

Language-agnostic, Automated Assessment of Listeners' Speech Recall Using Large Language Models

Trends in Hearing

Volume 29: 1–22

© The Author(s) 2025

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/23312165251347131

journals.sagepub.com/home/tiaBjörn Herrmann^{1,2}

Abstract

Speech-comprehension difficulties are common among older people. Standard speech tests do not fully capture such difficulties because the tests poorly resemble the context-rich, story-like nature of ongoing conversation and are typically available only in a country's dominant/official language (e.g., English), leading to inaccurate scores for native speakers of other languages. Assessments for naturalistic, story speech in multiple languages require accurate, time-efficient scoring. The current research leverages modern large language models (LLMs) in native English speakers and native speakers of 10 other languages to automate the generation of high-quality, spoken stories and scoring of speech recall in different languages. Participants listened to and freely recalled short stories (in quiet/clear and in babble noise) in their native language. Large language model text-embeddings and LLM prompt engineering with semantic similarity analyses to score speech recall revealed sensitivity to known effects of temporal order, primacy/recency, and background noise, and high similarity of recall scores across languages. The work overcomes limitations associated with simple speech materials and testing of closed native-speaker groups because recall data of varying length and details can be mapped across languages with high accuracy. The full automation of speech generation and recall scoring provides an important step toward comprehension assessments of naturalistic speech with clinical applicability.

Keywords

speech recall, recall, memory, multilanguage, generative artificial intelligence, large language models

Received: March 19, 2025; revised: May 11, 2025; accepted: May 16, 2025

Introduction

Many older people live with some form of hearing loss (Cruikshanks et al., 1998; Feder et al., 2015; Goman and Lin, 2016). Peripheral degeneration associated with hearing loss and resulting maladaptive neural plasticity (Auerbach et al., 2014; Herrmann and Butler, 2021; Salvi et al., 2017) impairs speech comprehension in noisy situations, such as in crowded restaurants. This often occurs well before pure-tone hearing thresholds worsen sufficiently for hearing loss to be diagnosed (Helfer and Jesse, 2021; Herrmann and Johnsrude, 2020a; Pichora-Fuller et al., 2016; Pichora-Fuller and Levitt, 2012). Persistent speech comprehension challenges increase the risk of social isolation (Bott and Saunders, 2021; Motala et al., 2024; Shukla et al., 2020) and cognitive decline (Chern and Golub, 2019; Lin and Albert, 2014; Livingston et al., 2020). The accurate assessment of speech-comprehension challenges is thus highly relevant for diagnostic and treatment outcomes.

Although individuals seek audiological help mainly because they have difficulties comprehending speech (McNeice et al., 2024; Portnuff and Bell, 2019), only one-third of audiologists use speech tests regularly (Billings et al., 2024; Fitzgerald et al., 2023; Portnuff and Bell, 2019). Speech testing adds 10–15 min to busy ~1-h audiological assessments (Billings et al., 2024), which is time-consuming and costly without automation. Even when speech tests are administered, they are typically available only in a

¹Rotman Research Institute, Baycrest Academy for Research and Education, North York, Ontario, Canada

²Department of Psychology, University of Toronto, Toronto, Ontario, Canada

Corresponding author:

Björn Herrmann, Rotman Research Institute, North York, Ontario, Canada M6A 2E1.

Emails: bjorn.herrmann@utoronto.ca; bherrmann@research.baycrest.org



country's dominant language. Speakers of other languages perform worse on speech-in-noise tests despite similar hearing function (Garcia Lecumberri et al., 2010; Kilman et al., 2015; Phillips et al., 2024; van Wijngaarden, 2001; van Wijngaarden et al., 2002). The generation of speech materials in different languages is costly because it requires voice actors with different language backgrounds to record speech. The lack of speech materials and ways to score them in different languages makes assessments less accessible, especially in multiethnic countries with a high number of immigrants. Recent advances in AI-based speech synthesizers may open new opportunities (Costa-jussà et al., 2024; van den Oord et al., 2016) because they enable the creation of highly naturalistic-sounding speech materials (Herrmann, 2023) that are increasingly used in research practice (Karunathilake et al., 2025; Herrmann, 2025b; Herrmann, 2025c; Pandey and Herrmann, in press) and may have high applicability in clinical settings (Polspoel et al., 2025). Modern synthesized speech can be generated in different languages, potentially reducing the need for costly recordings by voice actors (Herrmann, 2023; Polspoel et al., 2025; van den Oord et al., 2016).

Existing speech tests consist of simple words/sentences and verbatim word reports (Billings et al., 2024; Killion et al., 2004; Polspoel et al., 2025; Wilson et al., 2012; Wilson Richard et al., 2007) that poorly resemble the context-rich, story-like speech of typical conversations (Bohanek et al., 2009; Eisenberg, 1985; Jefferson, 1978; Mullen and Yi, 1995; Ochs et al., 1992). Moreover, the mental representations of conversations are not verbatim but reflect the gist integrated across multiple sentences and tens of seconds of speech (Fillenbaum, 1966; Gomulicki, 1956; Martinez, 2024; Mehler, 1963; Mehler and Miller, 1964; Panella et al., 2025; Raccach et al., 2024; Sachs, 1967; Schneider et al., 2002). Verbatim word reports with standard speech tests may fail to capture real-life comprehension challenges because individuals can understand the gist and be highly engaged in a conversation despite missing a significant number of words (Herrmann and Johnsrude, 2020b; Irsik et al., 2022). The focus on verbatim reports of words or short sentences may thus disconnect clinic testing from daily experiences, potentially explaining why hearing aid use is relatively low (Bainbridge and Wallhagen, 2014; Chien and Lin, 2013; Davidson et al., 2021). However, speech tests for naturalistic, story-like speech have thus far not been feasible because they would add ~30 min of manual scoring that audiologists cannot afford. Automating speech generation and speech-comprehension scoring in different languages could advance audiological practice and make assessments of hearing function more accessible.

Recent research suggests that scoring free recall data can be automatized using modern data-science approaches (Georgiou et al., 2025; Heusser et al., 2021; Martinez, 2024; Raccach et al., 2024; Shen et al., 2023). For example, prompt engineering with large-language models (LLMs)

(Chen et al., 2024) has been used to score the number of details a person recalls from short passages with a predefined number of details (Georgiou et al., 2025; Martinez, 2024), but it is unclear whether this is scalable to longer more naturalistic materials for which the details are not defined a priori. Another approach is to use either hidden Markov models or, more recently, LLMs to obtain text-embeddings as the basis for analysis (Chandler et al., 2021; Herrmann, 2025a; Heusser et al., 2021; Martinez, 2024; Raccach et al., 2024; Shen et al., 2023). Text-embeddings are high-dimensional vectors that represent the semantic meaning of text pieces (Cer et al., 2018; Devlin et al., 2019; Feng et al., 2022; Feuerriegel et al., 2025; Mikolov et al., 2013; Pennington et al., 2014). For example, the embedding vectors for two semantically related words, such as “sock” and “foot,” correlate higher than the embedding vectors for two semantically unrelated words, such as “sock” and “pool.” Correlations between embedding vectors can thus be used to assess the semantic similarity between segments of the speech materials and a person's recall. The approach enables assessing critical comprehension metrics, including recall magnitude, temporal order recall, and benefits from primacy and recency information in the stimulus materials (Heusser et al., 2021). However, the extent to which it can be used for spoken speech recall across languages is unclear.

The current study expands initial work (Heusser et al., 2021; Panella et al., 2025; Raccach et al., 2024; Shen et al., 2023) to explore a completely automated approach to assess speech comprehension of short stories under different speech-clarity conditions (speech in quiet/clear and in 12-talker babble) across 11 different languages. The work uses LLMs to generate stories, translate them into different languages, synthesize auditory stories in different languages, and automatically score speech recall in a listener's own native language (in which they also heard the stories). The approach proposed here proves to be sensitive to known comprehension effects (speech clarity, temporal order, primacy/recency) and generalizes to speakers of different languages. The current work has the potential to open new avenues for an accessible clinical practice that focuses on naturalistic speech listening.

Methods and Materials

Participants

Two groups of participants took part in the current study. One group included 29 native English speakers or individuals who grew up in English-speaking countries (mostly Canada) and have been speaking English since early childhood (<5 years of age) (mean age: 22.9 years, age range: 19–34 years, 6 male, 21 female, 1 transgender male, 1 nonbinary). The other group comprised 26 native speakers of not only a different language (mean age: 22.9 years, age range: 18–37 years, 7 male, 17 female, 1 genderfluid, 1 preferred not to answer) but who also spoke English sufficiently to understand and

comply with task instructions presented in English. Participants of this group were Danish ($N=2$), French ($N=3$), German ($N=1$), Italian ($N=1$), Polish ($N=1$), Portuguese ($N=2$), Russian ($N=4$), Spanish ($N=9$), Tagalog ($N=2$), and Ukrainian ($N=1$) native speakers. For simplicity, the two groups are referred to as English speakers and non-English speakers, respectively. The number of participants resulted from the availability of participants with different language backgrounds and is consistent with previous in-person speech-perception work observing meaningful effects with less than 30 participants per group (Ferguson et al., 2010; Gianakas Steven and Winn Matthew, 2025; Gordon et al., 2009; Peelle and Wingfield, 2005; Regev et al., 2025; Tsuji and Arai, 2025).

All participants reported having normal hearing abilities and no neurological disease. Participants gave written informed consent prior to the experiment and received \$20 per 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 (REB #23-11).

Data from three additional participants were recorded but excluded from the analysis. One English speaker reported having a neurological disease. One non-English speaker indicated that their first language is Polish, but they spoke English growing up and felt not very comfortable in Polish anymore. Another non-English speaker indicated that they are fluent in Lithuanian but identified as a native English speaker.

Speech Materials

Speech materials were auditory stories of about 2-min duration. The aim of the current study was to fully automate stimulus generation and data analysis approaches to test the feasibility of examining speech comprehension in individuals with different language backgrounds. Hence, stories were generated using generative AI for text generation and AI-based text-to-speech synthesizers (Georgiou et al., 2025; Herrmann, 2025c).

In detail, OpenAI's GPT-4o (v2024-05-13; OpenAI et al., 2023) was used to generate seven story texts (Georgiou et al., 2025; Herrmann, 2025c), one for practice and six for the main experiment. Each story text focused on a different topic (e.g., a boy stealing and being caught; a woman becoming a construction worker). The model was prompted using a zero-shot approach (Chen et al., 2024) with the following input: "Generate a 300 word engaging story about [story topic was inserted here]. The story should not start with "once upon a time" and should not contain quotations, dashes, colons, or semicolons. Generate a related story title." The number of words in the stories ranged between 314 and 355 (English).

Each story text was concatenated into a single paragraph for stimulus generation and analysis.

Auditory stories were generated from the story texts using OpenAI's Whisper (model: "tts-1-hd"; Radford et al., 2022) through OpenAI's application processing interface (API) using Python code. Whisper is an automatic speech recognition system that was trained on 680,000 h of multilingual and multitask supervised data collected from the internet (Radford et al., 2022). Whisper supports text-to-speech and speech-to-text applications in different languages. To generate auditory stories in English, each story text was directly fed into the Whisper API. To generate auditory stories in other languages, OpenAI's GPT-4o (v2024-05-13; API; OpenAI et al., 2023) was used to translate each story, using a zero-shot prompt (Chen et al., 2024): "Translate the following text to [language].;" followed by the story text. The translated text was fed into Whisper. Piloting with colleagues in the lab suggested accurate translation and high auditory quality for several languages. However, piloting also suggested that Asian languages, such as Cantonese, Mandarin, Bengali, and Hindi are more poorly converted to auditory stories using Whisper (despite accurate translations). The current study therefore focused on languages for which we were confident that the audio quality was high. The Whisper voice "alloy" was used to generate all stories, because this voice is naturalistic sounding and gender neutral. Indeed, individuals experience modern AI speech as very naturalistic, and speech perception in quiet and noise is highly similar for AI and human speech (Herrmann, 2023). The "alloy" voice was used for the synthesis of all stories in different languages, having the advantage that acoustic confounds are minimized that would arise if different voice actors had recorded the stories. Whisper generates audio files at a 24-kHz sampling frequency. Each story was up-sampled to a 44.1-kHz sampling frequency to enable use with the sound-delivery system.

To investigate the sensitivity of the recall scoring approach to different speech listening conditions, each participant was presented with three stories under clear conditions and three stories with added 12-talker babble (Bilger, 1984; Bilger et al., 1984) at a signal-to-noise ratio (SNR) of +2 dB. Twelve-talker babble simulates a crowded restaurant, while not permitting identification of individual words in the masker (Bilger et al., 1984; Mattys et al., 2012). Young normal-hearing listeners correctly hear about 80% of words of continuous speech masked by a 12-talker babble at +2 dB SNR (Irsik et al., 2022).

For each participant group (English, non-English), half of the participants listened to the stories with the topics "A boy being surprised," "A man going camping for the first time," and "A woman becoming a construction worker" under clear conditions (i.e., in quiet) and to the stories with the topics "A boy stealing and being caught," "A woman moving into a shared flat," and "A woman surprising her grandma" in babble. For the other half of the participants, the assignment of

speech-clarity conditions (clear, babble) was reversed; counterbalanced as much as possible across languages.

Experimental Procedures

Participants sat in a comfortable chair inside a single-wall sound booth (Eckel Industries). Experimental procedures were run using Psychtoolbox (v3.0.14) in MATLAB (MathWorks Inc.) on a Lenovo T450 s laptop with Microsoft Windows 7. The laptop screen was mirrored to a ViewSonic monitor within the sound booth. In each of the six blocks, participants listened to one 2-min story and afterward freely recalled the story. Participants were instructed to recall as many details as possible. Story order was randomized for each participant. Participants were familiarized with the task procedures in one training block with a separate 2-min story, prior to the six experimental blocks. Instructions for all participants were in English, and all participants indicated they understood the instructions (they were all proficient English speakers). Critically, stories were played to people in their native language, and participants were instructed to recall the story in their native language. Auditory stories were presented at a comfortable listening level that was fixed across participants (~70 dB SPL) via Sony Dynamic Stereo MDR-7506 headphones and an RME Fireface 400 external sound card. Participants' free recall was recorded via a Shure SM7B microphone.

