Neurobiology of Aging 134 (2024) 165–180

Contents lists available at ScienceDirect

Neurobiology of Aging

journal homepage: www.elsevier.com/locate/neuaging.org

Reliability and generalizability of neural speech tracking in younger and older adults

Ryan A. Panela [a,b], Francesca Copelli [a,b], Björn Herrmann [a,b,*]

[a] Rotman Research Institute, Baycrest Academy for Research and Education, M6A 2E1 North York, ON, Canada
[b] Department of Psychology, University of Toronto, M5S 1A1 Toronto, ON, Canada

ARTICLE INFO

Keywords:
Electroencephalography
Speech encoding
Story listening
Aging
Temporal response function
Test-retest reliability

ABSTRACT

Neural tracking of spoken speech is considered a potential clinical biomarker for speech-processing difficulties, but the reliability of neural speech tracking is unclear. Here, younger and older adults listened to stories in two sessions while electroencephalography was recorded to investigate the reliability and generalizability of neural speech tracking. Speech tracking amplitude was larger for older than younger adults, consistent with an age-related loss of inhibition. The reliability of neural speech tracking was moderate (ICC ~0.5–0.75) and tended to be higher for older adults. However, reliability was lower for speech tracking than for neural responses to noise bursts (ICC >0.8), which we used as a benchmark for maximum reliability. Neural speech tracking generalized moderately across different stories (ICC ~0.5–0.6), which appeared greatest for audiobook-like stories spoken by the same person. Hence, a variety of stories could possibly be used for clinical assessments. Overall, the current data are important for developing a biomarker of speech processing but suggest that further work is needed to increase the reliability to meet clinical standards.

1. Introduction

Understanding how speech is processed in the brain is important for clinical applications, such as age-related hearing loss, dementia, and stroke (Olichney et al., 2011; Schneider et al., 2002; Tyler and Marslen-Wilson, 2008). Traditional approaches to understanding speech processing have relied on word or sentence materials presented in random, disconnected order (Friederici et al., 1993; Herrmann et al., 2011; Marinkovic et al., 2003). Such materials lack a topical thread, are not very interesting to the listener, and thus may not capture how speech is processed in real life (Hamilton and Huth, 2020). Over the past decade, experimental and analytic approaches have substantially advanced to capture speech processing for more naturalistic, continuous speech, such as spoken stories (Brodbeck and Simon, 2020; Crosse et al., 2016; Ding and Simon, 2012; Hamilton and Huth, 2020; Herrmann and Johnsrude, 2020).

Possibly the most used approach to investigate the extent to which continuous, spoken speech is encoded in the brain is the temporal response function and related encoding/decoding models that, for example, predict electroencephalography/magnetoencephalography (EEG/MEG) activity from speech features (Brodbeck and Simon, 2020; Crosse et al., 2016; Crosse et al., 2021; Lalor and Foxe, 2010). Quantifying how well the brain tracks a specific speech feature provides an estimate of how well the feature is encoded. Especially the neural tracking of the low-frequency (<10 Hz) acoustic amplitude envelope of speech has been extensively studied, for example, in the context of selective attention (Brodbeck and Simon, 2020; Emily et al., 2022; Fiedler et al., 2019) and speech masked by background sound (Schmitt et al., 2022; Synigal et al., 2023; Yasmin et al., 2023). Investigating the neural tracking of the speech envelope is a valuable approach, because the envelope is important for speech intelligibility (Shannon et al., 1995) and envelope tracking predicts speech intelligibility to some extent (Ding et al., 2014; Lesenfants et al., 2019; Vanthornhout et al., 2018; see discussion in Gillis et al., 2022). Moreover, calculating the envelope is easy and available in various toolboxes (Crosse et al., 2016; Crosse et al., 2021), whereas other recent approaches are more complex (Broderick et al., 2018; Broderick et al., 2021).

Neural speech-tracking approaches are increasingly used to understand clinical phenomena, such as age-related hearing loss (Decruy et al., 2020b; Presacco et al., 2019; Schmitt et al., 2022) and stroke-related or dementia-related aphasias (Dial et al., 2021; Kries et al., 2023). For example, aging and hearing loss are associated with

* Correspondence to: Rotman Research Institute, Baycrest, 3560 Bathurst St, North York, ON M6A 2E1, Canada.
E-mail address: bherrmann@research.baycrest.org (B. Herrmann).

https://doi.org/10.1016/j.neurobiolaging.2023.11.007
Received 27 July 2023; Received in revised form 9 November 2023; Accepted 16 November 2023
Available online 21 November 2023

enhanced neural tracking of the speech envelope (Decruy et al., 2020b; Presacco et al., 2016b), which may be the result of a loss of cortical inhibition resulting from periphery deafferentation (Caspary et al., 2008; Herrmann and Butler, 2021a). Given the success of the speech-tracking approach, researchers have suggested that the neural-tracking response could be used as a biomarker for speech-processing pathologies (Gillis et al., 2022; Palana et al., 2022; Schmitt et al., 2022). Yet, for the neural-tracking response to be useful as a biomarker it must be reliable, but its reliability is currently unclear.

Reliability is investigated by conducting the same test procedures twice (Lockhart, 1998). The intra-class correlation (ICC; Koo and Li, 2016; McGraw and Wong, 1996; Shrout and Fleiss, 1979) – a metric that captures both the degree of correlation and agreement between two measurements – may be used to assess reliability, although some previous EEG/MEG research has used Spearman's or Pearson's correlation instead, which only assesses correlation (Cabral-Calderin and Henry, 2022; McEvoy et al., 2000; Tervaniemi et al., 1999b). Neural responses to tones, noises, and vowels have moderate to high reliability (Bidelman et al., 2018; Legget et al., 2017; Tervaniemi et al., 1999b), whereas little is known about the reliability of the neural tracking of continuous speech. Moreover, in clinical contexts, individuals would ideally listen to unique stories – for example, between appointments – to avoid influences of prior knowledge on the neural response. Hence, it would be beneficial if the neural-tracking response generalized across different stories.

The current study provides an extensive account of the reliability and generalizability of the neural-tracking response across different speakers, stories, and noise conditions in younger and older adults. The data are critical for research aiming to use the neural-tracking response as a biomarker for the assessment of auditory or cognitive impairments.

2. Methods and materials

2.1. Participants

Twenty-two younger adults (median: 22 years; range: 19–34 years) and 22 older adults (median: 70.5 years; range: 56–77 years) participated in the current study. Sixteen younger adults identified as female or women, five as male, and one as non-binary. Fourteen older adults identified as female and eight as male. Twelve younger and 20 older adults identified as native English speaker, whereas the other participants were highly proficient English speakers. Participants who indicated having a non-English first language, nevertheless, grew up in English-speaking countries (mostly Canada) and have been speaking English since early childhood (<5 years of age). Participants reported having no neurological disease, except for one older adult who indicated having trigeminal neuralgia (not affecting participation and results). Participants reported having normal hearing abilities. Participants did not wear hearing aids nor were they prescribed one.

Each participant took part in two sessions, separated by at least one week (median: 13 days; range: 7–59 days). Participants gave written informed consent prior to the experiment and were paid $7.50 CAD per half-hour for their participation. The study was conducted in accordance with the Declaration of Helsinki, the Canadian Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans (TCPS2–2014), and was approved by the Research Ethics Board of the Rotman Research Institute at Baycrest Academy for Research and Education.

2.2. Sound environment and stimulus presentation

Data collection was carried out in a sound-attenuating booth. Sounds were presented via Sennheiser (HD 25-SP II) headphones and a RME Fireface 400 external sound card. Stimulation was run using Psychtoolbox in MATLAB (v3.0.14; MathWorks Inc.) on a Lenovo T480 laptop with Microsoft Windows 7. Visual stimulation was projected into the sound booth via screen mirroring. All sounds were presented at about 75 dB SPL.

2.3. Hearing assessment

For each participant, audiometric thresholds were estimated for pure tones at frequencies of 0.25, 0.5, 1, 2, 4, and 8 kHz (Fig. 1A). Mean thresholds averaged across 0.25, 0.5, 1, 2, and 4 kHz were higher for older compared to younger adults ($t_{42} = 5.804$, $p = 7.6 \bullet 10^{-7}$, $d = 1.75$; Fig. 1). Although these thresholds are mostly clinically 'normal' for age according to the ISO-7029 standard (https://www.iso.org/standard/42916.html), elevated thresholds are consistent with the presence of mild-to-moderate hearing loss in the current sample of older adults, as would be expected (Herrmann et al., 2018, 2022; Moore, 2007; Plack, 2014; Presacco et al., 2016b).

2.4. Experimental procedures for noise-burst stimulation

The current study is concerned with the reliability of the neural-tracking response during story listening. To obtain a "benchmark" against which to compare neural-tracking responses during speech listening, participants passively listened to 132 repetitions of a 0.1-s white-noise burst (0.01 s fade-in and 0.01 s fade-out) in a separate ~3.5 min block of stimulation. The noise burst was presented with a mean onset-to-onset interval of 1.5 s (randomly selected between 1.2 and 1.8 s). We expected high reliability for responses to the noise burst. Noise-related reliability is expected to provide an upper reliability bound for story-related neural responses because noise bursts are spectrally broad, eliciting broad activity across auditory cortex.

2.5. Experimental procedures for story listening

Across the two sessions, participants listened to six stories from the story-telling podcast The Moth (https://themoth.org/). The Moth podcast features spoken stories about human experiences and life events. The stories are intended to create an absorbing and enjoyable experience for the listener. The Moth stories mirror speech in everyday life and, unlike an audiobook, include disfluencies, filler-words, sentence fragments, corrections, unintentional pauses, and more flexible grammar (Bortfeld et al., 2001; Tree, 1995). In the current study, we used the following six stories, each of which was approximately 7 min in duration: *The Loudest Whisper* by Devan Sandiford, *Speech Writers Lament* by Karen Duffin, *Lego Crimes* by Micaela Blei, *Do The Dishes And Leave* by Mitch Donaberger, *Priceless Mangos* by Saya Shamdasani, and *Teacher*

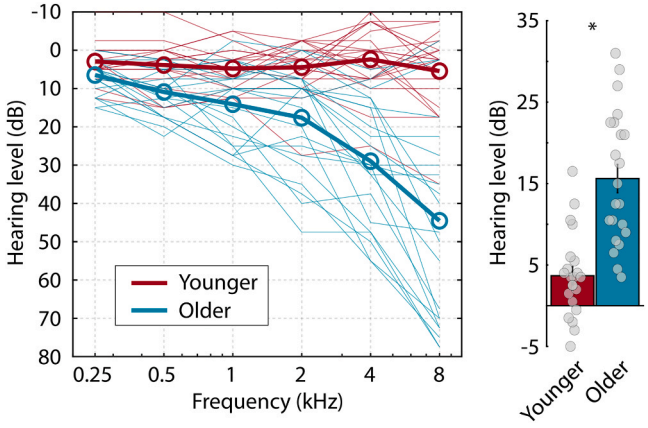

Fig. 1. Pure-tone audiometric thresholds. Left: Pure-tone audiometric thresholds for all frequencies. The thin lines reflect data from each individual participant. Thick lines reflect the mean across participants. Right: Pure-tone average hearing threshold (mean across 0.25, 0.5, 1, 2, and 4 kHz). Dots reflect the pure-tone average threshold for individual participants. *p ≤ 0.05.

Talent Show by Tim Manley. Participants also listened to two moderately engaging and easily comprehensible Storybook stories made for listeners at any level (Irsik et al., 2022a; Mathiesen et al., 2023). These stories were adapted from two print Storybooks, *Wave* by D.M. Ouellet and *Alibi* by Kristin Butcher, aimed at reluctant readers. The stories are high interest but have a simple vocabulary and a linear plot. They are somewhat similar to audiobooks, which have been used frequently for neural speech tracking (Broderick et al., 2018; Broderick et al., 2022; Lesenfants et al., 2019; Presacco et al., 2016a; b). The two stories were recorded in-lab by a male native English speaker and each story was about 10 min in duration.

In each of the two sessions, participants listened to four The Moth stories and one Storybook story. Two of The Moth stories were presented under clear conditions (i.e., without background noise), whereas 12-talker babble at a signal-to-noise ratio (SNR) of 9 dB was added to the other two The Moth stories (Bilger, 1984; Bilger et al., 1984; Wilson et al., 2012). Speech in background babble at 9 dB SNR is still highly intelligible (~90% of words; Irsik et al., 2022b), but may require more attention and effort by the listener than clear speech (Herrmann and Johnsrude, 2020; Yasmin et al., 2023). Moderate levels of background noise have also been shown to increase the neural tracking response (Yasmin et al., 2023). Storybook stories were presented under clear conditions. All stories were normalized to the same overall root-mean-square amplitude.

Stories were separated into three categories, henceforth referred to as SM stories (as in Same in both sessions, The Moth), DM stories (as in Different in both sessions, The Moth), and DB stories (as in Different in both sessions, Storybook; Fig. 2A). SM stories were repeated in session 1 and session 2 to investigate test-retest reliability in a strict sense. We used The Moth stories by Devan Sandiford and Karen Duffin. SM stories were presented in the first two blocks of each session to reduce variance that may be associated with the session duration. The story by Devan Sandiford was presented under clear conditions in both sessions, whereas the story by Karen Duffin was presented in background babble at 9 dB SNR in both sessions. The application of background babble was not counterbalanced across stories/sessions because this would have interfered with examining reliability. Story order was counter-balanced across the two sessions, such that if the story by Devan Sandiford was presented in the first block in session 1, it was presented in the second block of session 2, and vice versa (Fig. 2A).

DM stories were not repeated in session 1 and session 2. Two different stories from The Moth were presented in each session. DM stories enable the investigation of the generalizability of the neural-tracking response across sessions, stories, and speakers. In addition, by comparing the neural-tracking responses across stories within a session, we can investigate the generalizability across stories and speakers within a session. For DM stories, the stories by Micaela Blei, Mitch Donaberger, Saya Shamdasani, and Tim Manley were used. Two DM stories were always presented in blocks 3 and 4 of an experimental session (Fig. 2A). One of the DM stories per session was presented under clear conditions, whereas the other DM story was presented with added background babble at a 9-dB SNR. Story order and speech-clarity conditions (clear, background noise) were counter-balanced across sessions and participants (Fig. 2A). Counterbalancing enables us to make clearer inferences, compared to SM stories, about the effect of background babble on the neural-tracking response and its generalizability.

DB stories were Storybook stories and mirrored more closely an audiobook than The Moth stories. One DB story was presented in each experimental session, always in block 5. The specific Storybook story presented in a session was counter-balanced across participants (Fig. 2A). DB stories enable us to investigate the generalizability of neural speech tracking across sessions, but reveal the effect of maintaining the same speaker. Moreover, the analysis of responses to DB stories provides insights into the reliability of neural tracking under more standardized conditions, where recordings are made in-lab by the same speaker. Such stories are more easily created in high numbers if needed for clinical assessments.

