Van Atteveldt, N. M., V. C. Blau, L. Blomert, and R. Goebel. 2010. "fMR-Adaptation Indicates Selectivity to Audiovisual Content Congruency in Distributed Clusters in Human Superior Temporal Cortex." *BMC Neuroscience* 11: 11.
- Van Wassenhove, V., a. D. P. Grant, and D. Poeppel. 2005. "Visual Speech Speeds Up the Neural Processing of Auditory Speech." *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 4: 1181–1186.
- Varano, E., P. Guilleminot, and T. Reichenbach. 2023. "AVbook, a High-Frame-Rate Corpus of Narrative Audiovisual Speech for Investigating Multimodal Speech Perception." *Journal of the Acoustical Society of America* 153, no. 5: 3130.
- Varano, E., M. Thornton, D. Kolossa, S. Zeiler, and T. Reichenbach. 2025. "Delta-Band Cortical Speech Tracking Predicts Audiovisual Speech-in-Noise Benefit From Natural and Simplified Visual Cues." *NeuroImage* 325: 121654.
- Varano, E., K. Vougioukas, P. Ma, S. Petridis, M. Pantic, and T. Reichenbach. 2022. "Speech-Driven Facial Animations Improve Speech-in-Noise Comprehension of Humans." *Frontiers in Neuroscience* 15: 781196.
- Wagener, K., V. Kuehnel, and B. Kollmeier. 1999. "Entwicklung und Evaluation eines Satztests für die deutsche Sprache i: Design des Oldenburger Satztests." *Zeitschrift für Audiologie* 38, no. 1: 4–15.
- Whiteford, L. S., M. C. Baltzell, M. Chiu, et al. 2025. "Large-Scale Multi-Site Study Shows No Association Between Musical Training and Early Auditory Neural Sound Encoding." *Nature Communications* 16: 7152.