Language and Speech-Quality Ratings

Following the main experimental procedures, participants rated their experience (Likert scale from 1 to 5) with the spoken story materials (exact statements are described in Table 1). Ratings were obtained only for a subset of participants (19 English speakers; 18 non-English speakers) because the procedures were implemented after the study had already started. English and non-English speakers did not differ in their ratings of the overall impression of the stories, the naturalness of the spoken speech, nor the accurateness of the grammar (for all $p > 0.1$; Table 1). Non-English speakers rated speech understanding, the accurateness of the speech's meaning, and vocabulary usage slightly lower than English speakers (for both $p < 0.05$; Table 1), but the rating scores were relatively high for both groups. The most notable difference was that non-English speakers rated the speech pronunciation lower than English speakers and that they more frequently felt the speech was accented (for both $p < 0.001$; Table 1).

The results of the ratings are also reflected in qualitative reports from participants during debriefing. Non-English speakers reported that they liked the materials and could understand the speech, but that occasionally a word was presented in a context in which it would not typically occur (relating to the translation quality), and that the speech sometimes sounded like a person speaking with an English accent (relating to speech synthesis quality). The English accent of

OpenAI voices in non-English languages has been noted previously in developer forums (e.g., <https://community.openai.com/t/tts-voices-have-a-clear-us-accent/705394>) because the model was optimized for English (<https://platform.openai.com/docs/guides/text-to-speech>).

The group differences in ratings and qualitative reports highlight existing limitations of modern speech synthesis in different languages. Critically, the current study aims to automate the scoring of spoken recall of native speakers of different languages and is thus not directly concerned with the quality of the speech materials. Although differences in speech synthesis quality between groups could in principle influence recall performance, this does not seem to be the case in the current study (as described below). Nevertheless, potential limitations arising from group differences in rating scores will be detailed in the discussion section.

Preprocessing of Recall Data

Verbal recall data were converted to text using OpenAI's Whisper speech-to-text model (whisper-1) using the OpenAI API through Python. As outlined above, Whisper is a multilanguage model and can provide transcriptions for speech in different languages. For some transcriptions, punctuation was largely absent, because the model appeared to interpret the audio signal as one stream of thought. OpenAI's GPT-4o (v2024-05-13; temperature = 0; OpenAI et al., 2023) was used to correct punctuation using the following prompt: "The following text has no or minimal punctuation. Please repair missing punctuation where it seems appropriate. Use only periods, commas, and question marks. Avoid all other punctuation. Be moderate and do not overdo it. Do not add, replace, or remove any words from the text. Do not insert line breaks. Add punctuation, but otherwise reproduce the text exactly. This is the text for which punctuation needs to be repaired:" followed by the transcription. Punctuation correction was carried out for all recall data to avoid manual evaluation. A qualitative assessment suggested that when punctuation was already present, GPT-4o did not add unnecessary punctuation. More generally, the presence or absence of punctuation had little impact on recall analyses. The punctuation-corrected transcriptions in the original language were used as the recall data for analysis. That is, all story and recall texts used for analysis were in the language in which a person heard and recalled the stories. Analyses were also conducted for recall transcriptions translated to English using OpenAI's GPT-4o, but the results were largely comparable (Figures S1 and S2 in the Supplementary Materials).

Analysis of recall data was based on and extended different previous approaches that aimed to automate the analysis of free recall data (Georgiou et al., 2025; Heusser et al., 2021; Panella et al., 2025; Raccach et al., 2024; Shen et al., 2023). The approach implemented here uses modern LLMs and enables analyses across languages that were not possible with

Table 1. Ratings of Speech Materials.

Statement	English speakers	Non-English speakers	Statistic
My overall impression of the stories was good. (‘strongly disagree’ ... ‘strong agree’)	4.15 ± 0.49	3.83 ± 0.69	$t_{35} = 1.617, p = 0.115$
The spoken speech sounded naturalistic. (‘strongly disagree’ ... ‘strong agree’)	3.68 ± 0.98	3.05 ± 1.22	$t_{35} = 1.685, p = 0.101$
I could understand everything (when there was no background noise). (‘strongly disagree’ ... ‘strong agree’)	4.68 ± 0.57	4.17 ± 0.76	$t_{35} = 2.284, p = 0.029$
The pronunciation of the spoken speech was ... (‘Poor’ ... ‘Excellent’)	4.58 ± 0.67	3.06 ± 0.91	$t_{35} = 5.643, p < 0.001$
The language was grammatically/syntactically accurate. (‘Not at all’ ... ‘No mistakes’)	4.53 ± 0.99	4.0 ± 0.94	$t_{35} = 1.607, p = 0.117$
The meaning of the words and sentences was accurate. (‘Not at all’ ... ‘No mistakes’)	4.68 ± 0.57	4.17 ± 0.76	$t_{35} = 2.284, p = 0.029$
The vocabulary usage was ... (‘Very poor vocabulary, inappropriate or incorrect usage’ ... ‘Excellent vocabulary usage, fully appropriate’)	4.53 ± 0.59	3.72 ± 0.87	$t_{35} = 3.206, p = 0.003$
The spoken speech sounded accented. (‘Yes,’ ‘No’)	$p(\text{yes}) = 0.052$	$p(\text{yes}) = 0.889$	$\chi^2 = 22.76, p < 0.001$

Left column: The statement that participants (19 English speakers; 18 non-English speakers) rated on a 5-point scale (scale end points are indicated in parentheses) and the statement for which they answered “yes” or “no.” Middle columns: Mean and standard deviation for each statement and speaker group (English, non-English). Right column: Test statistics showing t-values (or chi-squared value for the “yes” vs. “no” question) and p-values. The text indicated in bold indicates a significant difference between English and non-English speakers ($p < 0.05$). Note that the identical descriptive and test statistics for the statements “The meaning of the words and sentences was accurate” and “I could understand everything” are not mistakes. The values for individual participants differ between the two statements; only the summary statistics happened to be identical.

previously used models. Analyses involve mapping the story texts and recall transcriptions onto high-dimensional numerical vectors (i.e., text-embeddings) to capture the semantic meaning of the texts (Cer et al., 2018; Devlin et al., 2019; Mikolov et al., 2013; Pennington et al., 2014; Radford et al., 2019).

Extensive previous work shows that individuals perceive, encode, and recall dynamic, naturalistic environments as discrete events or segments (Kurby and Zacks, 2008; Kurby and Zacks, 2018; Sargent et al., 2013; Sasmita and Swallow, 2022; Speer et al., 2004). Hence, each story text and recall transcription (both in the language spoken by the participant) was divided into 10 equally sized segments (based on the number of words in a story) with a length that achieved 20% overlap (Figures 1A and 1B). Segmentation also mitigates potential semantic losses or dilution resulting from an analysis that aims to capture the semantic space of the entire story by compressing it into a single embedding vector (Zhou et al., 2024). The 20% overlap was used to minimally smooth recall data over time and avoid artificial text breaks. Ten segments were selected as a compromise between temporally resolved recall scoring and a high number of short segments that may not capture recall appropriately. The impact of the number of segments was analyzed explicitly as described below. Please note that the results from the segmentation of stories and recalls into equally sized text segments are mostly comparable to the results from automated event segmentation based on meaningful units within a story/recall using GPT-4o with prompt engineering (Michelmann et al., 2025;

Panella et al., 2025; Figures S3 and S4 in the Supplementary Materials).

Each segment of the story texts and recall transcriptions was mapped onto one 768-dimensional embedding vector using the Language-agnostic BERT Sentence Embedding (LaBSE) model (Feng et al., 2022), where BERT refers to Google’s Bidirectional Encoder Representations from Transformers (Devlin et al., 2019). Language-agnostic BERT Sentence Embedding is a multilingual text-embedding model optimized for over 100 languages and is freely available through the Huggingface platform (<https://huggingface.co/sentence-transformers/LaBSE>). The model was imported into Python through the sentence-transformers module. Running the LaBSE model for each story segment and transcription segment resulted in 10 embedding vectors for each of the six story texts, and 10 embedding vectors for each of the six recall transcriptions per participant. The embedding vectors were used to calculate different metrics that assess recall performance.

Automated Scoring of Recall Data

For each participant and story, all-to-all Spearman correlations were calculated between the 10 embedding vectors reflecting the 10 story-text segments and the 10 embedding vectors reflecting the 10 recall-text segments (Figure 1C). Spearman correlation rather than Pearson’s correlation, cosine similarity, or Euclidean distance was used because it is less sensitive to outliers, if there were any (Herrmann, 2025a;

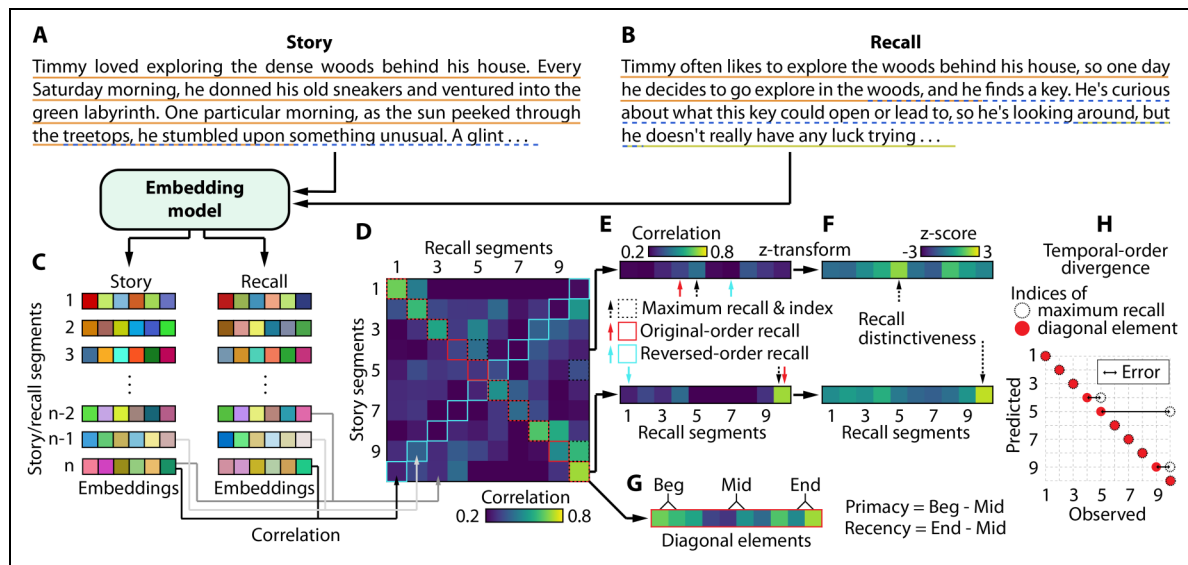


Figure 1. Schematic and Sample Data Displaying the Calculation of Recall Metrics. (A and B) Sample Story and Recall Texts from One English Speaker. The Colored Lines Below the Text Indicate the Story and Recall Segments, Dividing Both the Story and the Recall into 10 Equally Sized Segments (Based on the Number of Words; Separately for Story and Recall Texts) with 20% Overlap. (C) Story and Recall Segments are Fed into a Text-Embedding Model (LaBSE), Resulting in One High-Dimensional, Text-Embedding Vector for Each Story and Recall Segment; Here Schematically Shown, with Colors Symbolizing the Magnitude of Embedding Dimensions. (D) The Story x Recall Matrix Shows the Spearman Correlations (Transformed to Fisher's Z Correlations) Between the Embedding Vectors of the Story Segments and the Embedding Vectors of the Recall Segments for One Sample Participant. The Colored and Line-Dashed Square Boxes within the Matrix Indicate the Different Recall Metrics Specified in Panel E. (e) Several Recall Metrics can be Calculated from the Story x Recall Matrix. The "Maximum Recall Score" for a Story Segment is Calculated as the Maximum Correlation in the Corresponding Row of the Matrix. The Index Corresponding to the Maximum Correlation can be Used to Calculate a "Temporal-Order Divergence Score" Represented in Panel H. the "Original-Order Recall Scores" are the Correlation Values Along the Diagonal of the Story x Recall Matrix. This Score is High if a Participant Recalled the Story in the Original Temporal Order. The "Reversed-Order Recall Scores" are the Correlation Values Along the Reversed Diagonal of the Story x Recall Matrix. These Values can be Used as a Control Relative to the "Original-Order Recall Scores." (F) The "Recall Distinctiveness" is the Maximum Value of a Row of the Story x Recall Matrix After Standardizing the Correlation Values within the Row. A High Recall Distinctiveness Score Indicates that a Story Segment Correlated Strongly with One Specific Recall Segment, Rather than Being Similarly Related to Multiple Segments. (G) Calculation of Recall Scores for the Analysis of Primacy and Recency Effects by Averaging the First Two Diagonal Correlation Values (Primacy) and the Last Two Values (Recency) and Comparing them to the Average Across the Two Values in the Middle. (H) The Index Corresponding to the Recall Segment that Showed the Highest Correlation with a Story Segment ("Maximum Recall Score") can be Used to Calculate the "Temporal-Order Divergence Score" as the Root Mean Squared Error Between the Predicted Temporal Order (1, 2, 3, ..., N-1, N) and the Recalled Temporal Order. The Outlined Approach in Panels A-H can also be Used to Calculate the Similarity of Participants' Recall by Calculating Participant Recall x Participant Recall Matrices.

Panella et al., 2025; Zhelezniak et al., 2019). The correlation values of the resulting 10 story segments by 10 recall segments correlation matrix were Fisher Z-transformed (Figure 1D). A high value in the correlation matrix means that the semantic content of a story segment is highly similar to the semantic content of a recall segment. "Control" correlation matrices were calculated by pairing a specific story separately with the recall of the other five stories from the same participant. Overall, the procedure resulted in one story x recall matrix and five control story x recall matrices for each participant and for each of the six stories (three clear stories; three stories in background babble).