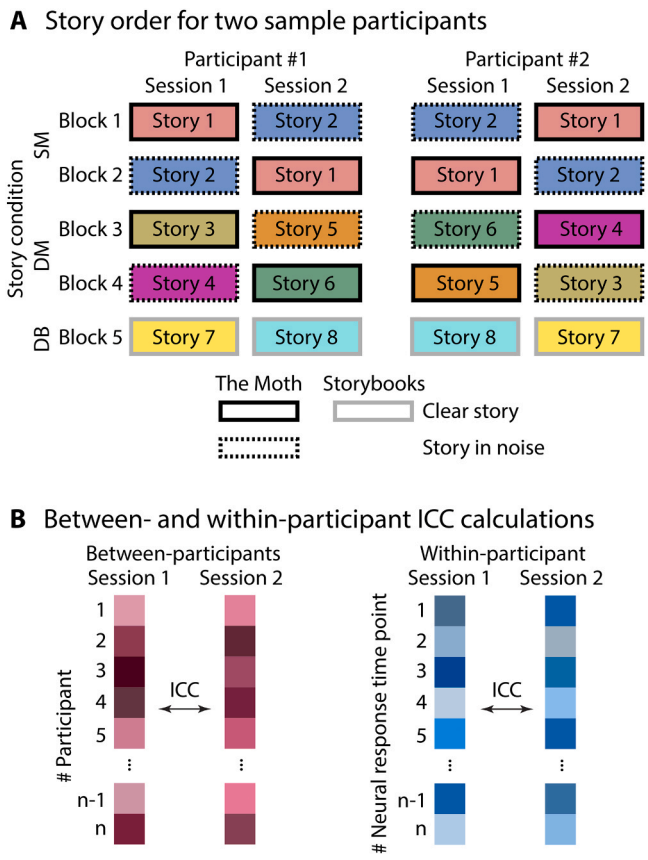

Fig. 2. Experimental design and measures A: Schematic of experimental design. Story order for two sample participants. Colors code the eight different stories and edge lines specify the speech-clarity conditions and story origin (solid – clear, dashed – babble noise; black – The Moth podcast, gray – Storybook). SM stories repeated across the two sessions (SM - same, The Moth), whereas DM stories did not repeat across sessions and were spoken by different speakers (DM – different, The Moth). SM and DM stories were sampled from The Moth podcast. DB stories mirror audiobook recordings. DB stories did not repeat across sessions (DB – different, Storybook), but were spoken by the same speaker. **B:** Schematic of the two types of intra-class correlation (ICC) calculations: between-participants ICC, and within-participant ICC. For between-participants ICC, observations are the participants, whereas, for within-participant ICC, observations are the neural response time points of a participant. Different colors schematically represent different response magnitudes.

After each story, participants answered ten comprehension questions about the story. Each comprehension question comprised four response options (chance level = 25%). Participants also rated the degree to which they were absorbed by and enjoyed the story, using four and two items from previous work, respectively (Herrmann and Johnsrude, 2020; Kuijpers et al., 2014; absorption: "I felt absorbed by the story.", "When I finished listening, I was surprised to see that time had gone by so fast.", "I felt connected with the main character(s) of the story.", "I could imagine what the world in which the story took place looked like."; enjoyment: "I listened to the story with great interest.", "I thought it was an exciting story."). Participants further rated the degree to which they had to invest effort to comprehend the story using two items ("I had to invest effort to understand what was said.", "Understanding the speaker was hard."). Absorption, enjoyment, and effort were rated on a 7-point scale ranging from 1 (completely disagree) to 7 (completely agree).

2.6. Analysis of behavioral story-listening data

Absorption, enjoyment, and effort rating scores were linearly rescaled such that they range from 0 to 1 in order to facilitate interpretation and similarity to comprehension scores. For each participant, scores and ratings were separately averaged across story comprehension, effort, enjoyment, and absorption items.

Responses to DM stories were analyzed to assess the effects of Speech Clarity and Age Group on the four measures. To this end, a repeated measures analysis of variance (rmANOVA) was calculated, separately for story comprehension, effort, enjoyment, and absorption. Session (session 1, session 2) and Speech Clarity (clear, noise) were within-participant factors, whereas Age Group (younger, older) was a between-participants factor. Session was included to explicitly examine whether participants changed their judgement criteria across multiple lab visits, although we did not expect an effect of session due to counterbalancing story order and speech-clarity conditions (Fig. 2A).

Responses to SM stories were analyzed to investigate the effects of Session and Age Group on the four measures. A rmANOVA was calculated separately for each behavioral measure (comprehension, effort, enjoyment, and absorption) using Session (session 1, session 2) as a within-participant factor and Age Group (younger, older) as a between-participants factor. Analyses were separately conducted for the clear story and the story in noise, because each speech-clarity condition for SM stories was associated with a specific story to facilitate reliability analyses as outlined above.

Responses to DB stories were analyzed to investigate the effects of Session and Age Group on the four measures for the audiobook-like stories. A rmANOVA was calculated separately for each behavioral measure (comprehension, effort, enjoyment, and absorption) using Session (session 1, session 2) as a within-participant factor and Age Group (younger, older) as a between-participants factor.

We also calculated the difference between enjoyment and absorption between The Moth stories (clear, session 1, SM & DM stories) and Storybook stories (clear, session 1, DB stories) using a rmANOVA with Story Condition (The Moth, Storybook) as a within-participant factor and Age Group (younger, older) as a between-participants factor.

2.7. Electroencephalography recordings and preprocessing

Electroencephalographical signals were recorded from 16 scalp electrodes (Ag/Ag–Cl-electrodes; 10–20 placement) and the left and right mastoids using a BioSemi system (Amsterdam, The Netherlands). The sampling frequency was 1024 Hz with an online low-pass filter of 208 Hz. Electrodes were referenced online to a monopolar reference feedback loop connecting a driven passive sensor and a common-mode-sense (CMS) active sensor, both located posteriorly on the scalp.

Offline analysis was conducted using MATLAB software. An elliptic filter was used to suppress power at the 60-Hz line frequency. Data were re-referenced by averaging the signal from the left and right mastoids and subtracting the average separately from each of the 16 channels. Re-referencing to the averaged mastoids was calculated to gain high signal-to-noise ratio for auditory responses at fronto-central-parietal electrodes, at the expense of losing information about auditory polarity reversal at the mastoids (Herrmann et al., 2013a). Data were filtered with a 0.7-Hz high-pass filter (length: 2449 samples, Hann window) and a 22-Hz low-pass filter (length: 211 samples, Kaiser window).

For the one block during which noise bursts were presented, EEG data were divided into epochs ranging from –1 to 2 s time-locked to noise onset and down-sampled to 512 Hz. Independent components analysis (runica method, Makeig et al., 1995; logistic infomax algorithm, Bell and Sejnowski, 1995; Fieldtrip implementation Oostenveld et al., 2011) was used to identify and remove components related to blinks and horizontal eye movements. Epochs for which the signal range exceeded 100 μV in any of the EEG electrodes were excluded from analysis.

For the five blocks during which stories were presented, EEG data were segmented into time series time-locked to story onset and down-sampled to 512 Hz. Independent components analysis was used to remove signal components reflecting blinks and eye movement (Bell and Sejnowski, 1995; Makeig, Oostenveld et al., 1995, 2011). Additional artifacts were removed after the independent components analysis by setting the voltage for segments in which the EEG amplitude varied more than 80 μV within a 0.2-s period in any channel to 0 μV (cf. Cohen and Parra, 2016; Dmochowski et al., 2014; Dmochowski et al., 2012; Irsik et al., 2022b; Yasmin et al., 2023). Data were low-pass filtered at 10 Hz (251 points, Kaiser window) because neural signals in the low-frequency range are most sensitive to acoustic features (Di Liberto et al., 2015; Yasmin et al., 2023; Zuk et al., 2021).

For one younger participant, EEG recordings for one SM story and, for another younger participant, EEG recordings for one DM story could not be analyzed because of a technical error during recording. EEG data from these participants were excluded for analyses that required the availability of data from both sessions but were otherwise included.

2.8. Analysis of event-related potentials to noise bursts

Time courses were averaged across trials, focusing on –0.15–0.5 s epochs time-locked to noise onset. Response time courses were baseline-corrected by subtracting the mean amplitude within the –0.15–0 s time window from the amplitude at each time point. Neural responses were also averaged across a fronto-central electrode cluster (F3, Fz, F4, C3, Cz, C4) known to be sensitive to neural activity originating from auditory cortex (Herrmann et al., 2018; Irsik et al., 2021; Näätänen and Picton, 1987; Picton et al., 2003).

Data analysis focused on the amplitude in the 0.03–0.06 s and the 0.08–0.12 s time windows that correspond to the P1 and N1 components of the event-related potentials (Herrmann et al., 2016; Herrmann et al., 2018; Näätänen and Picton, 1987; Picton et al., 1974). Mean amplitudes within a time window were used as dependent measure in a rmANOVA. Predictors were the within-participant factor Session (session 1, session 2) and the between-participants factor Age Group (younger, older).

Reliability of neural responses was calculated in two ways, focusing on between-participants reliability and within-participant reliability. For between-participants reliability (Fig. 2B, left), intra-class correlation (ICC) was calculated as the two-way mixed effects, absolute agreement, single rater (ICC(2,1); Koo and Li, 2016; McGraw and Wong, 1996; Shrout and Fleiss, 1979). Specifically, amplitudes were averaged, and ICC was calculated for 0.05 s sliding windows centered on each time point, separately for each age group. That is, the mean time-window response in session 1 and session 2 were the measures, while participants were the observations in this analysis. This resulted in one ICC time course per age group. The 0.05 s time window was chosen to account for some degree of variability in response latencies across participants (results are qualitatively similar for slightly shorter or longer time windows). We also calculated ICC values for the P1 and N1 time windows. ICC values indicate excellent, good, moderate, or poor reliability if they are greater than 0.9, between 0.75 and 0.9, between 0.5 and 0.75, or below 0.5, respectively (Koo and Li, 2016). In order to obtain an estimate of variability for between-participants ICC, we calculated the 95% confidence intervals using bootstrapping, such that between-participants ICC was calculated for 1000 resampled datasets with replacement (Efron, 1979; Koo and Li, 2016; Wasserman and Bockenholt, 1989).

For within-participant reliability (Fig. 2B, right), we calculated ICC values (ICC(2,1); Koo and Li, 2016; Shrout and Fleiss, 1979), separately for each participant as the reliability of the time course ranging from 0 to 0.4 s across sessions. That is, response time courses in session 1 and session 2 of the same participant were the two measures, while individual time points were the observations in this analysis. The 0–0.4 s time window was chosen because it comprises the major deflections of the event-related potential (Fig. 3). This resulted in one ICC value for each participant. Age-group differences in within-participant ICC were

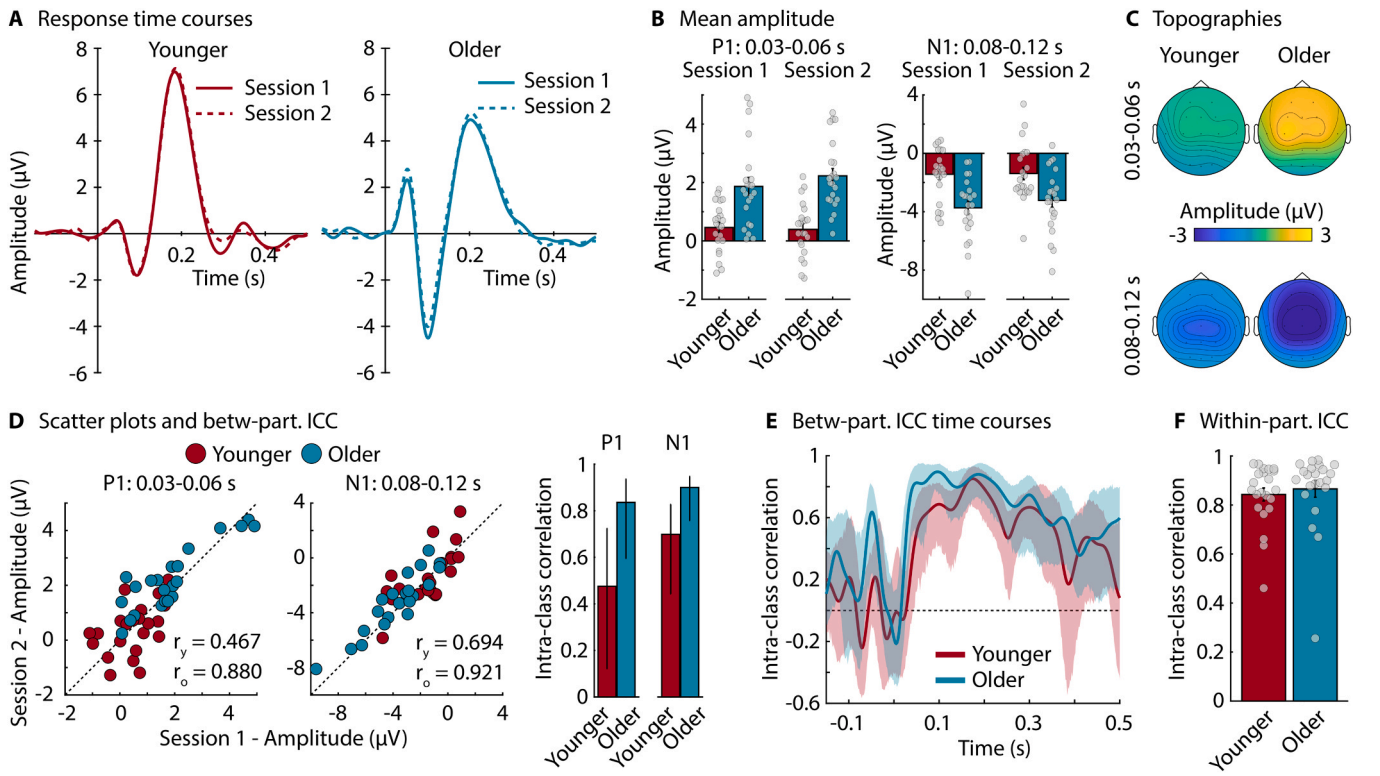

Fig. 3. Neural responses and reliability for noise bursts. A: Time courses of neural responses to noise bursts and topographical distributions for the 0.03–0.06 s and the 0.08–0.12 s time windows. **B:** Mean neural responses for the 0.03–0.06 s and the 0.08–0.12 s time windows. Dots reflect data from individual participants. **C:** Topographical distributions (averaged across sessions). **D:** Scatter plots for neural responses in the 0.03–0.06 s and the 0.08–0.12 s time windows. Pearson correlations are provided within each plot. The subscripts y and o indicate correlations for younger and older adults, respectively. Bar graphs reflect the between-participants ICC values for the P1 and N1 time windows. Error bars reflect the 95% confidence intervals from bootstrapping. **E:** Between-participants ICC time courses for younger and older adults. ICC was calculated for 0.05 s time windows centered on each time point. Shaded areas reflect the 95% confidence intervals from bootstrapping. **F:** Within-participant ICC, considering the time course from 0 s to 0.4 s. Error bar reflects the standard error of the mean. Dots reflect data from individual participants.

assessed using an independent-samples t-test. Sample time courses for the two sessions and different participants as well as corresponding within-participant ICC values are provided in Fig. S1 (supplementary materials).