Specific recall metrics were derived from these story x recall correlation matrices (metrics for the five control matrices were subsequently averaged). In detail, for each story segment (i.e., row of the story x recall correlation matrix), the

"maximum recall score" was taken as the maximum Fisher's Z correlation across the 10 recall segments (Figure 1E). The result was one maximum recall score for each story segment. Scores were averaged across the 10 story segments to obtain one maximum recall score reflecting the participant's overall recall for the specific story. A "recall-distinctiveness score" was calculated by z-transforming (standardizing) the Fisher's Z correlation values within a row (i.e., a story segment) of the story x recall correlation matrix and taking the maximum z-score (Figure 1F; Heusser et al., 2021). A high recall distinctiveness score indicates that a story segment correlated strongly with one specific recall segment, whereas a low value indicates it is similarly related across multiple of the 10 recall segments (Heusser et al., 2021). The index corresponding to the recall segment that showed the highest correlation with a story segment (i.e.,

the index of the maximum recall score) can be used to calculate a “temporal-order divergence score” (Figure 1H). To this end, for each story segment (i.e., row in the story \times recall matrix), the index corresponding to the highest correlation was extracted. The root mean squared error was then calculated between the original temporal order (i.e., 1, 2, 3, ..., $n-1$, n) and the recalled temporal order as indicated by the extracted indices. The root mean squared error was taken as the “temporal-order divergence score” (Figure 1H).

Previous research suggests that individuals tend to recall events in the temporal order in which they were experienced (Chen et al., 2017; Heusser et al., 2021; Howard and Kahana, 2002; Zhang et al., 2023). The diagonal elements in the story \times recall correlation matrix reflect the recall scores for story segments recalled in the original temporal order, that is, the order in which the story was played. Hence, for each story, the correlation values along the diagonal were averaged to obtain one “original-order recall score” for each story (Figure 1D and E). To examine whether participants indeed recalled a story in the original temporal order, the correlation values along the reversed diagonal were averaged to obtain one “reversed-order recall score” for each story (i.e., assuming that participants recalled the last story segment first and the first segment last) (Figure 1D and E). A higher original-order recall than reversed-order recall score would indicate that participants recalled a story in the temporal order in which the story was heard.

Previous memory research also highlights that information in the beginning (primacy) and at the end (recency) is better recalled than information in the middle of the stream of information (Murdock Jr, 1962; Howard and Kahana, 1999; Howard and Kahana, 2002; Tan and Ward, 2000; Zhang et al., 2023). To examine whether primacy plays a role in story recall, the first two scores of the “original-order recall scores” (i.e., the diagonal elements of the correlation matrix) were averaged and compared to the two averaged scores in the middle. The effect of recency was examined by averaging the last two scores of the diagonal elements and comparing them to the two averaged scores in the middle (Figure 1G).

All recall metrics—maximum recall, original-order recall, reversed-order recall, recall distinctiveness, temporal-order divergence, primacy, recency—were calculated for each participant and story (three clear; three in background babble). Recall metrics were averaged separately across the three clear stories and the three stories in background babble. To assess the degree to which meaningful information about story recall is represented in the story \times recall matrices (independent of speech clarity), two repeated-measures analysis of variance (rmANOVA) were calculated, using the within-participant factor Score Type (analysis 1: maximum recall vs. chance maximum recall; analysis 2: original-order recall vs. reverse-order recall; collapsed across speech-clarity levels) and the between-participant factor Speaker (English, Non-English). To assess the impact of background babble on story recall, an rmANOVA was calculated with the within-

participant factor Speech Clarity (clear, babble) and the between-participant factor Speaker (English, Non-English), separately calculated for the original-order recall score, recall distinctiveness, temporal-order divergence. To assess primacy and recency effects, the rmANOVA included the additional within-participant factor Time (beginning, middle, end). Post hoc comparison was calculated in case of significant effects, using Holm’s methods to correct for multiple comparisons (Holm, 1979). Statistical analyses were conducted in JASP software (JASP, 2024; version 0.19.1.0).

Intersubject Correlation of Recall

The recall metrics described in the previous section reflect how well the semantic content of a participant’s recall corresponds to the semantic content of a story. In order to investigate how well story recall in different languages agrees across participants, recall agreement metrics were calculated based on previously established intersubject correlation approaches (ISC; Dmochowski et al., 2012; Hasson et al., 2004; Hasson et al., 2008; Hasson et al., 2010; Irsik et al., 2022; Lee and Chen, 2022; Nastase et al., 2019; Nguyen et al., 2019). Calculations were similar to those outlined in Figure 1, with the exception that correlations were calculated between the recall transcriptions of two participants rather than between the story text and the recall transcription of a participant. Recall transcriptions were maintained in their original languages and were not translated to a common language prior to analysis.

In detail, for each participant, the all-to-all Spearman correlations (Fisher’s Z transformation) were calculated between the 10 embedding vectors corresponding to one participant’s recall and the 10 embedding vectors corresponding to another participant’s recall, leading to one recall \times recall matrix (using the transcription of the recall in each participant’s native language). For each participant, one correlation matrix was calculated relative to each of the other participants listening to the same story under the same speech clarity condition (clear or babble). For example, 14 English and 14 non-English speakers listened to the story “A boy being surprised” in babble noise. For one English speaker out of these 28 participants, 27 recall \times recall correlation matrices were calculated, 13 relating to the other English speakers and 14 relating to the non-English speakers. For each recall \times recall correlation matrix, we calculated the mean across the diagonal elements (“original-order score”; Figure 1E)—reflecting how much participants recalled the story in a similar temporal order—and the mean across the reversed-diagonal elements as a control (“reversed-order score”; Figure 1E). The “temporal-order divergence” was also calculated for each recall \times recall correlation matrix to assess the degree to which participants diverged from recalling the story in the same temporal order (Figure 1H). In the specific example from a few sentences above, this led to 13 scores per recall metric associated with the participant’s “own” group and 14 scores of the participant

relative to individuals from the “other” group. These scores were averaged separately for the participant’s “own” and “other” group, leading to two scores per metric. Note that “group” refers to individuals defined above as belonging to either the English speakers or the non-English speakers (and not unique languages).

To assess whether recall ISC is significant (independent from speech clarity), scores were averaged across speech-clarity conditions, and a rmANOVA was calculated using the within-participant factors Score Type (diagonal vs. reversed diagonal) and Reference Group (own, other), and the between-participant factor Speaker (English, non-English). To assess the impact of background babble, two rmANOVAs were calculated, one for ISC (diagonal of the recall \times recall matrix), and one for temporal-order divergence, with the within-participant factors Speech Clarity (clear, babble) and Reference Group (own, other), and the between-participant factor Speaker (English, Non-English). Post hoc comparisons were calculated in the case of significant effects, using Holm’s methods to correct for multiple comparisons (Holm, 1979).

Recall Scoring Using Similarity Ratings Through Prompt Engineering

Recall scoring through correlations between embedding vectors for story text segments and recall text segments has been used most for automated recall scoring (Heusser et al., 2021; Panella et al., 2025; Racciah et al., 2024; Shen et al., 2023; sometimes cosine similarity is used instead of correlation), and the current work expands the approach by using LLMs, novel metrics, and, most critical, scoring across languages. The embedding approach is computationally fast and is sensitive to experimental manipulations (see below). However, correlations between embedding vectors from multiword text segments can lead to a small dynamic range of values. Moreover, scores are rarely close to zero, even when embedding vectors from unrelated stories are correlated (e.g., 0.2; see below). For clinical purposes, scores should ideally be easily interpretable (meaningful dynamic range) and a low recall should lead to a score close to 0.

Previous work has explored prompt engineering with modern LLMs for shorter stories, such that the LLM provides a rating or accuracy score (Georgiou et al., 2025; Martinez, 2024). The current work extends this work to explore whether LLM prompt engineering provides a greater dynamic range of values and a score close to zero for a recall segment that is unrelated to a story segment. The following zero-shot prompt (Chen et al., 2024) was used to obtain a rating score between 0 and 100 for each of the 10×10 combinations of story text segments and recall text segments:

Act as an expert rater of speech contents. On a scale from 0 to 100, rate the degree to which a text segment from a story is

semantically captured by the text segment from a human who recalled the story. 0 refers to “not very” and 100 refers to “very.” If the recall segment mostly does not capture the story segment, give a 0. If the recall segment mostly captures the story segment, give a 100. This is the text segment from the story: “[story segment is inserted here].” This is the text segment from the human recall: “[recall segment is inserted here].” Please provide only the rating score and nothing else. Use the full range of the scale.

An LLM rating value for each story segment and recall segment combination was obtained three times and then averaged to reduce noise. Scores were divided by 100 to obtain scores akin to proportion correct responses.

The prompt was provided to OpenAI’s “gpt-4o-mini-2024-07-18” model (temperature = 0). The “mini” GPT model was used instead of GPT-4o (v2024-05-13) because each prompt fed to the model is associated with costs that were about 17 times cheaper for the “mini” GPT model. Calculating the 10×10 matrices and corresponding chance-level matrices (i.e., relating a story segment to a recall segment of a different story) required 10,800 prompt requests, costing about \$1.5 USD per participant for the “gpt-4o-mini-2024-07-18” model (the GPT-4o model would have led to >\$1200 overall). The practical implications of the approach will be discussed in the discussion section.

After obtaining the story \times recall matrices, calculation of recall metrics and statistical analyses mirrored the procedures described above for the embedding approach (Figure 1).

Intersubject correlation was not calculated for the GPT rating approach due to the high monetary and computational costs (several hours per participant) associated with the prompting approach for ISC (>30,000 prompt request), but pilot testing suggested similarly meaningful sensitivity to experimental manipulations. Practical solutions for future research will be discussed in the discussion section.

Evaluation of the Effect of Segment Numbers

The analyses described above, and the related results presented below are based on the segmentation of story texts and recall transcriptions into 10 segments. To evaluate whether the number of segments by which story texts and recall transcriptions are divided impacts the sensitivity of the automated assessment approach, calculations and analyses were carried out for 6, 10, 14, and 18 segments. Statistical assessments included similar analyses as described above, focusing on speech-clarity contrasts to assess the current approach’s sensitivity.

Results

Recall Accuracy: Embedding Approach

Figure 2A shows the story \times recall correlation matrices for English and non-English speakers. Highest values are present

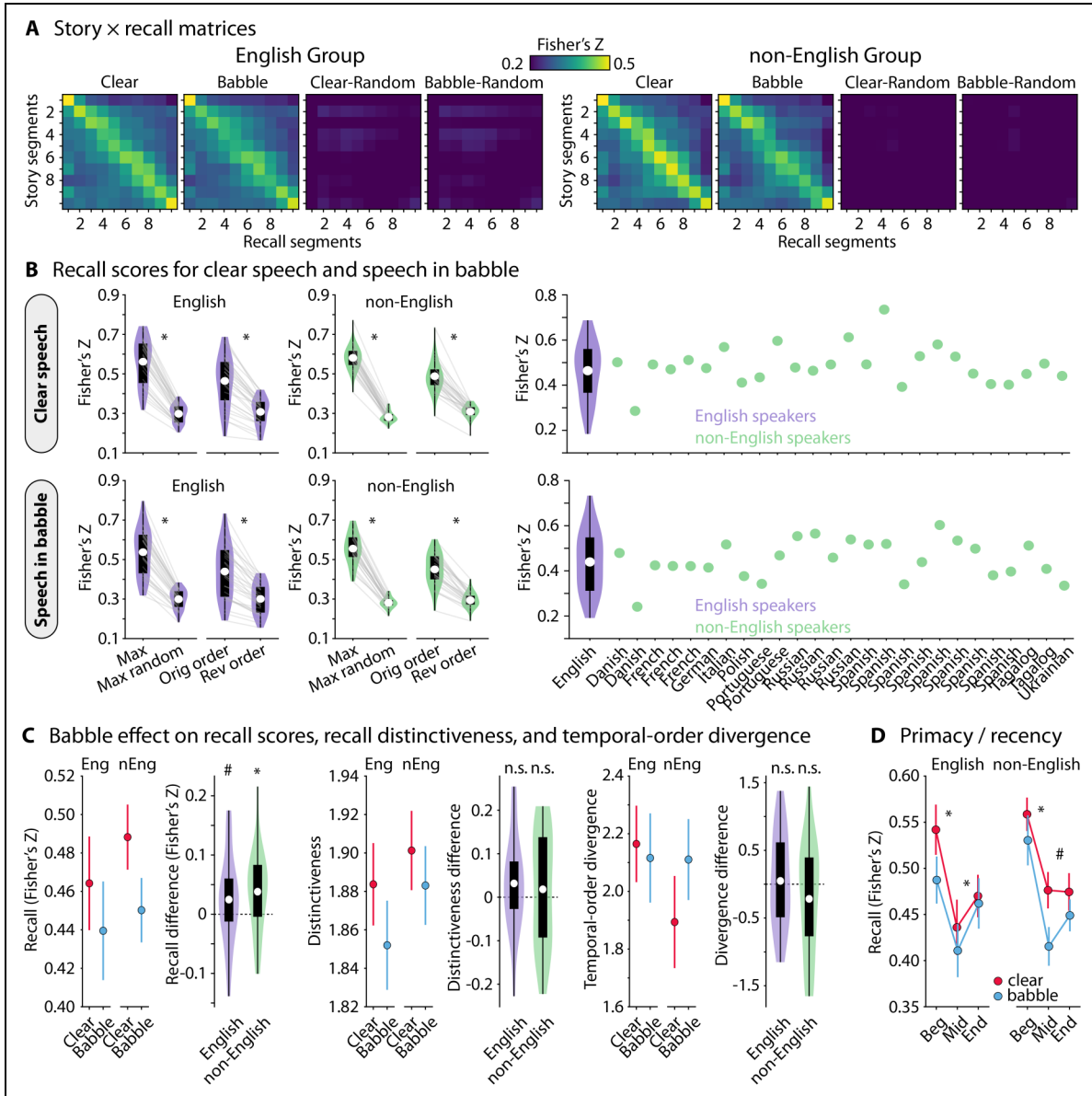


Figure 2. Story × Recall Correlation Matrices and Recall Scores (Embedding Approach). (A) Matrices (Averaged Across Participants) are Shown for Clear Stories and Stories in Babble and for the Actual Recall Scores and Chance Level (Recall Paired with an Unrelated Story). (B) Recall Scores for Clear Speech and Speech in Babble. Violin Plots, Boxplots, and the Mean Across Participants (White Circle) are Shown for the Maximum-Recall Score, Chance Maximum-Recall Score (Recall Paired with Unrelated Stories), Original-Order Recall Score, and Reverse-Order Recall Score. The Right Plot Shows Original-Order Recall Scores for Non-English Speakers and the Distribution for English Speakers, to Display the Overlap Between Speaker Groups. (C) Clear Speech Versus Speech in Babble Contrasts for Original-Order Recall Score (Left), Recall Distinctiveness (Middle), and Temporal-Order Divergence (Right). Violin Plots Show the Difference Between Clear Speech and Speech in Babble. (D) Recall Scores for the Beginning, Middle, and End of a Story, Assessing Primacy and Recency Effects. Eng – English Speakers; nEng – Non-English Speakers; N.s. – Not Significant, * $p < 0.05$, # $p < 0.1$.

along the diagonal of the matrix, indicating that participants recalled the stories in the original temporal order.