2.9. Calculation of neural tracking response and EEG reconstruction accuracy during story listening

We used a forward model based on the linear temporal response function (TRF; Crosse et al., 2016; Crosse et al., 2021) to quantify the relationship between the amplitude-onset envelope of a story and EEG activity. To this end, a cochleogram was calculated for each story using a simple auditory-periphery model with 30 auditory filters (McDermott and Simoncelli, 2011). The resulting amplitude envelope for each auditory filter was compressed by 0.6 to simulate inner ear compression (McDermott and Simoncelli, 2011). Such a computationally simple peripheral model has been shown to be sufficient, as compared to complex, more realistic models, for envelope-tracking approaches (Biesmans et al., 2017). Amplitude envelopes were averaged across auditory filters and low-pass filtered at 40-Hz filter (Butterworth filter). To obtain the amplitude-onset envelope, we calculated the first derivative and set all negative values to zero (Fiedler et al., 2017; Fiedler et al., 2019; Hertrich et al., 2012; Yasmin et al., 2023). The onset-envelope was down-sampled to match the sampling of the EEG data.

For the analysis, 100 30-s data snippets (Crosse et al., 2016; Crosse et al., 2021) were extracted randomly from the EEG data and corresponding onset-envelope per story and session. Each of the 100 EEG and onset-envelope snippets were held out once as a test dataset, while the remaining non-overlapping EEG and onset-envelope snippets were used as training datasets. That is, for each training dataset, linear regression with ridge regularization was used to map the onset-envelope onto the EEG activity to obtain a TRF model for lags ranging from 0 to 0.4 s (Crosse et al., 2016; Crosse et al., 2021; Hoerl and Kennard, 1970). The ridge regularization parameter lambda (λ), which prevents overfitting, was set to 10 based on previous work (Fiedler et al., 2017; Fiedler et al., 2019; Yasmin et al., 2023). Note that cross-validation to estimate λ yielded qualitative similar results compared to a λ of 10. Pre-selection of λ based on previous work avoids extremely low and high λ on some cross-validation iterations and avoids substantially longer computational time that may be unfeasible in clinical contexts. Pre-selection of λ may also be required clinically because assessments times need to be short, which limit the amount of data recorded (Crosse et al., 2021). The TRF model calculated for the training data was then used to predict the EEG signal for the held-out test dataset. The Pearson correlation between the predicted and the observed EEG data of the test dataset was used as a measure of EEG reconstruction accuracy (Crosse et al., 2016; Crosse et al., 2021). Model estimation and reconstruction accuracy were calculated separately for each of the 100 data snippets per story and session, and reconstruction accuracies were averaged across the 100 snippets. To investigate the neural-tracking response directly in addition to the reconstruction accuracy, we also calculated TRFs for each training dataset for a broader set of lags, ranging from −0.15–0.5 s, to enable similar analyses as for traditional event-related potentials (Yasmin et al., 2023). TRFs were averaged across the 100 training datasets.

Data analyses of response magnitude differences and reliability/generalizability of neural responses focused on a fronto-central electrode cluster (F3, Fz, F4, C3, Cz, C4) known to be most sensitive to auditory cortex activity (Irsik et al., 2021; Näätänen and Picton, 1987;

Picton et al., 2003). TRFs and reconstruction accuracies were averaged across the electrodes of this fronto-central electrode cluster prior to further analysis.

2.10. Analysis of the effects of speech clarity and age group on neural response magnitude

Before investigations into the reliability of neural responses, we first aimed to examine differences in response amplitude related to speech clarity and age group. To this end, we focused analyses on TRFs and reconstruction accuracy for DM Stories, for which speech-clarity conditions were counter-balanced across stories and sessions. TRFs and reconstruction accuracies were averaged across stories and sessions, separately for clear stories and stories in babble noise. Analyses of the TRF focused on P1 and N1 amplitudes as the average in the 0.03–0.06 s and the 0.9–0.13 s time windows, respectively. A slightly later time window was chosen for the N1 in response to speech compared to the N1 response to the noise bursts, because our and previous data indicate later N1 latencies for speech (Fiedler et al., 2019; Hertrich et al., 2012; Yasmin et al., 2023; possibly due to more graded acoustic onsets). A rmANOVA was calculated separately for TRF P1, TRF N1, and reconstruction accuracy, using the within-participant factor Speech Clarity (clear, noise) and the between-participants factor Age Group (younger, older).

2.11. Analysis of reliability of neural responses during story listening

Analyses of the reliability of the neural responses during story listening focused on SM stories, for which the identical stories (one clear, one with added babble) were presented in both sessions. Reliability analyses were calculated in two ways, similar to the neural responses to noise bursts, focusing on between-participants reliability and within-participant reliability (Fig. 2B).

For between-participants reliability, ICC was calculated as the two-way, mixed effects, absolute agreement (ICC(2,1); Koo and Li, 2016; Shrout and Fleiss, 1979). TRF amplitudes were averaged and ICC was calculated for 0.05 s sliding windows centered on each time point, separately for the clear story and the story with added background babble, and separately for the two age groups. The mean time-window response in session 1 and session 2 were the measures in this ICC analysis, while participants were the observations (Fig. 2B, left). This resulted in one ICC time course per speech-clarity condition and age group. ICC values were also calculated separately for TRF P1 and TRF N1 amplitudes. To obtain an estimate of variability for between-participants ICC, we calculated the 95% confidence intervals using bootstrapping (Efron, 1979; Koo and Li, 2016; Wasserman and Bockenholt, 1989).

For within-participant reliability (Fig. 2B, right), the TRF time courses (ranging from 0 to 0.4 s) of the two sessions were used to calculate ICC, separately for each participant and speech-clarity condition. Response time courses in the two sessions were the measures in this ICC analysis, while individual time points were the observations. Individual time courses for sample participants are shown in the supplementary materials (Fig. S1). Age-group differences in within-participant reliability (ICC) were assessed using an independent-samples t-test, separately for the clear story and the story in background babble (speech-clarity conditions were not directly compared because we did not counter-balance speech clarity across stories to reduce unwanted variance).

We also calculated between-participants reliability using ICC for reconstruction accuracy, separately for both speech-clarity conditions and both age groups. For this ICC analysis, the correlation values calculated using the leave-one-out procedure for session 1 and session 2, described above, were the measures, and participants were the observations (Fig. 2B, left). This resulted in one ICC value per speech-clarity condition and age group. Again, the 95% confidence intervals for between-participants ICC values were calculated using bootstrapping.

2.12. Generalizability of neural-speech tracking responses

Generalizability of story-related neural responses was investigated within sessions and between sessions. For the investigation of generalizability across stories and speakers within a session, we focused only on session 1 to avoid influences of the repetition of SM stories. ICC was calculated separately for clear stories and stories with background babble (i.e., using one SM and one DM story each), and separately for the two age groups. To this end, TRF amplitudes were averaged, and between-participants ICC was calculated for 0.05-s sliding windows centered on each time point. Between-participants ICC was also calculated for TRF P1 and TRF N1 amplitudes as well as for reconstruction accuracy, separately for each speech-clarity condition and age group. The 95% confidence intervals were calculated for between-participants ICC using bootstrapping. We further calculated within-participant ICC, for which the TRF time courses (ranging from 0 to 0.4 s) of the two stories of the same speech-clarity condition (clear, noise) were used to calculate ICC, separately for each participant. A rmANOVA was calculated using the within-participant ICC as the dependent variable, and Speech Clarity (clear, noise) and Age Group (younger, older) as independent variables.

For the investigation of the generalizability across sessions, stories, and speakers, ICC of neural responses to DM stories was calculated similarly as the reliability calculation for neural responses to SM stories. That is, TRF amplitudes were averaged and ICC was calculated for 0.05-s sliding windows centered on each time point, separately for clear stories and stories with added background babble, and separately for the two age groups. Between-participants ICC was also calculated for TRF P1 and TRF N1 amplitudes as well as for reconstruction accuracy, separately for both speech-clarity conditions and both age groups. The 95% confidence intervals were calculated for between-participants ICCs. We further calculated within-participant ICC, for which the TRF time courses (ranging from 0 to 0.4 s) of the responses to stories in the two sessions were used to calculate ICC separately for each participant and speech-clarity condition. A rmANOVA was calculated using the within-participant ICC as the dependent variable, and Speech Clarity (clear, noise) and Age Group (younger, older) as independent variables.

For the investigation of the generalizability of the neural-tracking response across sessions and speakers for audiobook-like stories, DB stories were used. Between-participants ICC time courses were calculated by averaging TRF amplitudes and computing ICC for 0.05-s sliding windows centered on each time point, separately for the two age groups. Between-participants ICC was also calculated for TRF P1 and TRF N1 amplitudes as well as for reconstruction accuracy, separately for both age groups. The 95% confidence intervals were calculated for between-participants ICCs. Within-participant ICC was further calculated (using the TRF time courses from 0 to 0.4 s) for each participant. An independent samples t-test was used to compare within-participant ICC between age groups.

2.13. Comparisons among reliability and generalizability

To compare whether ICC differed between different assessment types, we calculated a rmANOVA using within-participant ICC values as the dependent measure and Assessment Type as within-participant factor with five levels: ERP reliability, TRF reliability (SM stories across sessions), TRF generalizability within a session (SM stories vs DM stories), TRF generalizability across sessions (DM stories), and TRF generalizability across session for stories spoken by the same person (DB stories). Age Group (younger, older) was included as a between-participants factor. For this analysis, we focused on clear stories because the noise bursts for the ERPs were also presented under clear conditions.

2.14. Statistical analyses

All data analyses described were carried out using MATLAB (MathWorks) and JASP software (JASP, 2022; version 0.16.4.0). Effect sizes for rmANOVAs and t-tests are reported using omega squared (ω^2) and Cohen's d (d), respectively. Significant effects or interactions in rmANOVAs were resolved using post hoc tests with Holm's method for multi-comparison corrections (Holm, 1979).

3. Results

3.1. Neural responses to noise bursts

Responses to noise bursts in the 0.03–0.06 s and in the 0.08–0.12 s time windows were larger for older compared to younger adults (effect of Age Group: 0.03–0.06 s: $F_{1,42} = 26.616$, $p = 6.3 \bullet 10^{-6}$, $\omega^2 = 0.230$; 0.08–0.12 s: $F_{1,42} = 12.592$, $p = 9.7 \bullet 10^{-4}$, $\omega^2 = 0.119$; Fig. 2), but there was no difference between sessions (effect of Session: ps > 0.1) and no Session × Age Group interaction (ps > 0.05).

Between-participants reliability of the neural responses was moderate to good for younger adults (ICC ~0.6–0.8) and good for older adults

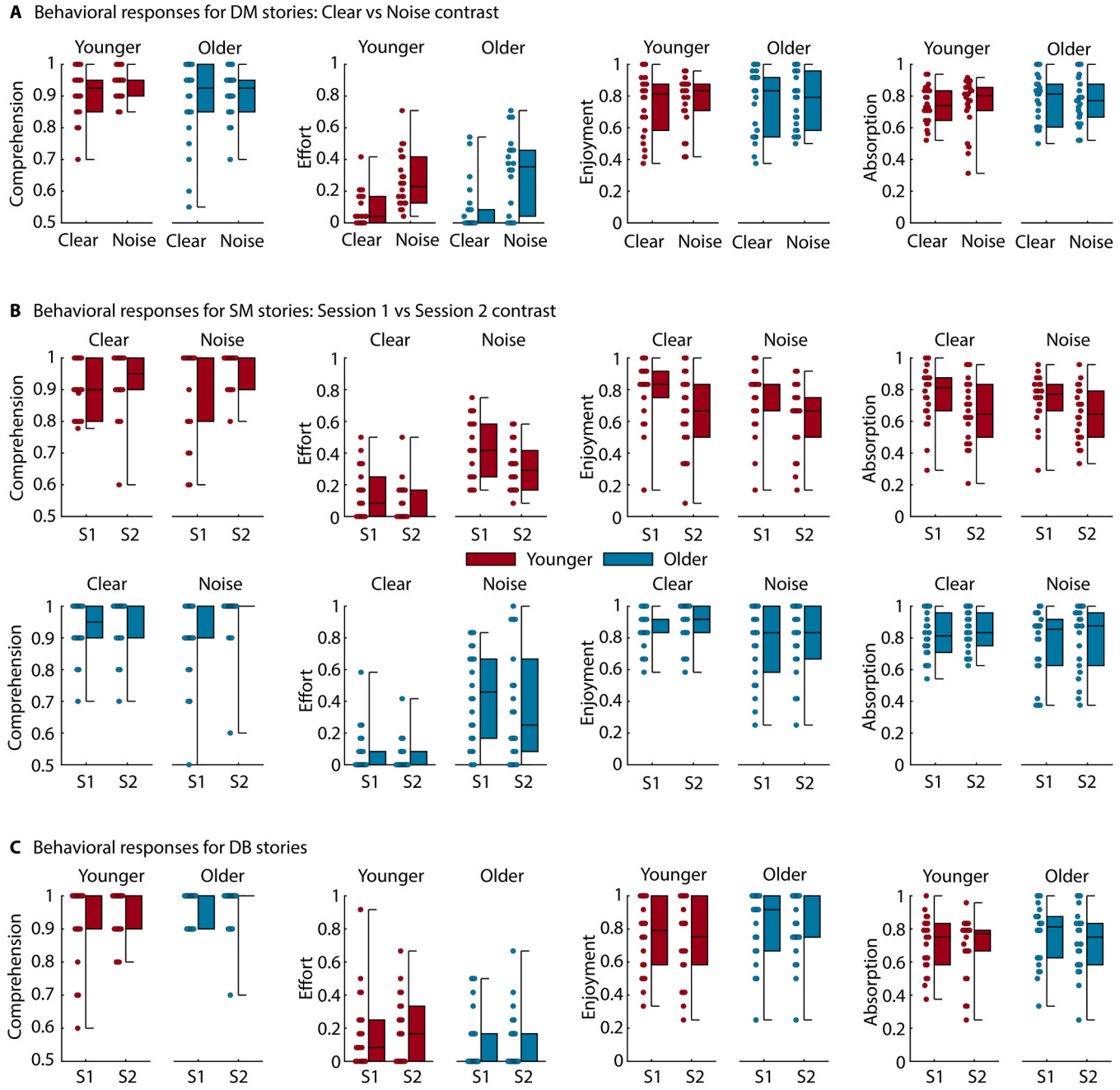

Fig. 4. Behavioral data for story listening. A: Story comprehension, listening effort, enjoyment, and absorption scores for DM stories (different The Moth stories in the two sessions), which enable comparisons between clear stories and stories in babble. **B:** Story comprehension, listening effort, enjoyment, and absorption scores for SM stories (same The Moth stories in both sessions), which enable comparison between sessions. First row in panel B shows data from younger adults, whereas the second row shows the data for older adults. **C:** Story comprehension, listening effort, enjoyment, and absorption scores for DB stories (different Storybook stories in the two sessions). Box plots are shown as well as data points from individual participants. S1 – session 1, S2 – session 2.

(ICC >0.8) in the 0.05–0.25 s time window and for the P1 and N1 amplitude (Koo and Li, 2016; Fig. 3D and E). Mean within-participant reliability, considering the time course from 0 s to 0.4 s, was good for both younger (ICC = 0.84) and older adults (ICC = 0.87), and did not differ between age groups ($t_{42} = 0.526$, $p = 0.601$, $d = 0.159$; Fig. 3F).