Repeated-measures ANOVAs revealed that the maximum-recall score was greater than the chance recall score (i.e., maximum-recall score for unrelated stories), and that the original-order score was greater than the reverse-order score (effect of Score Type; for both analyses: $F_{1,53} > 300$,

$p < 0.001$: Figure 2B), showing that there is meaningful information in the correlation matrices and that stories are recalled in the original temporal order. There was no effect of Speaker (for both analyses $p > 0.7$). There was a Score Type × Speaker interaction for the analysis contrasting maximum-recall score to chance recall ($F_{1,53} = 4.612$, $p = 0.036$, $\omega^2 = 0.012$) because the difference between the maximum-recall score and chance

level was slightly smaller for English than non-English speakers.

To provide a visual impression of the similarity of the recall scores across languages, the right-hand side of Figure 2B shows individual recall scores (original-order recall) for each non-English speaker and the distribution of scores for English speakers. All scores for non-English speakers fall within the distribution of scores for English speakers, confirming visually the absence of a difference between speakers. The results suggest that data are comparable across languages.

Figure 2C shows the effects of speech clarity (clear, babble) for different recall metrics and participant groups. The rmANOVA for the original-order recall revealed larger recall scores for stories under clear than under babble conditions (effect of Speech Clarity; $F_{1,53} = 10.314$, $p = 0.002$, $\omega^2 = 0.017$; Figure 2C, left). There was no difference between speakers and no interaction (for both $p > 0.4$). There were no effects for recall-distinctiveness and temporal-order divergence (for all $p > 0.1$; Figures 2C, middle and right).

The rmANOVA to investigate primacy and recency effects revealed again higher recall scores for clear stories than stories in babble (effect of Speech Clarity; $F_{1,53} = 13.011$, $p = 6.9 \cdot 10^{-4}$, $\omega^2 = 0.020$; Figure 2D). Recall scores were also greater for the beginning than the middle of a story (primacy: $t_{53} = 8.776$, $p_{\text{Holm}} = 2 \cdot 10^{-11}$, $d = 0.737$) and the end than the middle of a story (recency: $t_{53} = 3.836$, $p_{\text{Holm}} = 3.4 \cdot 10^{-4}$, $d = 0.227$; effect of Time: $F_{2,106} = 50.622$, $p = 3.7 \cdot 10^{-16}$, $\omega^2 = 0.104$; Figure 2D). No other main effects or interactions were significant (for all $p > 0.15$).

Intersubject Correlation: Embedding Approach

Figure 3A shows the ISC matrices. The rmANOVA revealed greater scores for diagonal elements than reversed-diagonal elements (effect of Score Type: $F_{1,53} = 1584.941$, $p = 3.6 \cdot 10^{-41}$, $\omega^2 = 0.666$; Figure 3B), showing that participants tend to recall a story in a similar temporal order. Correlations between recalls for speakers of their own group were greater than correlations between recalls with speakers from the other group (effect of Reference Group: $F_{1,53} = 37.458$, $p = 1.2 \cdot 10^{-7}$, $\omega^2 = 0.032$). There was no overall difference between English and non-English speakers (effect of Speaker: $F_{1,53} = 0.771$, $p = 0.384$, $\omega^2 < 0.001$), but the Reference Group \times Speaker, Score Type \times Reference Group, and the Score Type \times Reference Group \times Speaker interactions were significant (for both $F_{1,53} > 9$, $p < 0.005$, $\omega^2 > 0.0005$). Calculating the Score Type \times Reference Group rmANOVA separately for English and non-English speakers revealed an interaction only for non-English speakers ($F_{1,53} = 18.068$, $p = 5.6 \cdot 10^{-4}$, $\omega^2 = 0.014$), but not English speakers ($F_{1,53} = 0.082$, $p = 0.776$, $\omega^2 < 0.001$). For English speakers, both the diagonal and reversed-diagonal scores were greater for the own than the other group (for both $p_{\text{Holm}} < 0.001$), whereas for non-English speakers, this was only the case for the diagonal ($p_{\text{Holm}} = 0.023$), but not reversed-diagonal scores (p_{Holm}

$= 0.479$). Critically, English and non-English speakers did not significantly differ for any direct contrasts (for all $p > 0.2$; Figure 3B).

Figure 3C shows the effect of speech clarity (clear, babble) for ISC (diagonal elements). Participants' recalls were more similar for clear speech than speech in babble (effect of Speech Clarity: $F_{1,53} = 42.315$, $p = 2.8 \cdot 10^{-8}$, $\omega^2 = 0.084$) and more similar within their own speaker group than relative to the other speaker group (effect of Reference Group: $F_{1,53} = 31.769$, $p = 6.8 \cdot 10^{-7}$, $\omega^2 = 0.030$). There was no difference between English and non-English speakers (effect of Speaker: $F_{1,53} = 0.285$, $p = 0.596$, $\omega^2 < 0.001$), but the Speech Clarity \times Reference Group \times Speaker interaction was significant ($F_{1,53} = 4.421$, $p = 0.044$, $\omega^2 < 0.001$). Follow-up rmANOVAs revealed a Reference Group \times Speaker interaction for clear speech ($F_{1,53} = 7.308$, $p = 0.009$, $\omega^2 = 0.004$), such that the greater ISC for the "own" group compared to the "other" group was only significant for English speakers ($t_{53} = 6.571$, $p_{\text{Holm}} = 1.3 \cdot 10^{-7}$, $d = 0.453$), but not for non-English speakers ($t_{53} = 2.499$, $p_{\text{Holm}} = 0.078$, $d = 0.182$). For speech in babble, the Reference Group \times Speaker was not significant ($F_{1,53} = 0.269$, $p = 0.606$, $\omega^2 < 0.001$), showing greater "own" group than "other" group recall similarity for both English and non-English speakers (for both $p_{\text{Holm}} < 0.05$; main effect of Reference Group: $F_{1,53} = 21.406$, $p = 2.4 \cdot 10^{-5}$, $\omega^2 = 0.028$). None of the other interactions were significant (for all $p > 0.1$).

Figure 3D shows the effect of speech clarity (clear, babble) for temporal-order divergence. The rmANOVA revealed a greater divergence for speech in babble than clear speech (effect of Speech Clarity: $F_{1,53} = 20.264$, $p = 3.7 \cdot 10^{-5}$, $\omega^2 = 0.104$). There was no difference in temporal-order divergence between English and non-English speakers (effect of Speaker: $F_{1,53} = 0.006$, $p = 0.937$, $\omega^2 < 0.001$), but Reference Group and Speaker interacted ($F_{1,53} = 13.424$, $p = 5.7 \cdot 10^{-4}$, $\omega^2 = 0.026$). For English speakers, temporal-order divergence relative to speakers of the "other" group (non-English speakers) was lower than when calculated relative to speakers of their "own" group ($t_{53} = 2.166$, $p_{\text{Holm}} = 0.174$, $d = 0.224$), whereas in non-English speakers, temporal-order divergence relative to speakers of their "own" group was lower than when calculated relative to speakers of the "other" group (English speakers) ($t_{53} = 2.995$, $p_{\text{Holm}} = 0.025$, $d = 0.328$). In other words, the recall of both English and non-English speakers was temporally more similar (i.e., lower temporal-order divergence) relative to non-English speakers compared to English speakers. None of the other effects and interactions were significant (for all $p > 0.4$).

Recall Accuracy: GPT Ratings Through Prompt Engineering

Figure 4A shows the story \times recall rating matrices for English and non-English speakers for the prompt-engineering

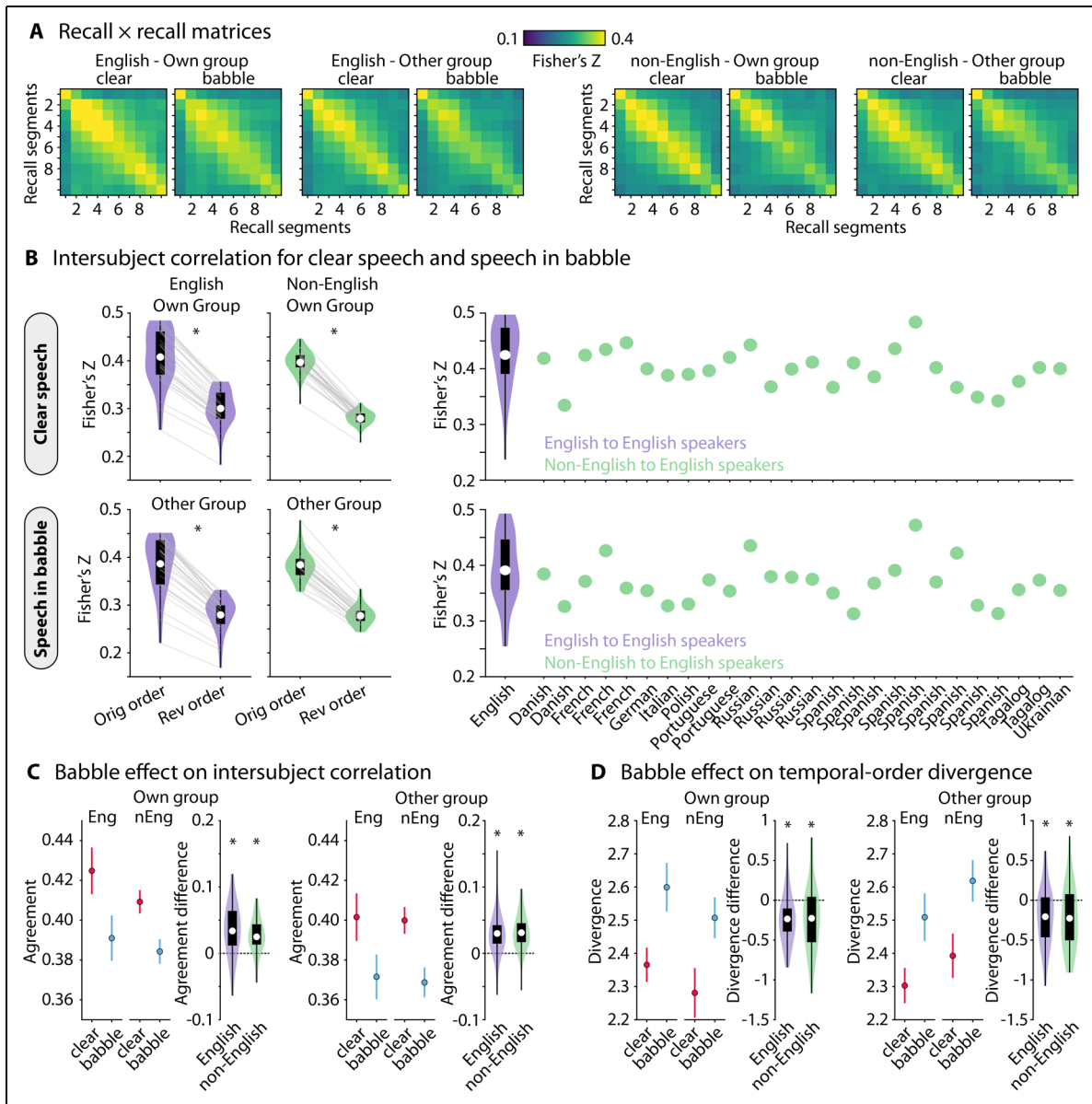


Figure 3. Recall \times Recall Intersubject Correlation Matrices and Scores (Embedding Approach). (A) Intersubject Correlation Matrices (Average Across Participants) are Shown for Clear Stories and Stories in Babble, For English and Non-English Speakers, and for Correlations Among Participants of their Speaker Own Group and Relative to the other Speaker Group. (B) Original Order (Diagonal of Recall \times Recall Matrix) and Reversed Order Scores. Violin Plots, Boxplots, and the Mean Across Participants (White Circle) are Shown. Plots on the Right Show the Correlation of Recall (Original-Order Scores, i.e., Diagonal Elements) Among English Speakers and Between Non-English and English Speakers. (C) Intersubject Correlation (Diagonal of Recall \times Recall Matrix) and Difference Between Clear Speech and Speech In Babble. (D) Same as in Panel C for Temporal-Order Divergence. Eng - English Speakers; nEng - Non-English Speakers, * $p < 0.05$.

analyses. The rmANOVAs showed that the maximum-recall score was greater than the chance recall score (i.e., maximum-recall score for unrelated stories) and that the original-order score was greater than the reverse-order score (effect of Score Type; for both analyses: $F_{1,53} > 400$, $p < 0.001$; Figure 4B), showing that stories are recalled more in the original order than in the reversed order. There was no effect of Speaker (for both $p > 0.15$) and no Score Type \times Speaker interaction (for both $p > 0.6$). Individual data points for

non-English speakers relative to the distribution of data for English speakers are shown in Figure 4B, right.