3.2. Behavioral data for story listening

We first focused on DM stories, that is, those that were not repeated and for which speech-clarity conditions were counterbalanced across sessions and stories. The behavioral data for DM stories enabled us to investigate the effects of Speech Clarity and Age Group on story comprehension, listening effort, enjoyment, and absorption as well as whether participants changed how they rated the metrics from session 1 to session 2 (Fig. 4A). Listening effort was higher for stories in background noise compared to clear speech (effect of Speech Clarity: $F_{1,42} = 46.933$, $p = 2.4 \bullet 10^{-8}$, $\omega^2 = 0.249$). No other effects nor interactions, for any of the measures, were significant (ps > 0.05). Hence, enjoyment, absorption, and comprehension were not significantly affected by background noise (Herrmann and Johnsrude, 2020), there were no evident differences between age groups (Mathiesen et al., 2023), and there was no overall tendency to rate the measures differently in the second compared to the first session (no effects of Session).

Analyses for SM stories (i.e., identical stories in both sessions) aimed to investigate the effects of Session and Age Group on story comprehension, listening effort, enjoyment, and absorption, separately for the clear story and the story in babble (Fig. 4B). Story comprehension did not differ between sessions under clear conditions ($p > 0.2$), but comprehension was higher in session 2 than session 1 under babble conditions ($F_{1,42} = 11.618$, $p = 0.001$, $\omega^2 = 0.089$). There were no effects of Age Group or interactions for comprehension (ps > 0.2). Effort decreased from session 1 to session 2 for the clear story ($F_{1,42} = 5.029$, $p = 0.030$, $\omega^2 = 0.022$) and the story in noise ($F_{1,42} = 4.689$, $p = 0.036$, $\omega^2 = 0.026$). There were again no effects of Age Group or interactions (ps > 0.1). For enjoyment, the rmANOVA revealed a Session × Age Group interaction for the clear story ($F_{1,42} = 14.948$, $p = 3.8 \bullet 10^{-4}$, $\omega^2 = 0.058$) and the story in noise ($F_{1,42} = 7.701$, $p = 0.008$, $\omega^2 = 0.022$), such that enjoyment was lower in session 2 than session 1 for younger adults (clear: $t_{42} = 4.353$, $p_{Holm} = 4.2 \bullet 10^{-4}$, $d = 0.806$; noise: $t_{42} = 2.970$, $p_{Holm} = 0.029$, $d = 0.486$), but not for older adults (for both $p_{Holm} > 0.5$). Results for absorption mirrored those for enjoyment. A Session × Age Group interaction was observed for the clear story ($F_{1,42} = 7.955$, $p = 0.007$, $\omega^2 = 0.029$) and the story in noise ($F_{1,42} = 9.050$, $p = 0.004$, $\omega^2 = 0.030$), showing lower absorption for session 2 than 1 for younger (clear: $t_{42} = 3.229$, $p_{Holm} = 0.012$, $d = 0.587$; noise: $t_{42} = 3.031$, $p_{Holm} = 0.025$, $d = 0.525$), but not for older adults (for both $p_{Holm} > 0.5$). Enjoyment and absorption were also higher for older compared to younger adults for the clear story (effect of Age Group: enjoyment: $F_{1,42} = 9.046$, $p = 0.004$, $\omega^2 = 0.086$; absorption: $F_{1,42} = 6.739$, $p = 0.013$, $\omega^2 = 0.063$), but not the story in noise (ps > 0.05).

For DB stories, there were no differences between sessions nor between age groups for any of the measures (ps > 0.05). Moreover, Storybook stories (DB stories) appeared to be as enjoyable and absorbing as The Moth stories (SM & DM stories; ps > 0.1; Fig. 4C).

In sum, the behavioral data show that speech in babble noise increases listening effort, but comprehension is higher and listening effort reduced when individuals listen to the same story again a week or more later than when listening to it for the first time. Enjoyment and absorption also decreased with story repetition, but this was only the case for younger adults. Older adults appeared to similarly enjoy and be absorbed by the stories in both sessions.

3.3. Age- and noise-related increases in neural-tracking response during story listening

Prior to analyses of neural-tracking reliability, we investigated the degree to which age group and speech clarity affect the TRF amplitude and reconstruction accuracy (Fig. 5). These analyses focused on DM stories for which speech-clarity conditions were counter-balanced across stories and sessions.

For the TRF analysis (averaged across stories and sessions) in the 0.03–0.06 s time window, amplitudes were larger for older compared to younger adults (effect of Age Group: $F_{1,42} = 43.841$, $p = 5.1 \bullet 10^{-8}$, $\omega^2 = 0.333$) and smaller for stories in background noise compared to clear stories (effect of Speech Clarity; $F_{1,42} = 15.221$, $p = 3.4 \bullet 10^{-4}$, $\omega^2 = 0.106$). The Speech Clarity × Age Group interaction was not significant ($F_{1,42} = 1.786$, $p = 0.189$, $\omega^2 = 0.006$; Fig. 5C).

For the 0.09–0.13 s time window, TRF amplitudes were larger (i.e., more negative) for stories in babble compared to clear stories (effect of Speech Clarity: $F_{1,42} = 104.036$, $p = 6.2 \bullet 10^{-13}$, $\omega^2 = 0.381$; Fig. 5A). There was no effect of Age Group ($F_{1,42} = 1.498$, $p = 0.228$, $\omega^2 = 0.006$) nor a Speech Clarity × Age Group interaction ($F_{1,42} = 2.888$, $p = 0.097$, $\omega^2 = 0.011$; Fig. 5C).

The rmANOVA for EEG reconstruction accuracy revealed larger accuracies for older compared to younger adults ($F_{1,42} = 14.027$, $p = 5.4 \bullet 10^{-4}$, $\omega^2 = 0.132$). The effect of Speech Clarity ($F_{1,42} < 0.001$, $p = 0.982$, $\omega^2 < 0.001$) and the Speech Clarity × Age Group interaction ($F_{1,42} = 0.005$, $p = 0.942$, $\omega^2 < 0.001$) were not significant.

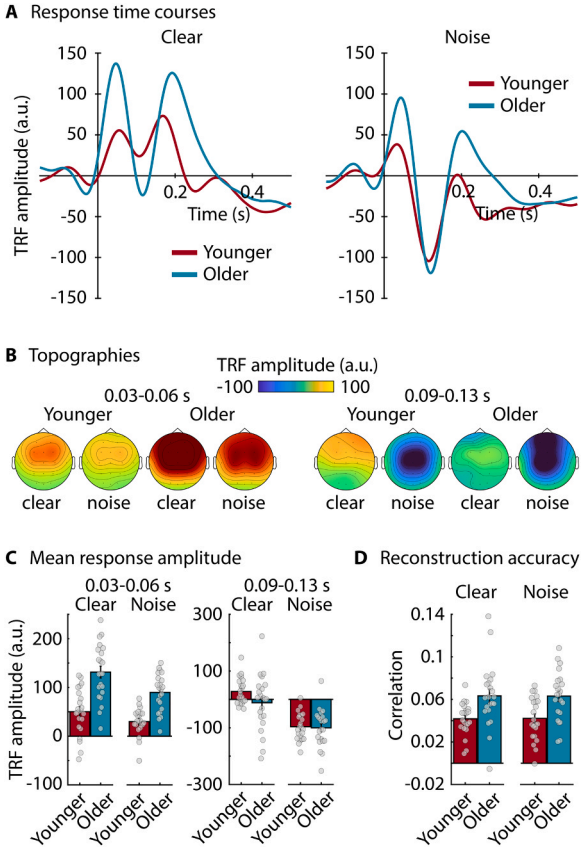

Fig. 5. Effects of Speech Clarity and Age Group on TRF amplitude and EEG reconstruction accuracy using DM stories. A: Temporal response functions (TRF) for each speech-clarity condition (clear, noise) and age group (younger, older). **B:** Topographies for the mean TRF amplitude in the 0.03–0.06 s and the 0.09–0.13 s time windows. **C:** Bar graphs show the mean TRF amplitude in the 0.03–0.06 s and the 0.09–0.13 s time windows. Dots reflect data from individual participants. **D:** EEG reconstruction accuracy for each speech-clarity condition (clear, noise) and age group (younger, older). Dots reflect data from individual participants.

3.4. Moderate reliability of neural-tracking response during story listening

The analyses reported in this section focus on SM stories and explore reliability in a strict sense (i.e., identical stories presented in both sessions). TRF amplitudes in the 0.03–0.06 s time window were larger for older compared to younger adults (clear: $F_{1,42} = 31.290$, $p = 1.6 \bullet 10^{-6}$, $\omega^2 = 0.265$; noise: $F_{1,42} = 22.764$, $p = 2.3 \bullet 10^{-5}$, $\omega^2 = 0.206$; Fig. 6), mirroring the results for DM stories (Fig. 5). There were no effects of Session, nor Session × Age Group interactions, nor any effects for the 0.09–0.13 s time window (ps > 0.15). EEG reconstruction accuracy was also greater in older compared to younger adults (clear: $F_{1,42} = 15.826$, $p = 2.8 \bullet 10^{-4}$, $\omega^2 = 0.150$; noise: $F_{1,42} = 11.250$, $p = 0.002$, $\omega^2 = 0.109$; Fig. 6E), but there were no effects of Session nor Session × Age Group interactions (ps > 0.4).

Fig. 7 shows reliability data for the neural responses elicited by SM stories. Time courses in Fig. A and B (middle) show the between-participants ICC for TRF amplitudes using 0.05-s sliding windows. ICC for P1 and N1 amplitudes are shown as well. Peak ICC at around 0.1–0.2 s was about 0.72 for younger and older adults for the clear story, but below 0.6 for most other time points. Peak ICC for the story in babble noise was about 0.75 for older and 0.5 for younger adults. ICC values between 0.5 and 0.75 are indicative of moderate reliability (Koo and Li, 2016).

Within-participant ICC – that is, the agreement of the 0–0.4-s time courses between the two sessions – was 0.58 (younger) and 0.65 (older) for the clear story and 0.45 (younger) and 0.62 (older) for the story in noise. Within-participant ICC was greater for older compared to younger adults for the story in noise ($t_{41} = 2.058$, $p = 0.046$, $d = 0.628$), but there was no age-group difference for the clear story ($t_{41} = 1.051$, $p = 0.299$, $d = 0.321$; Fig. 7A,B right).

Between-participants ICC for reconstruction accuracy of the clear story was 0.43 for younger and 0.58 for older adults. For the story in babble noise, between-participants ICC for reconstruction accuracy was 0.3 for younger and 0.64 for older adults (Fig. 7C).

3.5. Across-story generalizability: Within-sessions focus

Generalizability of neural responses across stories within session 1 was investigated by calculating ICC across SM stories and DM stories (TRFs and mean responses for DM stories are shown in Fig. 5 and Fig. S2 [supplementary materials]). Between-participants ICC for the TRF amplitude was about 0.6 at ~0.1 s where it peaked for clear stories and stories in noise (moderate; Koo and Li, 2016), although ICC for stories in noise for the younger adults was lower (<0.5; Fig. 8A, middle row).

Within-participant ICC for clear stories was 0.53 for younger and 0.54 for older adults ($t_{40} = 0.087$, $p = 0.931$, $d = 0.027$). Within-participant ICC for stories in babble was 0.46 and 0.61 for younger and older adults, respectively ($t_{40} = 1.839$, $p = 0.073$, $d = 0.568$; Fig. 8A, bottom row).

Between-participants ICC for the EEG reconstruction accuracy for younger adults was about 0.7 and 0.25 for clear stories and stories in noise, respectively. For older adults, between-participants ICC was about 0.55 and 0.75 for clear stories and stories in noise, respectively (Fig. 8A, bottom row).

3.6. Across-story generalizability: Between-sessions focus

DM stories were used to investigate the generalizability of neural responses across stories and sessions. That is, the same reliability analyses (using ICC) as for SM stories were calculated for DM stories. Time courses of between-participants ICC for TRF amplitudes are shown in Fig. 8B (middle row), showing peak ICC values of about 0.5–0.6

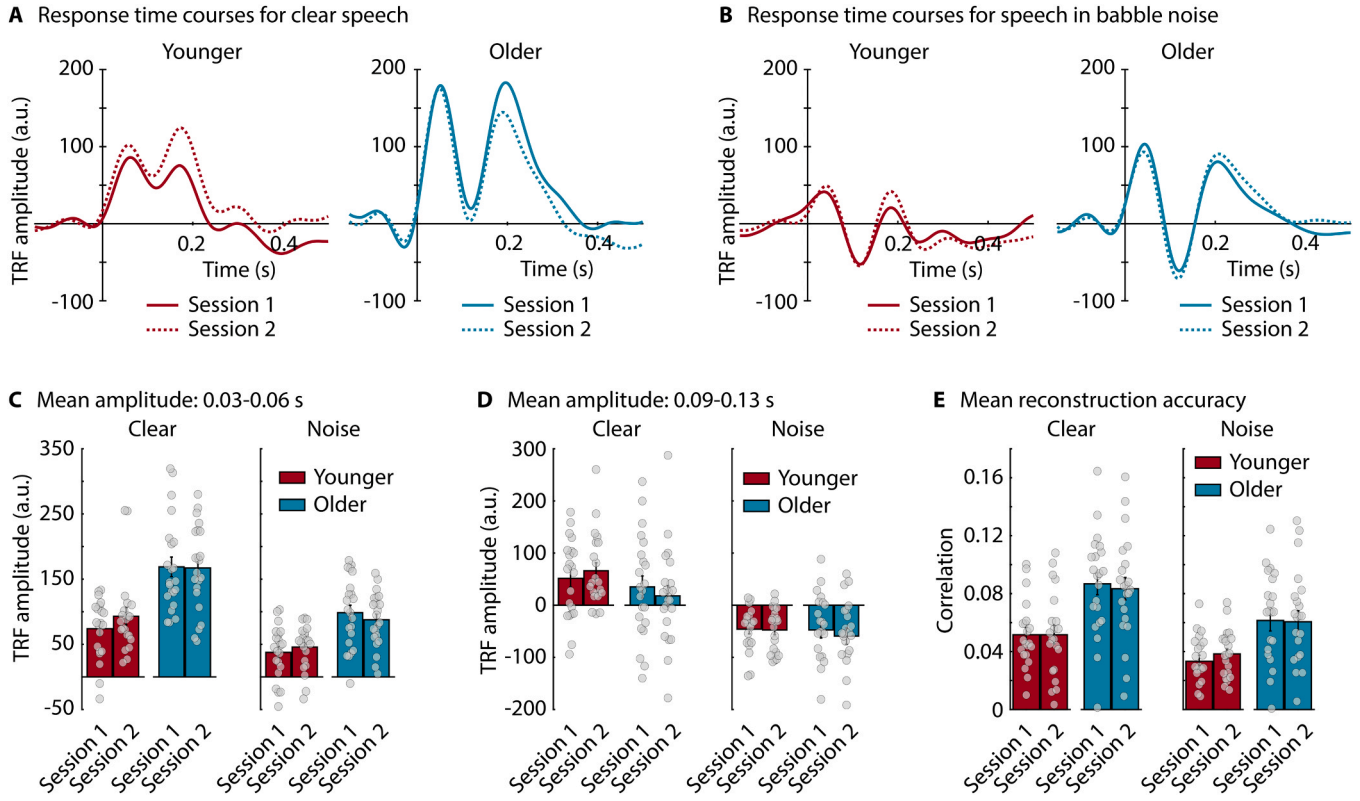

Fig. 6. Neural responses in session 1 and 2 for SM stories. A: Temporal response functions (TRFs) for sessions and age groups for the clear story. **B:** TRFs for sessions and age groups for the story in babble. **C:** Mean TRF amplitude in the 0.03–0.06 s time window. Bars reflect the mean across participants and gray dots data from individual participants. Error bars reflect the standard error of the mean. **D:** Same as in panel C for the TRF amplitude in the 0.09–0.13 s time window. **E:** Same as in panel C for EEG reconstruction accuracy.

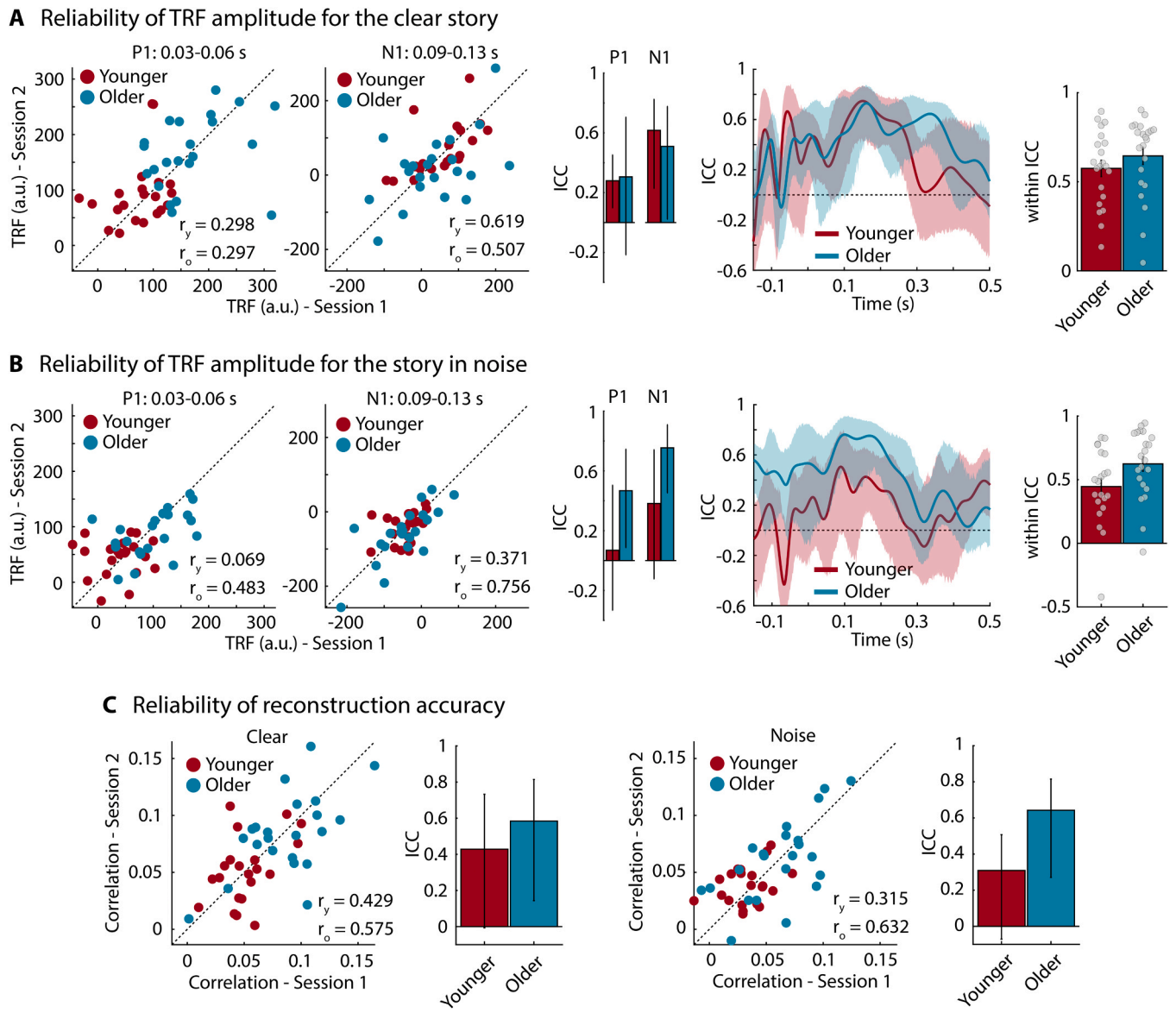

Fig. 7. Reliability of neural responses for SM stories. A: Reliability for TRF amplitude for the clear story. Left: Scatter plots for TRF amplitude in the 0.03–0.06 s (P1) and the 0.09–0.13 s (N1) time windows. Pearson correlations are provided. The subscripts y and o indicate correlations for younger and older adults, respectively. Middle-left: Bar graphs of between-participants ICC values for the P1 and N1 time windows. Error bars reflect the 95% confidence intervals from bootstrapping. Middle-right: Between-participants ICC time courses for both age groups (younger, older). ICC was calculated for 0.05 s time windows centered on each time point. Shaded areas reflect the 95% confidence intervals. Right: Within-participant ICC, considering the time course from 0 to 0.4 s. Bars reflect the mean and dots reflect data from individual participants. Error bars are the standard error of the mean. **B:** Same as in panel A for the story in babble noise. **C:** Reliability for EEG reconstruction accuracy for the clear story (left) and the story in noise (right). Scatter plots and between-participants ICC are shown. Error bars reflect the 95% confidence intervals.

(moderate) at around 0.1 s, although ICC for clear stories for younger adults was lower (<0.5).

The within-participant ICC was 0.48 for the clear story and the story in noise for younger adults (i.e., low reliability; Koo and Li, 2016). For older adults, within-participant ICC was 0.53 and 0.6 for the clear story and the story in noise, respectively. There were no significant differences between speech-clarity conditions or age groups (ps > 0.15; Fig. 8B, bottom row).

Between-participants ICC for the EEG reconstruction accuracy was about 0.6 for older adults under clear and noise conditions, whereas, for younger adults, it was about 0.15 and 0.6 for clear stories and stories in noise, respectively (Fig. 8B, bottom row).

3.7. Across-story generalizability: Between-sessions, same speaker focus

DB stories were used to investigate generalizability across sessions for stories that mirror an audiobook and were spoken by the same speaker. Neural responses are shown in Fig. 9. The rmANOVA for the 0.03–0.06 s time window revealed larger TRF amplitudes for older compared to younger adults ($F_{1,42} = 38.877$, $p = 1.8 \bullet 10^{-7}$, $\omega^2 = 0.306$; Fig. 9B), whereas the effect of Session and the Session × Age Group interaction were not significant (ps > 0.5). The rmANOVA for the 0.09–0.13 s time window revealed no effects (ps > 0.25). The rmANOVA for reconstruction accuracy revealed larger correlations for older compared to younger adults ($F_{1,42} = 15.868$, $p = 2.6 \bullet 10^{-4}$, $\omega^2 = 0.147$; Fig. 9C), whereas the effect of Session and the Session × Age Group interaction were not significant (ps > 0.05).

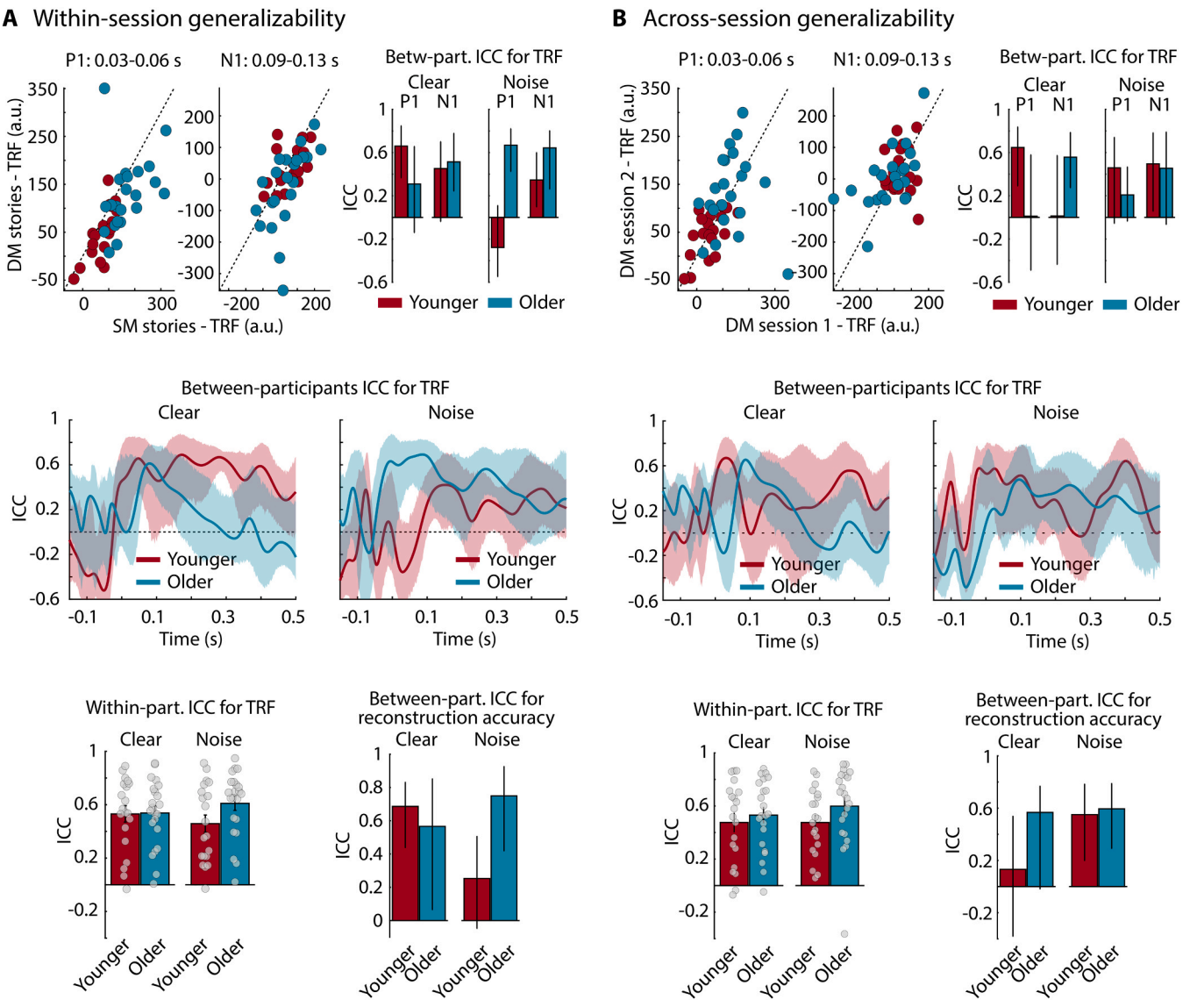

Fig. 8. Generalizability of neural responses across stories, within and across sessions. A: Generalizability within a session (i.e., SM stories vs DM stories in session 1): Top row: Scatter plots for TRF amplitude in the 0.03–0.06 s (P1) and the 0.09–0.13 s (N1) time windows. Bar graphs of between-participants ICC values for the P1 and N1 time windows. Error bars reflect the 95% confidence intervals from bootstrapping. Middle row: Between-participants ICC time courses for both speech-clarity conditions (clear, noise) and age groups (younger, older). ICC was calculated for 0.05 s sliding time windows centered on each time point. Shaded areas reflect the 95% confidence intervals. Bottom row, left: Within-participant ICC calculated using TRF time courses from 0 s to 0.4 s. Bars and error bars reflect the mean and standard error of the mean, respectively. Dots reflect data from individual participants. Bottom row, right: Between-participants reliability for EEG reconstruction accuracy. Error bars reflect the 95% confidence intervals. **B:** Generalizability across sessions (i.e., session 1 vs session 2 for DM stories): Same as in panel A.

Scatter plots of TRF amplitudes are shown in Fig. 9D. The time courses of between-participants ICC for TRF amplitudes in Fig. 9E show ICC values of about 0.5–0.7 (moderate) between 0 and 0.2 s, although ICC for younger adults was lower around 0.1 s (<0.5). Within-participant ICC was 0.62 and 0.74 for younger and older adults, respectively (i.e., moderate; Koo and Li, 2016; Fig. 9F) and there was no difference between age groups ($p > 0.05$). Between-participants ICC for the EEG reconstruction accuracy was about 0.55 and 0.63 for younger and older adults, respectively (Fig. 9G).

3.8. Comparing reliability and generalizability

We also assessed whether within-participant ICC values differed between assessment types: ERP reliability (session 1 vs session 2), TRF reliability (SM stories, session 1 vs session 2), within-session TRF generalizability across story/speaker (SM vs DM stories, session 1), TRF generalizability across session/story/speaker (DM stories, session 1 vs session 2), TRF generalizability across session/story (DB stories, session 1 vs session 2).