Figure 4C shows the effects of speech clarity (clear, babble) for different recall metrics and participant groups. The rmANOVA for the original-order recall revealed larger recall scores for stories under clear than under babble conditions (effect of Speech Clarity; $F_{1,53} = 11.647$, $p = 0.001$, $\omega^2 = 0.019$; Figure 4C, left). There was no difference between speakers and no interaction (for both $p > 0.6$). Recall-distinctiveness

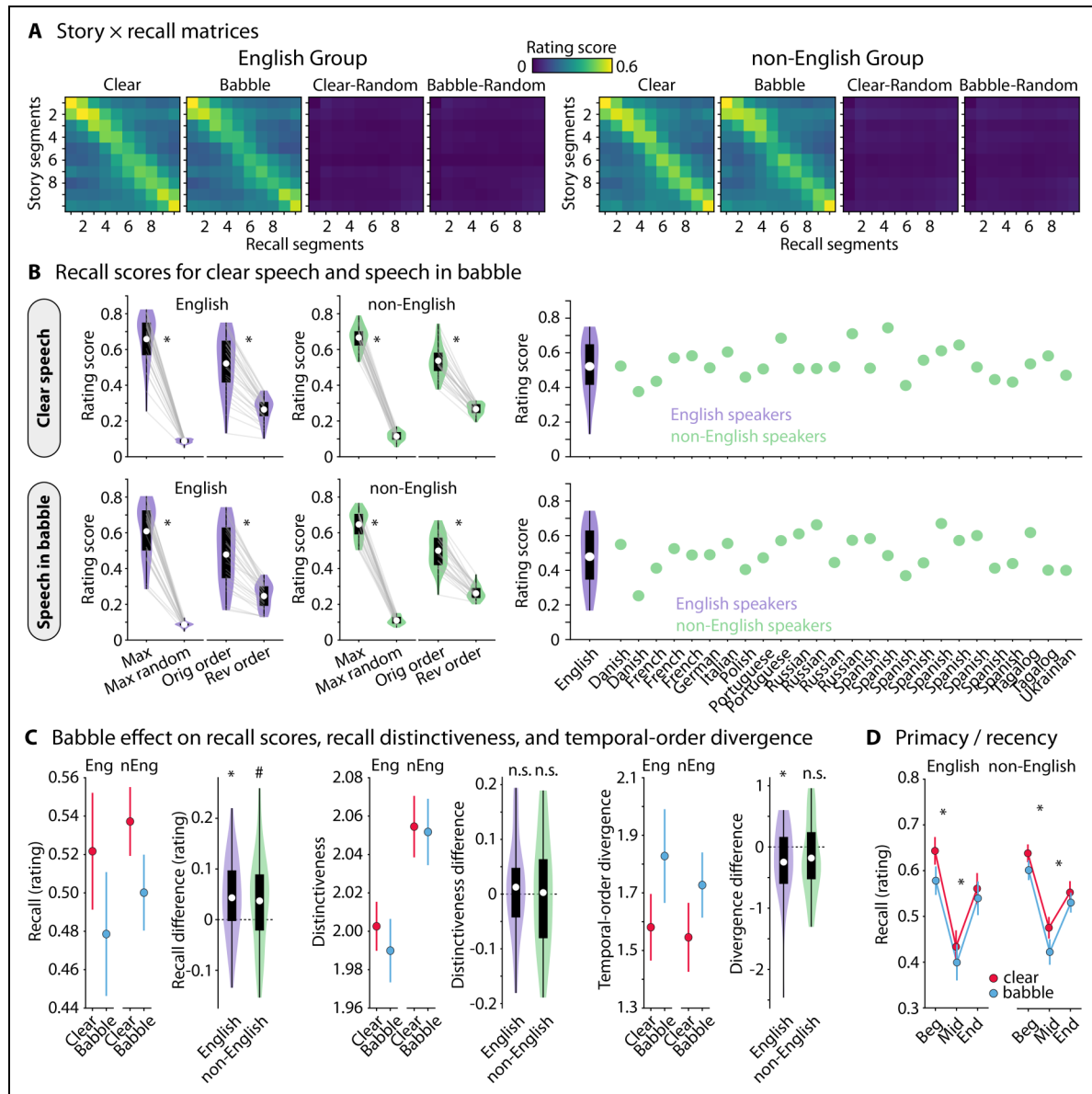


Figure 4. Story × Recall Correlation Matrices and Recall Scores (GPT Rating Approach). (A) Matrices (Average Across Participants) are Shown for Clear Stories and Stories in Babble and for the Actual Recall Scores and Chance Level (Recall Paired with an Unrelated Story). (B) Recall Scores for Clear Speech and Speech in Babble. Violin Plots, Boxplots, and the Mean Across Participants (White Circle) are Shown for the Maximum-Recall Score, Chance Maximum-Recall Score (Recall Paired with Unrelated Stories), Original-Order Recall Score, and Reverse-Order Recall Score. The Right Plot Shows Original-Order Recall Scores for Non-English Speakers and the Distribution for English Speakers, To Display the Overlap Between Speaker Groups. (C) Clear Speech Versus Speech in Babble Contrasts for Original-Order Recall Score (Left), Recall Distinctiveness (Middle), and Temporal-Order Divergence (Right). Violin Plots Show the Difference Between Clear Speech and Speech in Babble. (D) Recall Scores for the Beginning, Middle, and End of a Story/Recall, Assessing Primacy and Recency Effects. Eng – English Speakers; nEng – Non-English Speakers; N.s. – Not Significant, * $p < 0.05$, # $p < 0.1$.

was greater for non-English than English speakers ($F_{1,53} = 10.021$, $p = 0.003$, $\omega^2 = 0.077$; Figure 4C, middle), but there was no effect of Speech Clarity and no interaction (for both > 0.5). Temporal-order divergence was greater for speech in babble than clear speech (effect of Speech Clarity; $F_{1,53} = 6.377$, $p = 0.015$, $\omega^2 = 0.02$; Figure 4C, left), whereas there was no effect of Speaker and no interaction (for all $p > 0.6$; Figures 4C, right).

The rmANOVA to investigate primacy and recency effects revealed again higher recall scores for clear stories than stories in babble (effect of Speech Clarity; $F_{1,53} = 8.398$, $p = 0.005$, $\omega^2 = 0.016$; Figure 4D). Recall scores were also greater for the beginning than the middle of a story (primacy: $t_{53} = 15.249$, $p_{\text{Holm}} = 1.5 \cdot 10^{-20}$, $d = 1.150$) and the end than the middle of a story (recency: $t_{53} = 10.256$, $p_{\text{Holm}} = 6.9 \cdot 10^{-14}$, $d = 0.713$; effect of Time: $F_{2,106} = 140.834$, $p = 1.4 \cdot 10^{-30}$,

$\omega^2 = 0.220$; Figure 4D). No other main effects or interactions were significant (for all $p > 0.15$).

Evaluation of the Number of Segments

The analyses reported above were based on the segmentation of story texts and recall transcriptions into 10 segments. To evaluate whether the number of segments impacts the sensitivity of the automated recall scoring approach, the speech-clarity effects (clear vs. babble) reported in the previous sections were calculated for 6, 10, 14, and 18 story-text and recall-transcription segments.

For recall scoring using the embedding approach (Figure 5A), scores were greater for clear speech than speech in babble for all segment conditions (for all $F_{1,53} > 9$, $p < 0.005$). This speech-clarity effect did not differ between English and non-English speakers (for all $F_{1,53} < 0.7$, $p > 0.4$), and there was no overall difference between speaker groups (for all $F_{1,53} < 0.5$, $p > 0.5$). This indicates that recall scoring is sensitive to speech-clarity conditions in different languages regardless of whether the story texts and recall transcriptions are divided into 6, 10, 14, or 18 segments. However, the speech-clarity effect (clear minus babble) declined with the increase in the number of segments (linear trend; $t_{53} = -2.779$, $p = 0.008$, $d = 0.167$; Figure 5B).

Figure 5C displays the recall \times recall matrices (ISC) for transcriptions divided into 6, 10, 14, and 18 segments. For all segment numbers, ISC (i.e., the diagonal of recall \times recall matrices) was greater for clear speech than speech in babble (for all $F_{1,53} > 35$, $p < 0.001$) and greater among individuals of their own group compared to individuals from the other group (for all $F_{1,53} > 15$, $p < 0.001$; although for 14 and 18 segments, this was greater for English than non-English speakers: Reference Group \times Speaker interaction $p < 0.05$; whereas for 10 and 14 segments this difference was specifically greater for English speakers for clear speech: Speech Clarity \times Reference Group \times Speaker interaction $p < 0.05$). There was no overall difference in ISC between English and non-English speakers for any of the segment-number conditions (for all $F_{1,53} < 0.9$, $p > 0.35$). The speech-clarity effect (clear minus babble) decreased with the increase in the number of segments (linear trend: $t_{53} = -6.381$, $p = 4.5 \cdot 10^{-8}$, $d = 0.434$; Figure 5D). Temporal-order divergence was greater for speech in babble than clear speech for all segment conditions (for all $F_{1,53} > 20$, $p < 0.001$; Figure 5E). Similar to the results reported above (Figure 3D), temporal-order divergence was lower for calculations based on similarity with non-English than English speakers for all segment conditions (Reference Group \times Speaker interaction: for all $F_{1,53} > 10$, $p < 0.005$). There was no overall difference between speakers for any of the segment conditions (for all $F_{1,53} < 0.5$, $p > 0.5$). The speech-clarity effect (clear minus noise) increased with the increase in the number of segments (linear trend: $t_{53} = -4.787$, $p = 1.4 \cdot 10^{-5}$, $d = 0.358$; Figure 5E).

For recall scoring using the GPT ratings (Figures 5F and G), scores were greater for clear speech than speech in babble for all segment conditions (for all $F_{1,53} > 10$, $p < 0.003$). The speech-clarity effect did not differ between English and non-English speakers (for all $F_{1,53} < 0.2$, $p > 0.7$), and there was no overall difference between speaker groups (for all $F_{1,53} < 0.5$, $p > 0.4$). This indicates that recall scoring is sensitive to speech-clarity conditions in different languages regardless of whether the story texts and recall transcriptions are divided into 6, 10, 14, or 18 segments. There was no difference in the speech-clarity effect (clear minus babble) for different segment numbers ($F_{1,53} = 1.146$, $p = 0.332$, $\omega^2 < 0.001$; Figure 5G).

Overall, the analyses reported in this section show sensitivity to the effects of speech clarity for story texts and recall transcriptions divided into 6, 10, 14, and 18 segments, and no meaningful differences between English and non-English speakers. Nevertheless, the effect of speech clarity for recall scores and ISC (using the embedding approach) was greatest when fewer segments were used, whereas the speech-clarity effect for temporal-order divergence was greatest when more segments were used, suggesting an intermediate number of segments may be beneficial to maximize sensitivity for different metrics. GPT ratings were similarly sensitive to speech clarity for all segment numbers.

Discussion

The current study examined whether speech comprehension scoring of naturalistic stories can be automated for listeners with different language backgrounds. The data show that recall scoring using both LLM text-embeddings and LLM prompt engineering is sensitive to relevant speech-comprehension effects, including original temporal-order recall, benefits from primacy and recency information in stories, and reduced recall for speech in background noise. Scoring data for native speakers of various languages showed no meaningful differences, suggesting that LLMs provide a powerful way toward language-independent speech scoring. The current, fully automated approach—from story generation to recall scoring—across languages shows the feasibility of assessing speech comprehension with naturalistic speech materials in speakers with different language backgrounds. The work has the potential to open new avenues for an accessible clinical practice that captures challenges in naturalistic speech listening.

Large Language Models for Automated Comprehension Scoring

Manual scoring of verbal or written reports from participants or patients can be extremely time-consuming (Borrie et al., 2019; Bosker, 2021; Herrmann, 2025a; Levine et al., 2002; Martinez, 2024; van Genugten and Schacter, 2024). Recent work shows the applicability of LLMs to convert verbal