The rmANOVA revealed a significant main effect of Assessment Type ($F_{4160} = 22.958$, $p = 5.2 \bullet 10^{-15}$, $\omega^2 = 0.219$; Fig. 10). Post hoc comparisons showed that ICC for ERP reliability was greater than ICC for TRF reliability ($t_{41} = 6.012$, $p_{Holm} = 9.6 \bullet 10^{-8}$, $d = 1.062$), within-session TRF generalizability across story/speaker ($t_{41} = 7.760$, $p_{Holm} = 8.5 \bullet 10^{-12}$, $d = 1.371$), TRF generalizability across session/story/speaker ($t_{41} = 8.516$, $p_{Holm} = 1.1 \bullet 10^{-13}$, $d = 1.505$), and TRF generalizability across session/story ($t_{41} = 4.337$, $p_{Holm} = 1.8 \bullet 10^{-4}$, $d = 0.766$). ICC for TRF generalizability across session/story (DB stories) was also greater than ICC for within-session TRF generalizability across story/speaker ($t_{41} = 3.423$, $p_{Holm} = 0.004$, $d = 0.605$) and TRF generalizability across session/story/speaker ($t_{41} = 4.179$, $p_{Holm} = 2.8 \bullet 10^{-4}$, $d = 0.738$). No other effects, including age effects, were significant (ps >

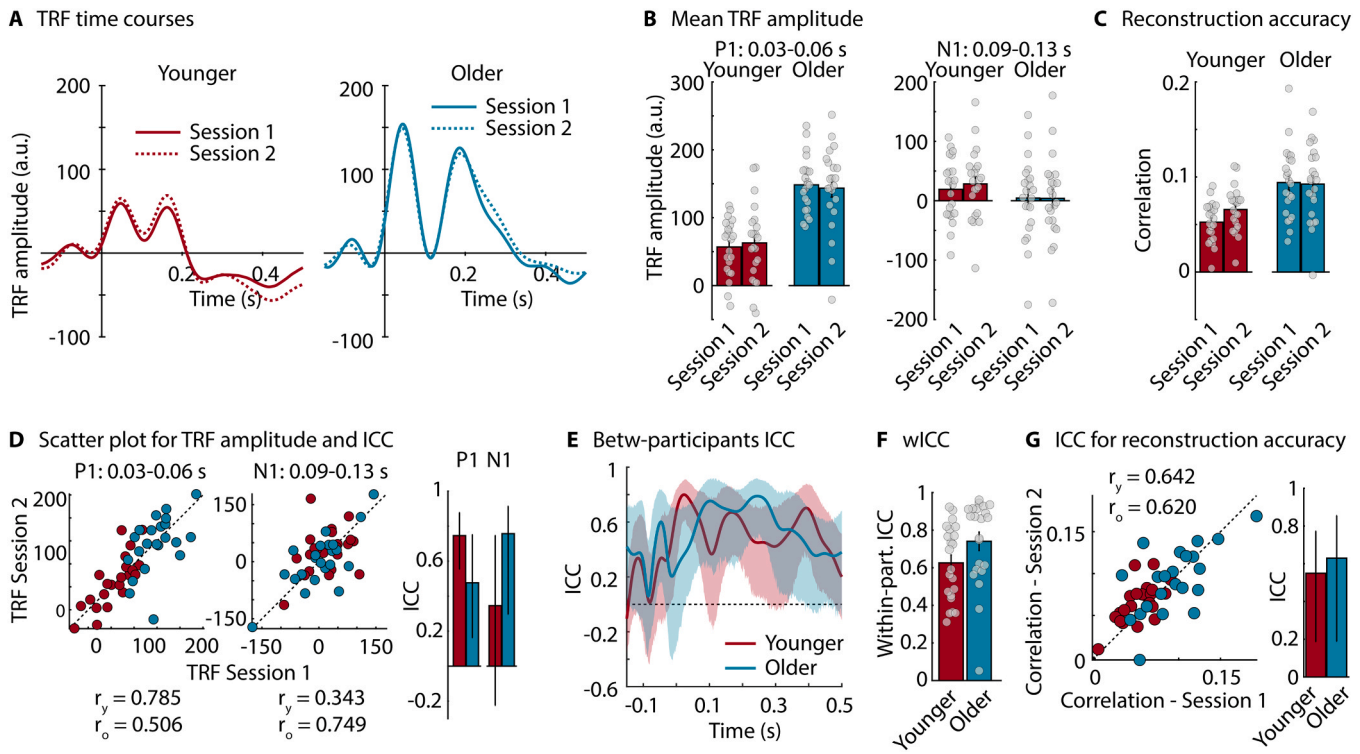

Fig. 9. Neural responses and intra-class correlation for DB stories. A: Temporal response functions (TRFs). **B:** Mean TRF amplitude for the 0.03–0.06 s and the 0.09–0.13 s time windows. **C:** EEG reconstruction accuracy. **D:** Scatter plots for the TRF amplitude in the 0.03–0.06 s and the 0.09–0.13 s time windows. Pearson correlations are provided below each plot. The subscripts y and o indicate correlations for younger and older adults, respectively. Bar graphs of between-participants ICC values for the P1 and N1 time windows. Error bars reflect the 95% confidence intervals from bootstrapping. **E:** Between-participants ICC calculated for 0.05 s sliding time windows centered on each time point. Shaded areas reflect the 95% confidence intervals. **F:** Within-participant ICC calculated using TRF time courses from 0 s to 0.4 s **G:** Scatter plot and ICC (including 95% confidence intervals) for reconstruction accuracy. Pearson correlations are provided in the scatter plot. The subscripts y and o indicate correlations for younger and older adults, respectively. Dots in the different panels reflect data from individual participants.

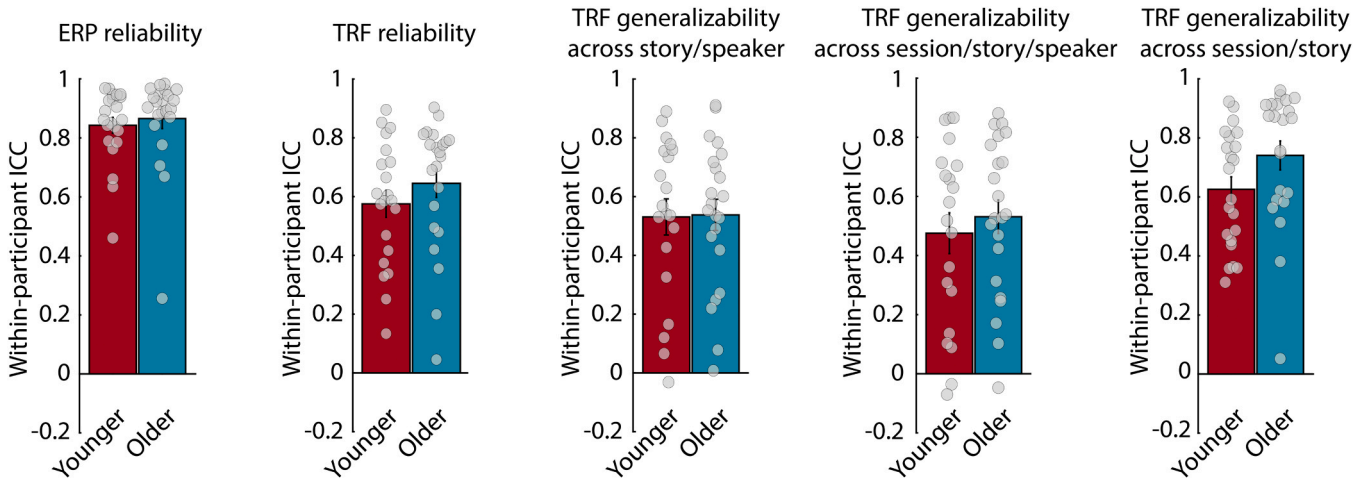

Fig. 10. Comparison of reliability and generalizability. Within-participant ICC using TRF time courses from 0 s to 0.4 s. ICC is shown for ERP reliability (responses to noise bursts; session 1 vs session 2), TRF reliability (responses to the clear SM story; session 1 vs session 2), within-session TRF generalizability across story/speaker (responses to clear stories in session 1; SM vs DM), TRF generalizability across session/story/speaker (responses to clear DM stories; session 1 vs session 2), TRF generalizability across session/story (responses to DB stories; session 1 vs session 2). Bar graphs show the mean ICC. Error bars reflect the standard error of the mean. Dots reflect data from individual participants.

0.05).

4. Discussion

Assessment of the neural tracking of continuous, naturalistic speech is increasingly used to understand speech encoding and considered a potential clinical biomarker, for example, for age-related hearing loss. However, to use neural speech tracking as a biomarker requires knowledge about its reliability. The current study investigated the reliability and generalizability of neural speech tracking in younger and older adults while they listened to stories and EEG was recorded in two separate sessions. Neural responses to noise bursts were used as a

benchmark for which we expected high reliability. Early responses to noise bursts (~0.05 and ~0.1 s), neural-speech tracking responses ~0.05 s (P1), and EEG reconstruction accuracy were larger for older compared to younger adults. Critically, reliability of neural speech tracking was moderate (ICC ~0.5–0.75) in younger and older adults, and there was a tendency for reliability to be larger in older adults for speech presented in moderate background babble. Reliability for responses to noise bursts was higher (ICC >0.8) than the speech-tracking reliability in both younger and older adults. Neural-speech tracking responses also moderately generalized across different stories (ICC ~0.5–0.6). Overall, the current study provides an important step in the development of an objective marker of speech encoding that can be used in clinical contexts.

4.1. Age-related enhancement of neural responses to white noise and stories

We observed larger responses for older compared to younger adults, for both noise bursts and spoken stories. This response enhancement has been observed frequently in older relative to younger adults for speech and non-speech sounds (Alain et al., 2012; Alain et al., 2014; Bidelman et al., 2014; Brodbeck et al., 2018; Decruy et al., 2019; Harris et al., 2022; Herrmann et al., 2016; Herrmann et al., 2013b; Millman et al., 2017; Presacco et al., 2016b; Tremblay et al., 2003), and is sometimes more prominent for the earlier (~0.05 s) than the later (~0.1 s) cortical response (Alain et al., 2014; Harris et al., 2022; Figs. 5 and 6).

A variety of possible mechanisms for the age-related response enhancement have been discussed, including a loss of cortical inhibition resulting from peripheral deafferentation (Auerbach et al., 2014; Chambers et al., 2016; Herrmann and Butler, 2021b; Resnik and Polley, 2017; Salvi et al., 2017), recruitment of additional cortical resources (Brodbeck et al., 2018; Gillis et al., 2022), and increased attention or effort (Decruy et al., 2020a; Gillis et al., 2022; Vanthornhout et al., 2019). However, cognitive factors are unlikely the sole contributors to the age-related response enhancement, given that it is present also under distracted listening conditions (Harris et al., 2022; Herrmann et al., 2016; Herrmann et al., 2018; Fig. 3).

4.2. Reliability of neural speech tracking

We investigated the reliability of neural speech tracking by recording EEG from individuals while they listened to the same stories twice in separate sessions. Younger and older adults reported lower listening effort for the masked story in session 2 compared to session 1, indicating that prior knowledge about speech can reduce listening effort (Herrmann and Johnsrude, 2020; Holmes et al., 2018; Obleser and Kotz, 2010; Pichora-Fuller et al., 1995; Signoret et al., 2011). However, younger adults found listening to the same story a second time less enjoyable and absorbing than when they listened to it for the first time, whereas older adults found them similarly enjoyable and absorbing both times (Fig. 4). Our observation that listening effort, enjoyment, and absorption are reduced when listening to the same story several times perhaps suggests that different stories should be used in clinical contexts when individuals are assessed repeatedly.

There were no differences in TRF amplitude or EEG reconstruction accuracy between session 1 and 2, and peak reliability of neural-tracking responses (TRF, accuracy) was moderate for both age groups (ICC ~0.5–0.75; Koo and Li, 2016; Fig. 7), although it appeared that in several analyses reliability was somewhat lower in younger than older adults (e.g., for EEG reconstruction accuracy, Fig. 7C). This may, in part, be due to larger responses for older compared to younger adults, but might also be related to the decrease in absorption and enjoyment from session 1 to session 2 for younger adults that was absent in older adults. The reliability for neural speech tracking was lower compared to the reliability for responses to noise bursts (showing good reliability; ICC >0.75; Koo and Li, 2016), which is consistent with the good reliability for simple sound stimuli observed previously (Bidelman et al., 2018; Legget et al., 2017; Tervaniemi et al., 1999b).

Neural speech tracking is increasingly used to investigate clinical phenomena, such as hearing loss (Decruy et al., 2020b; Presacco et al., 2019; Schmitt et al., 2022), and researchers have suggested that neural speech tracking could be an important biomarker (Gillis et al., 2022; Palana et al., 2022; Schmitt et al., 2022). However, a reliability of 0.7 or higher has been recommended for measures used in clinical research (Frost et al., 2007; Mokkink et al., 2022; Nunnally and Bernstein, 1994) and it thus appears that the moderate (ICC ~0.5–0.75) reliability for neural speech tracking in older adults may not be sufficiently high, or only in specific time windows, to meet this criterion.

The current data further help quantify the upper bound of how well neural speech tracking can correlate with or predict a clinical condition. That is, the maximum correlation between two measures is equal to the square root of the product of their reliabilities (sqrt(reliability of Measure A × reliability of Measure B); Bedeian, 2014; Bedeian et al., 1997; Goodwin and Leech, 2006). Standard clinical measures tend to have good test-retest reliability (ICC ~0.8; sometimes correlation instead of ICC is provided). For example, good-to-high reliability has been observed for audiometric assessments (McClannahan et al., 2021) and cognitive assessments (Montreal Cognitive Assessment; Gupta et al., 2019; Lee et al., 2022; Nasreddine et al., 2005). The degree of reliability in the current study was somewhat variable across neural-tracking measures, particularly, in younger adults and for reconstruction accuracy. Nevertheless, assuming a moderate ~0.6 reliability of the neural-tracking response and a reliability of 0.8 for audiological assessments, we would expect a maximum correlation of 0.69 between the two measures.

Critically, the current reliability assessment has focused on the neural tracking of the speech onset-envelope. The speech envelope and onset-envelope have been used most (Ding et al., 2015; Ding and Simon, 2012; Fiedler et al., 2021; Fiedler et al., 2019; Hertrich et al., 2012; Lalor and Foxe, 2010), can be easily calculated, and may thus be particularly useful as a clinical biomarker compared to recent approaches to assess tracking of semantic features of speech that require more complex analyses (Broderick et al., 2018; Broderick et al., 2021; Marlies et al., 2021; Yasmin et al., 2023). Nevertheless, future work should further investigate the reliability of the neural tracking of linguistic features of speech.

4.3. Generalizability of neural speech tracking across stories

Diagnosis or treatment of a clinical condition can involve repeated assessments of a person using the same measure or procedure, for example, when evaluating intervention progress for the treatment of hearing loss. A biomarker of speech processing should thus be independent of specific speech stimuli and instead generalize across stimuli to avoid prior knowledge affecting the measurement outcome. The current generalizability data show that across-story ICC did not differ from same-story ICC (strict reliability), although the former was numerically lower (Fig. 10). While ICC values for reliability (same story) and generalizability (different stories) were only moderate, there seems to be little indication that using neural speech tracking as an assessment tool would suffer from using different stories in the case that two or more assessments are needed.

Critically, our results suggest that generalizability may be highest if audiobooks (here Storybook stories) spoken by the same person are used as speech materials (Figs. 9 and 10). The Moth stories are highly engaging, enjoyable to participants, and reflect real-life speech with pauses, disfluencies, and other idiosyncrasies. Some work suggests The Moth stories are more enjoyable and absorbing than Storybook stories (Mathiesen et al., 2023), although this difference was not observed in the current study. We recommend that clinical research could perhaps rely on the more controlled audiobooks spoken by the same person if repeated assessments are needed, while enjoyment for such stories, relative to more real-life speech, is likely reduced only minimally.

4.4. Limitations

Larger neural responses were observed for older compared to younger adults. The data also appear to indicate numerically larger reliability values (ICC) in several analyses for older compared to younger adults. Although the current study was not designed to specifically investigate how larger responses relate to reliability, a larger response might be expected to give rise to a higher reliability. A larger neural response may result from a larger number of neurons being activated concurrently, the timing of the responses being more consistently related to the stimulus, or the stimulus eliciting a larger post-synaptic potential. These potential mechanisms are consistent with a loss of inhibition and increased excitation in older compared to younger adults (Auerbach et al., 2014; Caspary et al., 2008; Chambers et al., 2016; Herrmann and Butler, 2021b), and lead to better signal-to-noise ratio in electroencephalographic data. A higher signal-to-noise ratio and more consistent timing or more concurrent activity would help increase reliability, because less noise is represented in the neural response. Critically, the current study provides ICC values for different populations and may thus be informative for future studies aiming to use a neural biomarker of speech processing for a specific population under investigation.

The current study included 44 participants, 22 per age group, participating in two sessions (i.e., 88 EEG sessions were recorded). Assessing reliability and estimating the confidence intervals related to reliability benefits from a high number of participants. Some authors have suggested, as a rule of thumb, to obtain data from 30 or more participants whenever possible for reliability analyses, mainly because reliability could be lower for a low number of participants (Koo and Li, 2016). Critically, the number of participants in our study is comparable or higher relative to the number of participants included in many previous works assessing the reliability of neural responses (Easwar et al., 2020; Hamad et al., 2023; Hirano et al., 2020; Legget et al., 2017; Lu et al., 2007; McEvoy et al., 2000; McFadden et al., 2014; Tervaniemi et al., 1999a; Tervaniemi et al., 1999b; Williams et al., 2005; although see; Rentzsch et al., 2008; Walhovd and Fjell, 2002). Moreover, we show good reliability for neural responses to noise bursts (ICC >0.75), whereas reliability was lower for speech materials. Since the number of participants was the same for these analyses, we can perhaps be confident that the number of participants does not limit the chances of observing good reliability. We further provide evidence that stories that are more audiobook-like may result in higher reliability than naturalistic spoken speech. Hence, although the possibility of higher ICC values in a future study with a larger sample cannot be excluded, the current data suggest that the reliability of neural speech tracking is lower than for simple noise stimuli and may not be high enough for clinical standards.

5. Conclusions

The current study investigated the reliability and generalizability of neural speech tracking in younger and older adults. Participants listened to stories that either repeated in two sessions (to test reliability) or differed across sessions (to test generalizability). We observed larger neural responses for older compared to younger adults during story listening, consistent with a loss of inhibition in the aged auditory system. Reliability for the neural-tracking response was moderate (ICC ~0.5–0.75) and somewhat lower across stories and speech-clarity conditions in younger compared to older adults. Generalizability across stories appeared greatest when audiobook-like stories spoken by the same person were used as speech materials. The current data provide results critical for the development of an objective biomarker of speech processing, but also suggest that further work is needed to increase the reliability of the neural-tracking response to meet clinical standards.

CRediT authorship contribution statement

Ryan A. Panela: Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Francesca Copelli:** Methodology, Formal analysis, Data curation, Writing – review & editing. **Björn Herrmann:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgements

We thank Christie Tsagopoulos and Tazeen Atif for their help with data collection. The research was supported by the Canada Research Chair Program (CRC-2019-00156, 232733) and the Natural Sciences and Engineering Research Council of Canada (Discovery Grant: RGPIN-2021-02602).

Verification

This manuscript has not been published previously, it is not under consideration for publication elsewhere, its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out. If accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.neurobiolaging.2023.11.007.

References

Alain, C., McDonald, K., Van Roon, P., 2012. Effects of age and background noise on processing a mistuned harmonic in an otherwise periodic complex sound. Hear. Res. 283, 126–135. https://doi.org/10.1016/j.heares.2011.10.007.

Alain, C., Roye, A., Salloum, C., 2014. Effects of age-related hearing loss and background noise on neuromagnetic activity from auditory cortex. Front. Syst. Neurosci. 8, Artic., 8 https://doi.org/10.3389/fnsys.2014.00008.

Auerbach, B.D., Rodrigues, P.V., Salvi, R.J., 2014. Central gain control in tinnitus and hyperacusis (Article). Front. Neurol. 5, 206. https://doi.org/10.3389/fneur.2014.00206.

Bedeian, A.G., 2014. More than meets the eye": a guide to interpreting the descriptive statistics and correlation matrices reported in management research. Acad. Manag. Learn. Educ. 13 (1), 121–135. https://doi.org/10.5465/amle.2013.0001.

Bedeian, A.G., Day, D.V., Kelloway, E.K., 1997. Correcting for measurement error attenuation in structural equation models: some important reminders. Educ. Psychol. Meas. 57 (5), 785–799. https://doi.org/10.1177/0013164497057005004.

Bell, A.J., Sejnowski, T.J., 1995. An information maximization approach to blind separation and blind deconvolution. Neural Comput. 7, 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129.

Bidelman, G.M., Villafuerte, J.W., Moreno, S., Alain, C., 2014. Age-related changes in the subcorticalecortical encoding and categorical perception of speech. Neurobiol. Aging 35, 2526–2540. https://doi.org/10.1016/j.neurobiolaging.2014.05.006.

Bidelman, G.M., Pousson, M., Dugas, C., Fehrenbach, A., 2018. Test–retest reliability of dual-recorded brainstem versus cortical auditory-evoked potentials to speech. J. Am. Acad. Audio 29 (02), 164–174. https://doi.org/10.3766/jaaa.16167.

Biesmans, W., Das, N., Francart, T., Bertrand, A., 2017. Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. IEEE Trans. Neural Syst. Rehabil. Eng. 25 (5), 402–412. https://doi.org/10.1109/TNSRE.2016.2571900.

Bilger, R.C., 1984. Manual for the Clinical Use of the Revised SPIN Test. The University of Illinois, Champaign, IL, USA.

Bilger, R.C., Nuetzel, J.M., Rabinowitz, W.M., Rzeczkowski, C., 1984. Standardization of a test of speech perception in noise. J. Speech Lang. Hear. Res. 27 (1), 32–48. https://doi.org/10.1044/jshr.2701.32.

Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.F., Brennan, S.E., 2001. Disfluency rates in conversation: effects of age, relationship, topic, role, and gender. Lang. Speech 44 (2), 123–147. https://doi.org/10.1177/00238309010440020101.

Brodbeck, C., Simon, J.Z., 2020. Continuous speech processing. Curr. Opin. Physiol. 18, 25–31. https://doi.org/10.1016/j.cophys.2020.07.014.

Brodbeck, C., Presacco, A., Anderson, S., Simon, J.Z., 2018. Over-representation of speech in older adults originates from early response in higher order auditory cortex. Acta Acust. U. Acust. 104 (5), 774–777. https://doi.org/10.3813/aaa.919221.

Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., Lalor, E.C., 2018. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr. Biol. 28, 803–809. https://doi.org/10.1016/j.cub.2018.01.080.

Broderick, M.P., Di Liberto, G.M., Anderson, A.J., Rofes, A., Lalor, E.C., 2021. Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. Sci. Rep. 11, 4963. https://doi.org/10.1038/s41598-021-84597-9.

Broderick, M.P., Zuk, N.J., Anderson, A.J., Lalor, E.C., 2022. More than Words: neurophysiological correlates of semantic dissimilarity depend on comprehension of the speech narrative. Eur. J. Neurosci. 56 (8), 5201–5214. https://doi.org/10.1111/ejn.15805.

Cabral-Calderin, Y., Henry, M.J., 2022. Reliability of neural entrainment in the human auditory system. J. Neurosci. 42 (5), 894. https://doi.org/10.1523/JNEUROSCI.0514-21.2021.

Caspary, D.M., Ling, L., Turner, J.G., Hughes, L.F., 2008. Inhibitory neurotransmission, plasticity and aging in the mammalian central auditory system. J. Exp. Biol. 211, 1781–1791. https://doi.org/10.1242/jeb.013581.

Chambers, A.R., Resnik, J., Yuan, Y., Whitton, J.P., Edge, A.S., Liberman, M.C., Polley, D. B., 2016. Central gain restores auditory processing following near-complete cochlear denervation. Neuron 89, 867–879. https://doi.org/10.1016/j.neuron.2015.12.041.

Cohen, S.S., Parra, L.C., 2016. Memorable audiovisual narratives synchronize sensory and supramodal neural responses. eNeuro 3, e0203. https://doi.org/10.1523/ENEURO.0203-16.2016.

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

Crosse, M.J., Zuk, N.J., Di Liberto, G.M., Nidiffer, A.R., Molholm, S., Lalor, E.C., 2021. Linear modeling of neurophysiological responses to speech and other continuous stimuli: methodological considerations for applied research. Front. Neurosci. 15 https://doi.org/10.3389/fnins.2021.705621.

Decruy, L., Vanthornhout, J., Francart, T., 2019. Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. J. Neurophysiol. 122 (2), 601–615. https://doi.org/10.1152/jn.00687.2018.

Decruy, L., Lesenfants, D., Vanthornhout, J., Francart, T., 2020a. Top-down modulation of neural envelope tracking: The interplay with behavioral, self-report and neural measures of listening effort. Eur. J. Neurosci. 52 (5), 3375–3393. https://doi.org/10.1111/ejn.14753.

Decruy, L., Vanthornhout, J., Francart, T., 2020b. Hearing impairment is associated with enhanced neural tracking of the speech envelope. Hear. Res. 393, 107961 https://doi.org/10.1016/j.heares.2020.107961.

Di Liberto, Giovanni M., O'Sullivan, James A., Lalor, Edmund C., 2015. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr. Biol. 25 (19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030.

Dial, H.R., Gnanateja, G.N., Tessmer, R.S., Gorno-Tempini, M.L., Chandrasekaran, B., Henry, M.L., 2021. Cortical tracking of the speech envelope in logopenic variant primary progressive aphasia. Front. Hum. Neurosci. 14 https://doi.org/10.3389/fnhum.2020.597694.

Ding, N., Simon, J.Z., 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. 109, 11854–11859. https://doi.org/10.1073/pnas.1205381109.

Ding, N., Chatterjee, M., Simon, J.Z., 2014. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. NeuroImage 88, 41–46. https://doi.org/10.1016/j.neuroimage.2013.10.054.

Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2015. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19, 158–164. https://doi.org/10.1038/nn.4186.

Dmochowski, J.P., Sajda, P., Dias, J., Parra, L.C., 2012. Correlated components of ongoing EEG point to emotionally laden attention – a possible marker of engagement? Front. Hum. Neurosci. 6, Artic., 112 https://doi.org/10.3389/fnhum.2012.00112.

Dmochowski, J.P., Bezdek, M.A., Abelson, B.P., Johnson, J.S., Schumacher, E.H., Parra, L.C., 2014. Audience preferences are predicted by temporal reliability of neural processing. Nat. Commun. 29, 4567. https://doi.org/10.1038/ncomms5567.

Easwar, V., Scollie, S., Aiken, S., Purcell, D., 2020. Test-retest variability in the characteristics of envelope following responses evoked by speech stimuli. Ear Hear 41 (1). https://doi.org/10.1097/AUD.0000000000000739.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7 (1), 1–26. https://doi.org/10.1214/aos/1176344552.

Emily, S.T., Farhin, A., Edmund, C.L., 2022. Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment. J. Neurosci. 42 (4), 682. https://doi.org/10.1523/JNEUROSCI.1455-20.2021.

Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., Obleser, J., 2017. Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. J. Neural Eng. 14, 036020 https://doi.org/10.1088/1741-2552/aa66dd.

Fiedler, L., Wöstmann, M., Herbst, S.K., Obleser, J., 2019. Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. Neuroimage 186, 33–42. https://doi.org/10.1016/j.neuroimage.2018.10.057.

Fiedler, L., Ala, T.S., Graversen, C., Alickovic, E., Lunner, T., Wendt, D., 2021. Hearing aid noise reduction lowers the sustained listening effort during continuous speech in noise—a combined pupillometry and EEG study. Ear Hear. 42, 1590–1601. https://doi.org/10.1097/AUD.0000000000001050.

Friederici, A.D., Pfeifer, E., Hahne, A., 1993. Event-related brain potentials during natural speech processing: effects of semantic, morphological and syntactic violations. Cogn. Brain Res. 1, 183–192. https://doi.org/10.1016/0926-6410(93)90026-2.

Frost, M.H., Reeve, B.B., Liepa, A.M., Stauffer, J.W., Hays, R.D., the Mayo/FDA Patient-Reported Outcomes consensus Meeting Group, 2007. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health 10 (*s*2), S94–S105. https://doi.org/10.1111/j.1524-4733.2007.00272.x.

Gillis, M., Van Canneyt, J., Francart, T., Vanthornhout, J., 2022. Neural tracking as a diagnostic tool to assess the auditory pathway. Hear. Res. 426, 108607 https://doi.org/10.1016/j.heares.2022.108607.

Goodwin, L.D., Leech, N.L., 2006. Understanding correlation: factors that affect the size of r. J. Exp. Educ. 74 (3), 249–266. https://doi.org/10.3200/JEXE.74.3.249-266.

Gupta, M., Gupta, V., Nagar Buckshee, R., Sharma, V., 2019. Validity and reliability of hindi translated version of Montreal cognitive assessment in older adults. Asian J. Psychiatry 45, 125–128. https://doi.org/10.1016/j.ajp.2019.09.022.

Hamad, H., Washnik, N.J., Suresh, C.H., 2023. Next-generation auditory steady-state responses in normal-hearing adults: a pilot test-retest reliability study. J. Otorhinolaryngol., Hear. Balance Med. 4 (2), 6. https://doi.org/10.3390/ohbm4020006.

Hamilton, L.S., Huth, A.G., 2020. The revolution will not be controlled: natural stimuli in speech neuroscience. Lang., Cogn. Neurosci. 35 (5), 573–582. https://doi.org/10.1080/23273798.2018.1499946.

Harris, K.C., Dias, J.W., McClaskey, C., M., Rumschlag, J., Prisciandaro, J., Dubno, J.R., 2022. Afferent Loss, GABA, and Central Gain in Older Adults: Associations with Speech Recognition in Noise. J. Neurosci. 42 (38), 7201. https://doi.org/10.1523/JNEUROSCI.0242-22.2022.

Herrmann, B., Butler, B.E., 2021a. Aging auditory cortex: The impact of reduced inhibition on function. In: Martin, C.R., Preedy, V.R., Rajendram, R. (Eds.), Assessments, Treatments and Modelling in Aging and Neurological Disease: The Neuroscience of Aging. Academic Press, pp. 183–192.

Herrmann, B., Butler, B.E., 2021b. Hearing loss and brain plasticity: the hyperactivity phenomenon. Brain Struct. Funct. 226, 2019–2039. https://doi.org/10.1007/s00429-021-02313-9.

Herrmann, B., Johnsrude, I.S., 2020. Absorption and enjoyment during listening to acoustically masked stories. Trends Hear. 24, 1–18. https://doi.org/10.1177/2331216520967850.

Herrmann, B., Maess, B., Friederici, A.D., 2011. Violation of syntax and prosody - Disentangling their contributions to the early left anterior negativity (ELAN). Neurosci. Lett. 490, 116–120. https://doi.org/10.1016/j.neulet.2010.12.039.

Herrmann, B., Henry, M.J., Obleser, J., 2013a. Frequency-specific adaptation in human auditory cortex depends on the spectral variance in the acoustic stimulation. J. Neurophysiol. 109, 2086–2096. https://doi.org/10.1152/jn.00907.2012.

Herrmann, B., Henry, M.J., Scharinger, M., Obleser, J., 2013b. Auditory filter width affects response magnitude but not frequency specificity in auditory cortex. Hear. Res. 304, 128–136. https://doi.org/10.1016/j.heares.2013.07.005.

Herrmann, B., Henry, M.J., Johnsrude, I.S., Obleser, J., 2016. Altered temporal dynamics of neural adaptation in the aging human auditory cortex. Neurobiol. Aging 45, 10–22. https://doi.org/10.1016/j.neurobiolaging.2016.05.006.