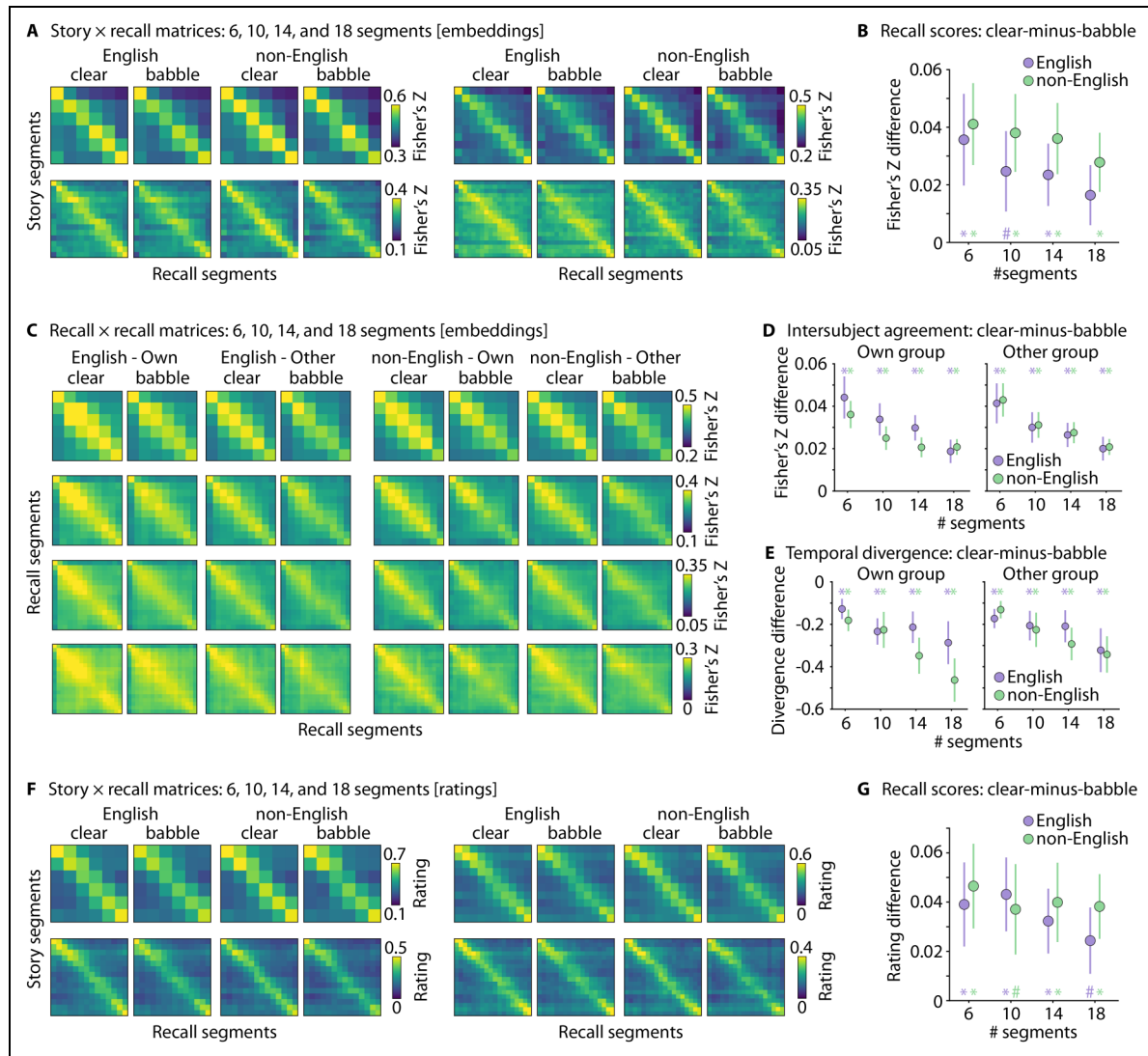


Figure 5. Results for Different Segment Numbers. (A) Story × Recall Matrices (from the Embedding Approach) Based on Story Texts and Recall Transcriptions Divided into 6, 10, 14, or 18 Segments (Top Left, Top Right, Bottom Left, and Bottom Right, Respectively). (B) The Effect of Background Babble (Clear Minus Babble) on Recall Scores (Original-Order Recall; Diagonal Elements). The Speech-Clarity Effect Decreased for Higher Segment Numbers. (C) Recall × Recall Matrices (from the Embedding Approach) Based on Recall Transcriptions Divided into 6, 10, 14, or 18 Segments. (D) The Effect of Background Babble (Clear Minus Babble) on Intersubject Correlation (Diagonal Elements). The Speech-Clarity Effect Decreased for Higher Segment Numbers. (E) The Effect of Background Babble (Clear Minus Babble) on Temporal-Order Divergence Scores. The Speech-Clarity Effect Increased for Higher Segment Numbers. (F and G) Same as in Panels A and B for Story × Recall Matrices from the GPT Rating Approach Through Prompt Engineering. All Matrices Reflect the Average Across Participants. Error Bars Reflect the Standard Error of the Mean. The Colored * and # Indicate $P < 0.05$ and $P < 0.1$ Significance Levels, Respectively, For Tests Against Zero.

speech to text for speech-recognition use in the speech, language, and hearing sciences (Ballier et al., 2024; Slaney and Fitzgerald, 2024; Zhao et al., 2025). Verbatim word report scoring can also be automatized using LLM text-embeddings (Herrmann, 2025a), but this work has still been limited to words or short sentences.

The scoring of free recall data is substantially more challenging (Levine et al., 2002; Martinez, 2024; van Genugten and Schacter, 2024), especially across languages, because

people's mental representations for naturalistic stimuli are not verbatim (Fillenbaum, 1966; Gomulicki, 1956; Mehler, 1963; Mehler and Miller, 1964; Raccach et al., 2024; Sachs, 1967). Each person uses somewhat different words and descriptions when recalling the same story information (Raccach et al., 2024). Scoring in different languages would require a human scorer to be fluent in numerous languages, which makes scalability to a large number of languages not feasible. The current data show that scoring speech recall

for speakers with different language backgrounds can be automated using semantic similarity analyses (text-embeddings) and rating scores derived through prompt engineering.

A few other recent works have shown the feasibility of using text-embeddings on segmented recall data for scoring movie comprehension (Heusser et al., 2021; Shen et al., 2023), and recently for story reading (Martinez, 2024; Panella et al., 2025) and listening comprehension (Chandler et al., 2021; Martinez, 2024; Racciah et al., 2024). Some of these works show that scores from text-embeddings correlate well with scores made by humans (Chandler et al., 2021; Martinez, 2024; Panella et al., 2025), including the text-embedding approach employed here (Panella et al., 2025). The current work extends these previous works by fully automating story generation and scoring in different languages and testing new assessment metrics—that is, temporal-order divergence and ISC—that show sensitivity to the clarity of speech. Other recent works used LLM prompting to generate short passages with a predefined number of details and LLM prompting to quantify the number of details correctly reported by participants (Georgiou et al., 2025; Martinez, 2024). This work focused on sentence clauses, which are unlikely the unit of mental representations of information encountered in everyday conversations. Nevertheless, scores from prompt engineering showed a high correlation with human raters (Georgiou et al., 2025; Martinez, 2024). Scoring approaches using text-embeddings—despite showing meaningful comprehension sensitivity—can have the disadvantage that even unrelated multisentence passages can lead to nonzero correlation/similarity scores (Figures 2A and 2B), thereby reducing the dynamic range and interpretability. By combining the general pipeline of story/recall text segmentation and obtaining metrics from story \times recall matrices with LLM prompt engineering, the current study demonstrates a wider dynamic range and chance-level scores close to zero (Figure 4B), while also showing sensitivity to relevant speech-comprehension effects across speakers with different language backgrounds. Combining text segmentation with LLM-based speech-similarity ratings may thus provide the most fruitful avenue for research and clinical practice.

Sensitivity of LLMs to Different Aspects of Speech Comprehension

The current approach is sensitive to speech comprehension as indicated in several ways. Scores were greater than chance (i.e., when recall data were paired with an unrelated story), and scores were higher for recall in the original than reversed temporal order (Figures 2B and 4B), indicating that participants reported story segments in the order in which they heard a story, which is consistent with previous work (Chen et al., 2017; Heusser et al., 2021; Howard and Kahana, 2002; Zhang et al., 2023). Moreover, recall scores were greater for story segments in the beginning and end compared to the

middle of a story, showing the well-known primacy and recency recall pattern (Murdock Jr, 1962; Howard and Kahana, 1999; Tan and Ward, 2000; Howard and Kahana, 2002; Zhang et al., 2023). There were no meaningful differences between English speakers and non-English speakers. In fact, the ISC analyses demonstrate a similar distribution of correlation scores among the group of English speakers and correlation scores of non-English with English speakers. The automated scoring is thus language-independent.

The current scoring approach also shows sensitivity to the clarity of speech. Recall scores were lower for speech in the presence of background noise compared to clear speech. A reduction in speech intelligibility due to background noise is well-known (Gordon et al., 2009; Helfer and Jesse, 2021; Mattys et al., 2012; McArdle and Wilson Richard, 2009; Miller, 1947; Pandey and Herrmann, in press; Pichora-Fuller et al., 2016; Summerfield, 1987) and translates to reduced story comprehension in the current study. Moreover, the temporal order in which participants recalled the stories was reduced for speech in background noise, which appeared to be most prominent in the ISC analyses (Figure 3D). Temporal-order effects cannot be revealed well with traditional speech-perception assessments, because these assessments rely on individual words or short sentences (Billings et al., 2024; Gustafsson and Arlinger, 1994; Killion et al., 2004; Polspoel et al., 2025; Wilson et al., 2012; Wilson Richard et al., 2007). Automatically scoring the comprehension of naturalistic, story speech thus offers additional avenues to understand speech-comprehension challenges in everyday life.

Young normal-hearing listeners correctly hear about 80% of words of continuous speech masked by the level of background babble used in the current study (+2 dB SNR; Irsik et al., 2022). Participants tend to be as engaged in listening to stories under this level of background babble as they are when listening to clear stories (Herrmann and Johnsrude, 2020b; Irsik et al., 2022), suggesting that missing or incorrectly hearing about 20% of words does not impede comprehension success. This could explain why speech clarity did not affect all recall metrics, and why the significance of the speech-clarity effects varied across analysis types (e.g., recall score vs. ISC).

Limitations and Practical Considerations

The current study provides a language-independent approach to automating the scoring of free-recall data of individuals listening to short stories. The approach is sensitive to critical speech-comprehension effects, but there are practical considerations and limitations.

Stories were generated, translated, and synthesized in different languages using modern large language models (LLM), but the rating scores by non-English native speakers highlighted two limitations. Translations occasionally included words that would typically not be used in the context in which

they were presented. This limitation is consistent with recent advocacy for including cultural competence beyond text-based information into the training of models for machine translations to improve translations (Feuerriegel et al., 2025; Tenzer et al., 2024). Moreover, the OpenAI voice used in the current study (i.e., Alloy) led to speech that was spoken in a somewhat American-English accent, because the model was optimized for English. The voice was selected because it is naturalistic-sounding and controls voice acoustics across languages. Other synthesizers for different languages are available, such as Google's text-to-speech (<https://cloud.google.com/text-to-speech/docs/voices>). However, American-English was the only highly naturalistic Google voice available at the time the research was conducted, whereas voices in other languages still relied on older models that resulted in less naturalistic speech. The OpenAI voice did not work well with Asian or Arabic languages (despite good translations), leading to incomprehensible auditory materials. This will likely not be an issue for long, given the fast-developing AI space. In fact, since this research was conducted, new models that appear less English-accented and more naturalistic have been made available by both OpenAI and Google. Despite the language and acoustic limitations in the current study, participants with different language backgrounds found the stories engaging and there were no meaningful impacts on the automated scoring.

The current study compared speech listening and comprehension scoring between people with different language backgrounds. The results suggest that full speech-assessment automation can be achieved across languages using modern speech synthesizers and LLM. However, speech perception can also be affected by regional dialects/accents (Flocchia et al., 2006; Goslin et al., 2012; McCullough et al., 2017; Nissen Shawn et al., 2013; Wang et al., in press), but modern speech synthesizers that generate regional dialects/accents have not emerged (with the exception of Google's English US, UK, Australian, and Indian accents). Future technological advances may fill this space.

The automated recall scoring involved dividing story and recall texts into segments. This choice was based on previous recall-scoring work (Chen et al., 2017; Heusser et al., 2021; Racciah et al., 2024) and cognitive research showing that individuals perceive, encode, and recall everyday environments as discrete events (Lee and Chen, 2022; Pitts et al., 2023; Richmond et al., 2017; Sasmita and Swallow, 2022; Speer et al., 2004; Zacks et al., 2007; Zacks and Swallow, 2007). Segmentation was also used to avoid semantic dilution (Zhou et al., 2024). The current work explored dividing the story and recall texts into different number of segments, showing—for the text-embedding approach—that maximizing the sensitivity to speech-clarity effects may benefit from an intermediate number of segments. If overall recall accuracy is the main focus, fewer segments may be beneficial, whereas sensitivity to effects on temporal-order divergence may be greater for a higher number of segments. The latter may be

the case because more segments provide a higher temporal resolution. For LLM prompting, the number of segments appeared to matter less. Eventually, the choice for the number of segments may also depend on the duration of the story, with longer stories possibly requiring more segments.

In the current study, ~2 min spoken stories were used to examine whether speech comprehension scoring can be automated using LLMs. In theory, longer stories could pose more challenges for scoring participants' recall because more information is integrated, which could lead to semantic dilution (Zhou et al., 2024). Nevertheless, our recent work in English speakers shows that the current assessment approach works well also for 10-min stories (Panella et al., 2025). The segmentation into story snippets appears to be critical, especially for longer speech materials and recalls to avoid semantic dilution. For longer stories, participants may also increasingly recall only story gist rather than individual details (Georgiou et al., 2025), potentially counteracting methodological issues related to semantic dilution, but this requires further research. More generally, for a specific clinical test, story generation will require validation to ensure that the generated stories afford similar recall scores under ideal conditions.

The languages included in the current study all favor chronological narrative structures. Although for most (if not all) languages, narratives are temporally forward-moving when spoken, there may be nuance in how narratives are structured that could interact with the assessment across languages. This could affect ISC analysis, but should not affect maximum recall scores because this measure is independent of the temporal order in which a story is recalled. Moreover, there is a chance that linguistic complexity is altered during the translation of stories, which in turn may be reflected in participants' recall. If a researcher was concerned about these potential issues, they could translate the recall to English prior to analysis (assuming translations would somewhat account for structural differences). In the current study, no meaningful differences were observed when using the story and recall transcriptions in the original language compared to English translations (Figures S1 and S2 in the Supplementary Materials).

The current automated scoring approach leveraged the LaBSE language-independent text-embedding model (LaBSE; Devlin et al., 2019; Feng et al., 2022). Other multi-language models, such as USE-CMLM-Multilingual (Yang et al., 2020), were also explored, but initial evaluations suggested poorer performance. Examinations further suggest that the translation of story and recall texts to English using OpenAI's GPT-4o and then using English-based models or LaBSE lead to qualitatively similar results as the ones reported here (Figures S1 and S2 in the Supplementary Materials). Moreover, scoring through LLM prompt engineering did not require translation. It was sufficient to use English instructions with text pieces in other languages. Whether scoring performance could be improved somewhat by using

non-English prompts for non-English recall scoring may be worth exploring in the future, but it appears that state-of-the-art models simply map speech across different languages.