Herrmann, B., Maess, B., Johnsrude, I.S., 2018. Aging affects adaptation to sound-level statistics in human auditory cortex. J. Neurosci. 38, 1989–1999. https://doi.org/10.1523/JNEUROSCI.1489-17.2018.

Herrmann, B., Maess, B., Johnsrude, I.S., 2022. A neural signature of regularity in sound is reduced in older adults. Neurobiol. Aging 109, 1–10. https://doi.org/10.1016/j.neurobiolaging.2021.09.011.

Hertrich, I., Dietrich, S., Trouvain, J., Moos, A., Ackermann, H., 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. https://doi.org/10.1111/j.1469-8986.2011.01314.x.

Hirano, Y., Nakamura, I., Tamura, S., Onitsuka, T., 2020. Long-term test-retest reliability of auditory gamma oscillations between different clinical EEG systems. Front. Psychiatry 11. https://doi.org/10.3389/fpsyt.2020.00876.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67. https://doi.org/10.2307/1267351.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6 (2), 65–70. ⟨http://www.jstor.org/stable/4615733⟩.

Holmes, E., Folkeard, P., Johnsrude, I.S., Scollie, S., 2018. Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. Int. J. Audiol. 57, 483–492. https://doi.org/10.1080/14992027.2018.1432901.

Irsik, V.C., Almanaseer, A., Johnsrude, I.S., Herrmann, B., 2021. Cortical responses to the amplitude envelopes of sounds change with age. J. Neurosci. 41, 5045–5055. https://doi.org/10.1523/JNEUROSCI.2715-20.2021.

Irsik, V.C., Johnsrude, I.S., Herrmann, B., 2022a. Age-related deficits in dip-listening evident for isolated sentences but not for spoken stories. Sci. Rep. 12, 5898. https://doi.org/10.1038/s41598-022-09805-6.

Irsik, V.C., Johnsrude, I.S., Herrmann, B., 2022b. Neural activity during story listening is synchronized across individuals despite acoustic masking. J. Cogn. Neurosci. 34, 933–950. https://doi.org/10.1162/jocn_a_01842.

JASP, 2022. JASP [Computer software], ⟨https://jasp-stats.org/⟩.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kries, J., De Clercq, P., Gillis, M., Vanthornhout, J., Lemmens, R., Francart, T., Vandermosten, M., 2023. Exploring neural tracking of acoustic and linguistic speech representations in individuals with post-stroke aphasia. BioRxiv.

Kuijpers, M.M., Hakemulder, F., Tan, E.S., Doicaru, M.M., 2014. Exploring absorbing reading experiences: Developing and validating a self-report scale to measure story world absorption. Sci. Study Lit. 4, 89–122. https://doi.org/10.1075/ssol.4.1.05kui.
Lalor, E.C., Foxe, J.J., 2010. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur. J. Neurosci. 31, 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x.
Lee, Y.-C., Lin, Y.-T., Chiu, E.-C., 2022. A comparison of test-retest reliability of four cognitive screening tools in people with dementia. Disabil. Rehabil. 44 (15), 4090–4095. https://doi.org/10.1080/09638288.2021.1891466.
Legget, K.T., Hild, A.K., Steinmetz, S.E., Simon, S.T., Rojas, D.C., 2017. MEG and EEG demonstrate similar test-retest reliability of the 40Hz auditory steady-state response. Int. J. Psychophysiol. 114, 16–23. https://doi.org/10.1016/j.ijpsycho.2017.01.013.
Lesenfants, D., Vanthornhout, J., Verschueren, E., Decruy, L., Francart, T., 2019. Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations. Hear. Res. 380, 1–9. https://doi.org/10.1016/j.heares.2019.05.006.
Lockhart, R.S., 1998. Introduction to Statistics and Data Analysis for the Behavioral Sciences. Worth Publishers, New York, USA.
Lu, B.Y., Edgar, J.C., Jones, A.P., Smith, A.K., Huang, M.-X., Miller, G.A., Cañive, J.M., 2007. Improved test–retest reliability of 50-ms paired-click auditory gating using magnetoencephalography source modeling. Psychophysiology 44 (1), 86–90. https://doi.org/10.1111/j.1469-8986.2006.00478.x.
Makeig, S., Bell, A.J., Jung, T.-P., Sejnowski, T.J., 1995. Independent component analysis of electroencephalographic data. In: Touretzky, D., Mozer, M., Hasselmo, M. (Eds.), Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA, USA, pp. 145–151.
Marinkovic, K., Dhond, R.P., Dale, A.M., Glessner, M., Carr, V., Halgren, E., 2003. Spatiotemporal Dynamics of Modality-Specific and Supramodal Word Processing. Neuron 38, 487–497. https://doi.org/10.1016/s0896-6273(03)00197-1.
Marlies, G., Jonas, V., Jonathan, Z.S., Tom, F., Christian, B., 2021. Neural markers of speech comprehension: measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. J. Neurosci. 41 (50), 10316. https://doi.org/10.1523/JNEUROSCI.0812-21.2021.
Mathiesen, S.L., Van Hedger, S.C., Irsik, V.C., Bain, M.M., Johnsrude, I.S., Herrmann, B., 2023. Exploring age differences in absorption and enjoyment during story listening. PsyArXiv.
McClannahan, K.S., Chiu, Y.-F., Sommers, M.S., Peelle, J.E., 2021. Test-retest reliability of audiometric assessment in individuals with mild dementia. JAMA Otolaryngol. Neck Surg. 147 (5), 442–449. https://doi.org/10.1001/jamaoto.2021.0012.
McDermott, Josh H., Simoncelli, Eero P., 2011. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron 71 (5), 926–940. https://doi.org/10.1016/j.neuron.2011.06.032.
McEvoy, L.K., Smith, M.E., Gevins, A., 2000. Test–retest reliability of cognitive EEG. Clin. Neurophysiol. 111 (3), 457–463. https://doi.org/10.1016/S1388-2457(99)00258-8.
McFadden, K.L., Steinmetz, S.E., Carroll, A.M., Simon, S.T., Wallace, A., Rojas, D.C., 2014. Test-retest reliability of the 40 Hz EEG auditory steady-state response. PLoS One 9 (1), e85748. https://doi.org/10.1371/journal.pone.0085748.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1 (1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30.
Millman, R.E., Mattys, S.L., Gouws, A.D., Prendergast, G., 2017. Magnified neural envelope coding predicts deficits in speech perception in noise. J. Neurosci. 37, 7727–7736. https://doi.org/10.1523/JNEUROSCI.2722-16.2017.
Mokkink, L.B., de Vet, H., Diemeer, S., Eekhout, I., 2022. Sample size recommendations for studies on reliability and measurement error: an online application based on simulation studies. Health Serv. Outcomes Res. Methodol. https://doi.org/10.1007/s10742-022-00293-9.
Moore, B.C.J., 2007. Cochlear Hearing Loss: Physiological, Psychological and Technical Issues. John Wiley & Sons, Ltd, West Sussex, Engand.
Näätänen, R., Picton, T.W., 1987. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology 24, 375–425. https://doi.org/10.1111/j.1469-8986.1987.tb00311.x.
Nasreddine, Z.S., Phillips, N.A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H., 2005. The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. J. Am. Geriatr. Soc. 53, 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x.
Nunnally, J.C., Bernstein, I.H., 1994 (ed). Psychometric theory, 3rd ed. McGraw-Hill,, New York, USA.
Obleser, J., Kotz, S.A., 2010. Expectancy constraints in degraded speech modulate the language comprehension network. Cereb. Cortex 20, 633–640.
Olichney, J.M., Yang, J.-C., Taylor, J., Kutas, M., 2011. Cognitive event-related potentials: biomarkers of synaptic dysfunction across the stages of Alzheimer's disease. J. Alzheimer's Dis. 26, 215–228. https://doi.org/10.3233/JAD-2011-0047.
Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci., 156869 https://doi.org/10.1155/2011/156869.
Palana, J., Schwartz, S., Tager-Flusberg, H., 2022. Evaluating the use of cortical entrainment to measure atypical speech processing: a systematic review. Neurosci. Biobehav. Rev. 133, 104506 https://doi.org/10.1016/j.neubiorev.2021.12.029.
Pichora-Fuller, M.K., Schneider, B.A., Daneman, M., 1995. How young and old adults listen to and remember speech in noise. J. Acoust. Soc. Am. 97, 593–608. https://doi.org/10.1121/1.412282.
Picton, T.W., Hillyard, S.A., Krausz, H.I., Galambos, R., 1974. Human auditory evoked potentials. I: Evaluation of components. Electroencephalogr. Clin. Neurophysiol. 36, 179–190. https://doi.org/10.1016/0013-4694(74)90155-2.
Picton, T.W., John, S.M., Dimitrijevic, A., Purcell, D.W., 2003. Human auditory steady-state responses. Int. J. Audiol. 42, 177–219. https://doi.org/10.3109/14992020309101316.
Plack, C.J., 2014. The Sense of Hearing. Psychology Press, New York, USA.
Presacco, A., Simon, J.Z., Anderson, S., 2016a. Effect of informational content of noise on speech representation in the aging midbrain and cortex. J. Neurophysiol. 116, 2356–2367. https://doi.org/10.1152/jn.00373.2016.
Presacco, A., Simon, J.Z., Anderson, S., 2016b. Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. J. Neurophysiol. 116, 2346–2355. https://doi.org/10.1152/jn.00372.2016.
Presacco, A., Simon, J.Z., Anderson, S., 2019. Speech-in-noise representation in the aging midbrain and cortex: Effects of hearing loss. PLoS ONE 14, e0213899. https://doi.org/10.1371/journal.pone.0213899.
Rentzsch, J., Jockers-Scherübl, M.C., Boutros, N.N., Gallinat, J., 2008. Test–retest reliability of P50, N100 and P200 auditory sensory gating in healthy subjects. Int. J. Psychophysiol. 67 (2), 81–90. https://doi.org/10.1016/j.ijpsycho.2007.10.006.
Resnik, J., Polley, D.B., 2017. Fast-spiking GABA circuit dynamics in the auditory cortex predict recovery of sensory processing following peripheral nerve damage. eLife 6, e21452. https://doi.org/10.7554/eLife.21452.
Salvi, R., Sun, W., Ding, D., Chen, G.-D., Lobarinas, E., Wang, J., Radziwon, K., Auerbach, B.D., 2017. Inner Hair Cell Loss Disrupts Hearing and Cochlear Function Leading to Sensory Deprivation and Enhanced Central Auditory Gain (Article). Front. Neurosci. 10, 621. https://doi.org/10.3389/fnins.2016.00621.
Schmitt, R., Meyer, M., Giroud, N., 2022. Better speech-in-noise comprehension is associated with enhanced neural speech tracking in older adults with hearing impairment. Cortex 151, 133–146. https://doi.org/10.1016/j.cortex.2022.02.017.
Schneider, B.A., Daneman, M., Pichora-Fuller, K.M., 2002. Listening in AGING ADULTS: FROM DISCOURSE COMPREHENSION TO PSYCHOAcoustics. Can. J. Exp. Psychol. 56, 139–152. https://doi.org/10.1037/h0087392.
Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech RECOGNITION WITH PRIMARILY TEMPORAL CUes. Science 270 (5234), 303–304. https://doi.org/10.1126/science.270.5234.303.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull. 86, 420–428. https://doi.org/10.1037/0033-2909.86.2.420.
Signoret, C., Johnsrude, I.S., Classon, E., Rudner, M., 2011. Combined effects of form- and meaning-based predictability on perceived clarity of speech. J. Exp. Psychol.: Hum. Percept. Perform. 44, 277–285. https://doi.org/10.1037/xhp0000442.
Synigal, S.R., Anderson, A.J., Lalor, E.C., 2023. Electrophysiological indices of hierarchical speech processing differentially reflect the comprehension of speech in noise. BioRxiv.
Tervaniemi, M., Kujala, T., Alho, K., Virtanen, J., Ilmoniemi, R.J., Näätänen, R., 1999a. Functional specialization of the human auditory cortex in processing phonetic and musical sounds: a magnetoencephalographic (MEG) study. NeuroImage 9, 330–336. https://doi.org/10.1006/nimg.1999.0405.
Tervaniemi, M., Lehtokoski, A., Sinkkonen, J., Virtanen, J., Ilmoniemi, R.J., Näätänen, R., 1999b. Test-retest reliability of mismatch negativity for duration, frequency and intensity changes. Clin. Neurophysiol. 110, 1388–1393. https://doi.org/10.1016/s1388-2457(99)00108-x.
Tree, J.E.F., 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. J. Mem. Lang. 34 (6), 709–738. https://doi.org/10.1006/jmla.1995.1032.
Tremblay, K.L., Piskosz, M., Souza, P., 2003. Effects of age and age-related hearing loss on the neural representation of speech cues. Clin. Neurophysiol. 114, 1332–1343. https://doi.org/10.1016/s1388-2457(03)00114-7.
Tyler, L.K., Marslen-Wilson, W.D., 2008. Fronto-temporal brain systems supporting spoken language comprehension. Philos. Trans. R. Soc. B 363, 1037–1054. https://doi.org/10.1098/rstb.2007.2158.
Vanthornhout, J., Decruy, L., Wouters, J., Simon, J.Z., Francart, T., 2018. Speech intelligibility predicted from neural entrainment of the speech envelope. J. Assoc. Res. Otolaryngol. 19 (2), 181–191. https://doi.org/10.1007/s10162-018-0654-z.
Vanthornhout, J., Decruy, L., Francart, T., 2019. Effect of task and attention on neural tracking of speech. Front. Neurosci. 13 https://doi.org/10.3389/fnins.2019.00977.
Walhovd, K.B., Fjell, A.M., 2002. One-year test–retest reliability of auditory ERPs in young and old adults. Int. J. Psychophysiol. 46, 29–40. https://doi.org/10.1016/s0167-8760(02)00039-9.
Wasserman, S., Bockenholt, U., 1989. Bootstrapping: applications to Psychophysiology. Psychophysiology 26 (2), 208–221. https://doi.org/10.1111/j.1469-8986.1989.tb03159.x.
Williams, L.M., Simms, E., Clark, C.R., Paul, R.H., Rowe, D., Gordon, E., 2005. The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: "neuromarker. Int. J. Neurosci. 115 (12), 1605–1630. https://doi.org/10.1080/00207450590958475.
Wilson, R.H., McArdle, R., Watts, K.L., Smith, S.L., 2012. The Revised Speech Perception in Noise Test (R-SPIN) in a Multiple Signal-to-Noise Ratio Paradigm. J. Am. Acad. Audio 23 (08), 590–605. https://doi.org/10.3766/jaaa.23.7.9.
Yasmin, S., Irsik, V.C., Johnsrude, I.S., Herrmann, B., 2023. The effects of speech masking on neural tracking of acoustic and semantic features of natural speech. Neuropsychologia 186, 108584. https://doi.org/10.1016/j.neuropsychologia.2023.108584.
Zuk, N.J., Murphy, J.W., Reilly, R.B., Lalor, E.C., 2021. Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. PLOS Comput. Biol. 17 (9), e1009358 https://doi.org/10.1371/journal.pcbi.1009358.