Using LLM can come with computational and direct financial costs. OpenAI's LLMs were used in the current study to generate stories, translate speech, and synthesize speech. The costs associated with these tasks involved less than two dollars and were performed within a few minutes for all participants (using a Dell XPS14 laptop). The LaBSE text-embedding model is freely available, and the calculation of the correlation matrices took less than a minute per participant (calculations of ISC took a few minutes per participant). The LLM zero-shot prompt engineering approach was computationally and financially more expensive even though GPT-4o mini was used. Fees across the calculations conducted were about \$80 USD (including chance level) and calculations took about 15 min per participant. Nevertheless, GPT prompt engineering showed somewhat better performance than the embedding approach, such that scores had a greater dynamic range and chance-level scores were near zero. To minimize costs in future work, only the diagonal of the story \times recall matrix could be calculated, reducing the number of computations 10-fold (for a 10×10 matrix), and calculations for chance level matrices could be omitted. Fees could be avoided by using open-access models like LLaMA 3.1, but this may come with reduced accuracy (LLaMA performs more poorly than GPT; e.g., Panella et al., 2025) and may require higher computational capacity in-house (e.g., computers with GPUs). Other new models such as DeepSeek v3 (<https://www.deepseek.com/>) may offer a 10 times lower price at the same performance level. Considerations of data sharing, ethical use, and carbon footprint may also determine the model choice (Luccioni et al., 2023; Rillig et al., 2023). The AI landscape is evolving fast, but the current approach outlines the general architecture of automated recall scoring across languages that could likely be used with any future model.

Conclusions

The current study investigated the feasibility of a fully automated approach to story generation, speech synthesis, and speech-recall scoring across different languages. A group of English speakers and a group of native speakers of a non-English language (10 non-English languages) listened and recalled short stories, both in their native language. The results showed that scoring recall with LLM text-embeddings and prompt engineering is sensitive to known speech-comprehension effects, including temporal-order recall, primacy and recency benefits, and sensitivity to reductions in speech clarity. The LLM prompt engineering approach may be preferable over the LLM text-embedding approach because the scores for the former have a wider dynamic range and are more interpretable. Critically, there were no

meaningful differences between English and non-English speakers, demonstrating automated, language-independent scoring of speech recall. The current approach may open fruitful avenues for clinical assessments with naturalistic speech.

Acknowledgments

The author thanks Lisa D'Souza for her help with data collection and Ryan Panella for his comments on an earlier version of the manuscript. The research was supported by the Canada Research Chair program (CRC-2023-00383), the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant: RGPIN-2021-02602), and the Data Sciences Institute at the University of Toronto.

Author Contributions

Björn Herrmann: Conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualization, supervision, project administration, and funding acquisition.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Natural Sciences and Engineering Research Council of Canada, University of Toronto, Canada Research Chairs, (RGPIN-2021-02602, Data Sciences Institute, CRC-2023-00383, respectively).

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability

Data are available at <https://osf.io/qdg89/>. Three participants declined to share their data publicly (we employ separate consent forms for study participation and data sharing in line with Canadian Tri-Council Policies for Ethical Conduct for Research Involving Humans—TCPS 2 from 2022). Their data are thus not made available.

Supplemental Material

Supplemental material for this article is available online.

References

- Auerbach, B. D., Rodrigues, P. V., & Salvi, R. J. (2014). Central gain control in tinnitus and hyperacusis. *Frontiers in Neurology*, 5, Article 206. <https://doi.org/10.3389/fneur.2014.00206>
- Bainbridge, K. E., & Wallhagen, M. I. (2014). Hearing loss in an aging American population: Extent, impact, and management. *Annual Review of Public Health*, 35(1), 139–152. <https://doi.org/10.1146/annurev-publhealth-032013-182510>
- Ballier, N., Arnold, T., Méli, A., Thurston, T., & Yunès, J.-B. (2024). Whisper for L2 speech scoring. *International Journal of Speech Technology*, 27(4), 923–934. <https://doi.org/10.1007/s10772-024-10141-5>

- Bilger, R. C. (1984). *Manual for the clinical use of the revised SPIN Test*. The University of Illinois.
- Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rieczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech, Language, and Hearing Research*, 27(1), 32–48. <https://doi.org/10.1044/jshr.2701.32>
- Billings, C. J., Olsen, T. M., Charney, L., Madsen, B. M., & Holmes, C. E. (2024). Speech-in-noise testing: An introduction for audiologists. *Seminars in Hearing*, 45(1), 55–82. <https://doi.org/10.1055/s-0043-1770155>
- Bohanek, J. G., Fivush, R., Zaman, W., Lepore, C. E., Merchant, S., & Duke, M. P. (2009). Narrative interaction in family dinnertime conversations. *Merrill-Palmer Quarterly*, 55(4), 488–515. <https://doi.org/10.1353/mpq.0.0031>
- Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America*, 145(1), 392–399. <https://doi.org/10.1121/1.5087276>
- Bosker, H. R. (2021). Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies. *Behavior Research Methods*, 53(5), 1945–1953. <https://doi.org/10.3758/s13428-021-01542-4>
- Bott, A., & Saunders, G. (2021). A scoping review of studies investigating hearing loss, social isolation and/or loneliness in adults. *International Journal of Audiology*, 60(Suppl 2), 30–46. <https://doi.org/10.1080/14992027.2021.1915506>
- Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strophe, B., & Kurzweil, R. (2018). Universal sentence encoder. *arXiv*. <https://doi.org/10.48550/arXiv.1803.11175>
- Chandler, C., Holmlund, T. B., Foltz, P. W., Cohen, A. S., & Elvevåg, B. (2021). Extending the usefulness of the verbal memory test: The promise of machine learning. *Psychiatry Research*, 297, 113743. <https://doi.org/10.1016/j.psychres.2021.113743>
- Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1), 115–125. <https://doi.org/10.1038/nn.4450>
- Chen, B., Zhang, Z., Langrene, N., & Zhu, S. (2024). Unleashing the potential of prompt engineering: a comprehensive review. *arXiv*. <https://doi.org/10.48550/arXiv.2310.14735>
- Chern, A., & Golub, J. S. (2019). Age-related hearing loss and dementia. *Alzheimer Disease and Associated Disorders*, 33(3), 285–290. <https://doi.org/10.1097/WAD.0000000000000325>
- Chien, W., & Lin, F. R. (2013). Prevalence of hearing aid use among older adults in the United States. *Archives of Internal Medicine*, 173(2), 292–293. <https://doi.org/10.1001/archinternmed.2011.1408>
- Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heaffield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., & ... Wang, J. (2024). Scaling neural machine translation to 200 languages. *Nature*, 630(8018), 841–846. <https://doi.org/10.1038/s41586-024-07335-x>
- Cruikshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, B. E. K., Klein, R., Mares-Perlman, J. A., & Nondahl, D. M. (1998). Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin. *American Journal of Epidemiology*, 148(9), 879–886. <https://doi.org/10.1093/oxfordjournals.aje.a009713>
- Davidson, A., Marrone, N., Wong, B., & Musiek, F. (2021). Predicting hearing aid satisfaction in adults: A systematic review of speech-in-noise tests and other behavioral measures. *Ear & Hearing*, 42(6), 1485–1498. <https://doi.org/10.1097/AUD.0000000000001051>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. <https://doi.org/10.48550/arXiv.1810.04805>
- 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? *Frontiers in Human Neuroscience*, 6(Article 112). <https://doi.org/10.3389/fnhum.2012.00112>
- Eisenberg, A. R. (1985). Learning to describe past experiences in conversation. *Discourse Processes*, 8(2), 177–204. <https://doi.org/10.1080/01638538509544613>
- Feder, K., Michaud, D., Ramage-Morin, P., McNamee, J., & Beauregard, Y. (2015). Prevalence of hearing loss among Canadians aged 20 to 79: Audiometric results from the 2012/2013 Canadian health measures survey. *Health Reports*, 26(7), 18–25.
- Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. *arXiv*. <https://doi.org/10.48550/arXiv.2007.01852>
- Ferguson, S. H., Jongman, A., Sereno, J. A., & Keum, K. A. (2010). Intelligibility of foreign-accented speech for older adults with and without hearing loss. *Journal of the American Academy of Audiology*, 21(3), 153–162. <https://doi.org/10.3766/jaaa.21.3.3>
- Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B., & Van Bavel, J. J. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 4(2). <https://doi.org/10.1038/s44159-024-00392-z>
- Fillenbaum, S. (1966). Memory for gist: Some relevant variables. *Language and Speech*, 9(4), 217–227. <https://doi.org/10.1177/002383096600900403>
- Fitzgerald, M. B., Gianakas, S. P., Qian, Z. J., Losorelli, S., & Swanson, A. C. (2023). Preliminary guidelines for replacing word-recognition in quiet with speech in noise assessment in the routine audiologic test battery. *Ear and Hearing*, 44(6). <https://doi.org/10.1097/AUD.0000000000001409>
- Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance*, 32(5), 1276–1293. <https://doi.org/10.1037/0096-1523.32.5.1276>
- Garcia Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, 52(11-12), 864–886. <https://doi.org/10.1016/j.specom.2010.08.014>

- Georgiou, A., Can, T., Katkov, M., & Tsodyks, M. (2025). Large-scale study of human memory for meaningful narratives. *Learning & Memory*, 32(2), a054043. <https://doi.org/10.1101/lm.054043.124>
- Gianakas Steven, P., & Winn Matthew, B. (2025). Advance contextual clues alleviate listening effort during sentence repair in listeners with hearing aids. *Journal of Speech, Language, and Hearing Research*, 68(4), 2144–2156. https://doi.org/10.1044/2025_JSLHR-24-00184
- Goman, A. M., & Lin, F. R. (2016). Prevalence of hearing loss by severity in the United States. *American Journal of Public Health*, 106(10), 1820–1822. <https://doi.org/10.2105/AJPH.2016.303299>
- Gomulicki, B. R. (1956). Recall as an abstractive process. *Acta Psychologica*, 12, 77–94. [https://doi.org/10.1016/0001-6918\(56\)90009-9](https://doi.org/10.1016/0001-6918(56)90009-9)
- Gordon, M. S., Daneman, M., & Schneider, B. A. (2009). Comprehension of speeded discourse by younger and older listeners. *Experimental Aging Research*, 35(3), 277–296. <https://doi.org/10.1080/03610730902769262>
- Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language*, 122(2), 92–102. <https://doi.org/10.1016/j.bandl.2012.04.017>
- Gustafsson, HÅ, & Arlinger, S. D. (1994). Masking of speech by amplitude-modulated noise. *The Journal of the Acoustical Society of America*, 95(3), 518–529. <https://doi.org/10.1121/1.408346>
- Hasson, U., Furman, O., Clark, D., Dudai, Y., & Davachi, L. (2008). Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. *Neuron*, 57(3), 452–462. <https://doi.org/10.1016/j.neuron.2007.12.009>
- Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, 14(1), 40–48. <https://doi.org/10.1016/j.tics.2009.10.011>
- Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664), 1634–1640. <https://doi.org/10.1126/science.1089506>
- Helfer, K. S., & Jesse, A. (2021). Hearing and speech processing in midlife. *Hearing Research*, 402, 108097. <https://doi.org/10.1016/j.heares.2020.108097>
- Herrmann, B. (2023). The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults. *International Journal of Speech Technology*, 26, 395–415. <https://doi.org/10.1007/s10772-023-10027-y>
- Herrmann, B. (2025a). Leveraging natural language processing models to automate speech-intelligibility scoring. *Speech, Language and Hearing*, 28(1), 2374160. <https://doi.org/10.1080/2050571X.2024.2374160>
- Herrmann, B. (2025b). Age-related differences in the impact of background noise on neural speech tracking. *bioRxiv*. <https://doi.org/10.1101/2025.01.09.632269>
- Herrmann, B. (2025c). Enhanced neural speech tracking through noise indicates stochastic resonance in humans. *eLife*, 13, RP100830. <https://doi.org/10.7554/eLife.100830.3>
- Herrmann, B., & Butler, B. E. (2021). Hearing loss and brain plasticity: The hyperactivity phenomenon. *Brain Structure & Function*, 226(7), 2019–2039. <https://doi.org/10.1007/s00429-021-02313-9>
- Herrmann, B., & Johnsrude, I. S. (2020a). A model of listening engagement (MoLE). *Hearing Research*, 397, 108016. <https://doi.org/10.31234/osf.io/y3cr9>
- Herrmann, B., & Johnsrude, I. S. (2020b). Absorption and enjoyment during listening to acoustically masked stories. *Trends in Hearing*, 24, 1–18. <https://doi.org/10.1177/2331216520967850>
- Heusser, A. C., Fitzpatrick, P. C., & Manning, J. R. (2021). Geometric models reveal behavioural and neural signatures of transforming experiences into memories. *Nature Human Behaviour*, 5(7), 905–919. <https://doi.org/10.1038/s41562-021-01051-6>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <https://doi.org/http://www.jstor.org/stable/4615733>
- Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4), 923–941. <https://doi.org/10.1037/0278-7393.25.4.923>
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3), 269–299. <https://doi.org/10.1006/jmps.2001.1388>
- Irsik, V. C., Johnsrude, I. S., & Herrmann, B. (2022). Neural activity during story listening is synchronized across individuals despite acoustic masking. *Journal of Cognitive Neuroscience*, 34(6), 933–950. https://doi.org/10.1162/jocn_a_01842
- JASP. (2024). JASP [Computer software]. In: <https://jasp-stats.org/>.
- Jefferson, G. (1978). Sequential aspects of storytelling in conversation. In J. I. M. Schenkein (Ed.), *Studies in the organization of conversational interaction* (pp. 219–248). Academic Press.
- Karunathilake, I. M. D., Brodbeck, C., Bhattasali, S., Resnik, P., & Simon, J. Z. (2025). Neural dynamics of the processing of speech features: Evidence for a progression of features from acoustic to sentential processing. *The Journal of Neuroscience*, 45(11), e1143242025. <https://doi.org/10.1523/JNEUROSCI.1143-24.2025>
- Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 116(4 Pt 1), 2395–2405. <https://doi.org/10.1121/1.1784440>
- Kilman, L., Zekveld, A., Hällgren, M., & Rönnberg, J. (2015). Native and non-native speech perception by hearing-impaired listeners in noise- and speech maskers. *Trends in Hearing*, 19, 1–12. <https://doi.org/10.1177/2331216515579127>
- Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2), 72–79. <https://doi.org/10.1016/j.tics.2007.11.004>
- Kurby, C. A., & Zacks, J. M. (2018). Preserved neural event segmentation in healthy older adults. *Psychology and Aging*, 33(2), 232–245. <https://doi.org/10.1037/pag0000226>
- Lee, H., & Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, 13(1), 4235. <https://doi.org/10.1038/s41467-022-31965-2>

- Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: Dissociating episodic from semantic retrieval. *Psychology and Aging, 17*(4), 677–689. <https://doi.org/10.1037/0882-7974.17.4.677>
- Lin, F. R., & Albert, M. (2014). Hearing loss and dementia—who is listening? *Aging & Mental Health, 18*(6), 671–673. <https://doi.org/10.1080/13607863.2014.915924>
- Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S. G., Dias, A., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Kivimäki, M., Larson, E. B., Ogunniyi, A., & ... Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet, 396*(10248), 413–446. [https://doi.org/10.1016/S0140-6736\(20\)30367-6](https://doi.org/10.1016/S0140-6736(20)30367-6)
- Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research, 24*(1), 1–15. <https://doi.org/10.5555/3648699.3648952>
- Martinez, D. (2024). Scoring story recall for individual differences research: Central details, peripheral details, and automated scoring. *Behavior Research Methods, 56*(8), 8362–8378. <https://doi.org/10.3758/s13428-024-02480-7>
- Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7-8), 953–978. <https://doi.org/10.1080/01690965.2012.705006>
- McArdle, R., & Wilson Richard, H. (2009). Speech perception in noise: The basics. *Perspectives on Hearing and Hearing Disorders: Research and Diagnostics, 13*(1), 4–13. <https://doi.org/10.1044/hhd13.1.4>
- McCullough, E. A., Clopper, C. G., & Wagner, L. (2017). Regional dialect perception across the lifespan: Identification and discrimination. *Language and Speech, 62*(1), 115–136. <https://doi.org/10.1177/0023830917743277>
- McNeice, Z., Tomlin, D., Timmer, B., Short, C. E., Nixon, G., & Galvin, K. (2024). A scoping review exploring how adults self-describe and communicate about the listening difficulties they experience. *International Journal of Audiology, 63*(3), 163–170. <https://doi.org/10.1080/14992027.2023.2245136>
- Mehler, J. (1963). Some effects of grammatical transformation on the recall of English sentences. *Journal of Verbal Learning & Verbal Behavior, 2*(4), 346–351. [https://doi.org/10.1016/S0022-5371\(63\)80103-6](https://doi.org/10.1016/S0022-5371(63)80103-6)
- Mehler, J., & Miller, G. A. (1964). Retroactive interference in the recall of simple sentences. *British Journal of Psychology, 55*(3), 295–301. <https://doi.org/10.1111/j.2044-8295.1964.tb00913.x>
- Michelmann, S., Kumar, M., Norman, K. A., & Toneva, M. (2025). Large language models can segment narrative events similarly to humans. *Behavior Research Methods, 57*(1), 39. <https://doi.org/10.3758/s13428-024-02569-z>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. <https://doi.org/10.48550/arXiv.1301.3781>
- Miller, G. A. (1947). The masking of speech. *Psychological Bulletin, 44*(2), 105–129. <https://doi.org/10.1037/h0055960>
- Motala, A., Johnsrude, I. S., & Herrmann, B. (2024). A longitudinal framework to describe the relation between age-related hearing loss and social isolation. *Trends in Hearing, 28*(2), 1–18. <https://doi.org/10.1177/23312165241236041>
- Mullen, M. K., & Yi, S. (1995). The cultural context of talk about the past: Implications for the development of autobiographical memory. *Cognitive Development, 10*(3), 407–419. [https://doi.org/10.1016/0885-2014\(95\)90004-7](https://doi.org/10.1016/0885-2014(95)90004-7)
- Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*, 482–488. <https://doi.org/10.1037/h0045106>
- Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience, 14*(6), 669–687. <https://doi.org/10.1093/scan/nsz037>
- Nguyen, M., Vanderwal, T., & Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *NeuroImage, 184*, 161–170. <https://doi.org/10.1016/j.neuroimage.2018.09.010>
- Nissen Shawn, L., Harris Richard, W., Channell Ron, W., Richardson Nathan, E., Garlick Jamie, A., & Eggett Dennis, L. (2013). The effect of dialect on speech audiometry testing. *American Journal of Audiology, 22*(2), 233–240. [https://doi.org/10.1044/1059-0889\(2013\)12-0077](https://doi.org/10.1044/1059-0889(2013)12-0077)
- Ochs, E., Taylor, C., Rudolph, D., & Smith, R. (1992). Storytelling as a theory-building activity. *Discourse Processes, 15*(1), 37–72. <https://doi.org/10.1080/01638539209544801>
- OpenAI, et al. (2023). GPT-4 Technical Report. *arXiv*. <https://doi.org/10.48550/arXiv.2303.08774>
- Pandey, P. R., & Herrmann, B. (2025). The influence of semantic context on the intelligibility benefit from speech glimpses in younger and older adults. *Journal of Speech, Language, and Hearing Research, 68*(5), 2499–2516. https://doi.org/10.1044/2025_JSLHR-24-00588
- Panella, R. A., Barnett, A. J., Barense, M. D., & Herrmann, B. (2025). Event segmentation applications in large language model enabled automated recall assessments. *arXiv*. <https://doi.org/10.48550/arXiv.2502.13349>
- Peelle, J. E., & Wingfield, A. (2005). Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech. *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1315–1330. <https://doi.org/10.1037/0096-1523.31.6.1315>
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP):1532-1543*.
- Phillips, I., Bieber Rebecca, E., Dirks, C., Grant Ken, W., & Brungart Douglas, S. (2024). Age impacts speech-in-noise recognition differently for nonnative and native listeners. *Journal of Speech, Language, and Hearing Research, 67*(5), 1602–1623. https://doi.org/10.1044/2024_JSLHR-23-00470

- Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear & Hearing, 37*(Suppl 1), 5S–27S. <https://doi.org/10.1097/aud.0000000000000312>
- Pichora-Fuller, M. K., & Levitt, H. (2012). Speech comprehension training and auditory and cognitive processing in older adults. *American Journal of Audiology, 21*(2), 351–357. [https://doi.org/10.1044/1059-0889\(2012/12-0025\)](https://doi.org/10.1044/1059-0889(2012/12-0025))
- Pitts, B. L., Eisenberg, M. L., Bailey, H. R., & Zacks, J. M. (2023). Cueing natural event boundaries improves memory in people with post-traumatic stress disorder. *Cognitive Research: Principles and Implications, 8*(1), 26. <https://doi.org/10.1186/s41235-023-00478-x>
- Polspoel, S., Moore, D. R., Swanepoel, D. W., Kramer, S. E., & Smits, C. (2025). Automatic development of speech-in-noise hearing tests using machine learning. *Scientific Reports, 15*(1), 12878. <https://doi.org/10.1038/s41598-025-96312-z>
- Portnuff, C., & Bell, B. (2019). Effective use of speech-in-noise testing in the clinic. *The Hearing Journal, 72*(5), 40–43. <https://doi.org/10.1097/01.HJ.0000559502.51932.b1>
- Racchah, O., Chen, P., Gureckis, T., Poeppel, D., & Vo, V. (2024). The naturalistic free recall dataset: Four stories, hundreds of participants, and high-fidelity transcriptions. *Scientific Data, 11*(1), 1317. <https://doi.org/10.1038/s41597-024-04082-6>
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv*. <https://doi.org/10.48550/arXiv.2212.04356>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. <https://d4mucfksywwcloudfrontnet/better-language-models/language-models.pdf>
- Regev, J., Oxenham, A. J., Relaño-Iborra, H., Zaar, J., & Dau, T. (2025). Evaluating the role of age on speech-in-noise perception based primarily on temporal envelope information. *Hearing Research, 460*, 109236. <https://doi.org/10.1016/j.heares.2025.109236>
- Richmond, L. L., Gold, D. A., & Zacks, J. M. (2017). Event perception: Translations and applications. *Journal of Applied Research in Memory and Cognition, 6*(2), 111–120. <https://doi.org/10.1016/j.jarmac.2016.11.002>
- Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and benefits of large language models for the environment. *Environmental Science & Technology, 57*(9), 3464–3466. <https://doi.org/10.1021/acs.est.3c01106>
- Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics, 2*(9), 437–442. <https://doi.org/10.3758/BF03208784>
- 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. *Frontiers in Neuroscience, 10*(Article 621). <https://doi.org/10.3389/fnins.2016.00621>
- Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., Eisenberg, M. L., & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition, 129*(2), 241–255. <https://doi.org/10.1016/j.cognition.2013.07.002>
- Sasmita, K., & Swallow, K. M. (2022). Measuring event segmentation: An investigation into the stability of event boundary agreement across groups. *Behavior Research Methods, 55*(1), 428–447. <https://doi.org/10.3758/s13428-022-01832-5>
- Schneider, B. A., Daneman, M., & Pichora-Fuller, K. M. (2002). Listening in aging adults: From discourse comprehension to psychoacoustics. *Canadian Journal of Experimental Psychology, 56*(3), 139–152. <https://doi.org/10.1037/h0087392>
- Shen, X., Houser, T., Smith, D. V., & Murty, V. P. (2023). Machine-learning as a validated tool to characterize individual differences in free recall of naturalistic events. *Psychonomic Bulletin & Review, 30*(1), 308–316. <https://doi.org/10.3758/s13423-022-02171-4>
- Shukla, A., Harper, M., Pedersen, E., Goman, A., Suen, J. J., Price, C., Applebaum, J., Hoyer, M., Lin, F. R., & Reed, N. S. (2020). Hearing loss, loneliness, and social isolation: A systematic review. *Otolaryngology - Head and Neck Surgery, 162*(5), 622–633. <https://doi.org/10.1177/0194599820910377>
- Slaney, M., & Fitzgerald, M. B. (2024). Comparing human and machine speech recognition in noise with QuickSIN. *JASA Express Letters, 4*(9), 095202. <https://doi.org/10.1121/10.0028612>
- Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2004). Perceiving narrated events. *Proceedings of the Annual Meeting of the Cognitive Science Society 26:1637*. <https://doi.org/doi>
- Summerfield, Q. (1987). Speech perception in normal and impaired hearing. *British Medical Bulletin, 43*(4), 909–925. <https://doi.org/10.1093/oxfordjournals.bmb.a072225>
- Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1589–1625. <https://doi.org/10.1037/0278-7393.26.6.1589>
- Tenzer, H., Feuerriegel, S., & Piekkari, R. (2024). AI machine translation tools must be taught cultural differences too. *Nature, 630*(8018), 820–820. <https://doi.org/10.1038/d41586-024-02091-4>
- Tsuji, S., & Arai, T. (2025). The impact of reverberation on speech intelligibility and adaptation for people with unilateral hearing loss. *The Journal of the Acoustical Society of America, 157*(4), 3070–3086. <https://doi.org/10.1121/10.0036462>
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. In *Proc. 9th ISCA workshop on speech synthesis workshop (SSW 9)* (pp. 125).
- van Genugten, R. D. I., & Schacter, D. L. (2024). Automated scoring of the autobiographical interview with natural language processing. *Behavior Research Methods, 56*(3), 2243–2259. <https://doi.org/10.3758/s13428-023-02145-x>

- van Wijngaarden, S. J. (2001). Intelligibility of native and non-native Dutch speech. *Speech Communication*, 35(1-2), 103–113. [https://doi.org/10.1016/S0167-6393\(00\)00098-4](https://doi.org/10.1016/S0167-6393(00)00098-4)
- van Wijngaarden, S. J., Steeneken, H. J. M., & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *The Journal of the Acoustical Society of America*, 111(4), 1906–1916. <https://doi.org/10.1121/1.1456928>
- Wang, S., Wong Lena, L. N., & Shen, X. (in press). The effect of dialect and accent on digit perception in noise in young listeners with normal hearing. *Journal of Speech, Language, and Hearing Research*, 68(5), 2584–2596. https://doi.org/10.1044/2025_JSLHR-24-00472
- Wilson, R. H., McArdle, R. A., & Smith, S. L. (2007). An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. *Journal of Speech, Language, and Hearing Research*, 50(4), 844–856. [https://doi.org/10.1044/1092-4388\(2007\)059](https://doi.org/10.1044/1092-4388(2007)059)
- Wilson, R. H., McArdle, R. A., Watts, K. L., & Smith, S. L. (2012). The revised speech perception in noise test (R-SPIN) in a multiple signal-to-noise ratio paradigm. *Journal of the American Academy of Audiology*, 23(8), 590–605. <https://doi.org/10.3766/jaaa.23.7.9>
- Yang, Z., Yang, Y., Cer, D., Law, J., & Darve, E. (2020). Universal sentence representation learning with conditional masked language model. *arXiv*. <https://doi.org/10.48550/arXiv.2012.14388>
- Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2), 273–293. <https://doi.org/10.1037/0033-2909.133.2.273>
- Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, 16(2), 80–84. <https://doi.org/10.1111/j.1467-8721.2007.00480.x>
- Zhang, Q., Griffiths, T. L., & Norman, K. A. (2023). Optimal policies for free recall. *Psychological Review*, 130(4), 1104–1124. <https://doi.org/10.1037/rev0000375>
- Zhao, R., Choi, A. S. G., Koenecke, A., & Rameau, A. (2025). Quantification of automatic speech recognition system performance on d/deaf and hard of hearing speech. *The Laryngoscope*, 135(1), 191–197. <https://doi.org/10.1002/lary.31713>
- Zhelezniak, V., Savkov, A., Shen, A., & Hammerla, N. (2019). Correlation coefficients and semantic textual similarity. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 951–962).
- Zhou, Y., Dai, S., Cao, Z., Zhang, X., & Xu, J. (2024). Length-induced embedding collapse in transformer-based models. *arXiv*. <https://doi.org/10.48550/arXiv.2410.24200>