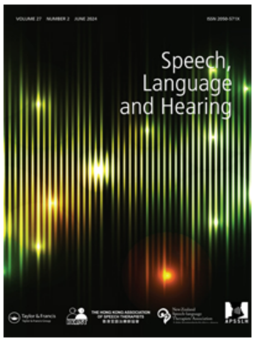

Taylor & Francis
Taylor & Francis Group

Speech, Language and Hearing

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/yslh20

Leveraging natural language processing models to automate speech-intelligibility scoring

Björn Herrmann

To cite this article: Björn Herrmann (07 Jul 2024): Leveraging natural language processing models to automate speech-intelligibility scoring, Speech, Language and Hearing, DOI: 10.1080/2050571X.2024.2374160

To link to this article: https://doi.org/10.1080/2050571X.2024.2374160

SPEECH, LANGUAGE AND HEARING
https://doi.org/10.1080/2050571X.2024.2374160

Taylor & Francis
Taylor & Francis Group

Leveraging natural language processing models to automate speech-intelligibility scoring

Björn Herrmann[a,b]

[a]Baycrest Academy for Research and Education, Rotman Research Institute, North York, Canada; [b]Department of Psychology, University of Toronto, Toronto, Canada

ABSTRACT

Assessment of speech intelligibility in noise is critical for measuring the impact of age-related hearing loss. However, quantifying intelligibility often requires a human to manually process responses provided by a participant or patient to obtain a speech-intelligibility score – typically the proportion of correctly heard words. This manual process can be time-consuming and thus costly. The current study investigates whether state-of-the-art Natural Language Processing (NLP) models from Google and OpenAI could be used to calculate speech-intelligibility scores as an alternative to human scoring. It was specifically tested whether NLP models capture common speech-in-noise perception phenomena in younger and older adults (N = 144) listening to speech masked by modulated or unmodulated babble noise. The results show that NLP speech-intelligibility scores closely matched intelligibility scores from a human scorer (r ∼0.95). The main difference is, on average, ∼2% underestimation of NLP intelligibility scores relative to human intelligibility scores for moderate to high signal-to-noise ratios. This underestimation results from participants making minor errors related to misspellings, gender, or tense, to which NLP models are sensitive, but human scorers typically correct prior to scoring. Critically, NLP models capture the known age-related reduction in intelligibility and the age-related reduction in the benefit from a modulated relative to an unmodulated masker. OpenAI's ADA2 appears to perform the best out of the tested NLP models, showing no difference in the speech-in-noise phenomena compared to human scoring. The current study suggests that modern NLP models can be used to score speech-intelligibility data.

ARTICLE HISTORY
Received 8 November 2023
Accepted 24 June 2024

KEYWORDS
Speech intelligibility; speech in noise; aging; natural language processing; Google; OpenAI

Introduction

Assessing a person's ability to understand speech is critical in basic science and clinical contexts, for example, for the understanding and assessment of the impact of age-related hearing loss (Bilger, Nuetzel, Rabinowitz, & Rzeczkowski, 1984; Nielsen & Dau, 2009; Nilsson, Soli, & Sullivan, 1994; Parmar, Rajasingam, Bizley, & Vickers, 2022; Wilson, 2003). A common approach to investigate a person's ability to understand speech is to quantify how well the person can understand individual words of a sentence, which is referred to as speech intelligibility (Dupuis & Pichora-Fuller, 2014; Ferguson, Jongman, Sereno, & Keum, 2010; Gustafsson & Arlinger, 1994; Hirsh, Reynolds, & Joseph, 1954; Holmes, Folkeard, Johnsrude, & Scollie, 2018b; Irsik, Johnsrude, & Herrmann, 2022a; 2022b; Mattys, Davis, Bradlow, & Scott, 2012; Miller, 2013; Nielsen & Dau, 2009; Nilsson et al., 1994; Ohlenforst et al., 2017; Ritz, Wild, & Johnsrude, 2022; Winn & Teece, 2021). However, scoring speech-intelligibility data from a participant or patient can be time-consuming (Borrie, Barrett, & Yoho, 2019; Bosker, 2021). The current study investigates whether state-of-the art, artificial-intelligence based tools can be used to automate speech-intelligibility scoring.

Typical experimental or audiological procedures involve a participant or patient listening to one sentence at the time and subsequently reporting back exactly what they heard (Cooke, Mayo, & Valentini-Botinhao, 2013; Herrmann, 2023; Irsik, Johnsrude, & Herrmann, 2022b; Miller, 2013). Sentences may be presented under varying degrees of background noise to assess the individual's ability to understand speech under unfavorable conditions that resemble, for example, a busy restaurant (Gustafsson & Arlinger, 1994; Herrmann, 2023; Irsik et al., 2022b; Winn & Teece, 2021). Individuals may report back verbally what they heard, which can be scored either immediately by a human experimenter/clinician or analyzed at a later stage (Dupuis & Pichora-Fuller, 2014; Gustafsson & Arlinger, 1994; Miller, 2013; Winn & Teece, 2021). In other settings, individuals interact with a computer, either by selecting words on a computer screen or by typing what they heard using a computer keyboard, e.g., 'type out what you hear' (Aoki, Cohn, & Zellou, 2022; Herrmann, 2023;

CONTACT Björn Herrmann bherrmann@research.baycrest.org Baycrest Academy for Research and Education, Rotman Research Institute, 3560 Bathurst St, North York, ON M6A 2E1, Canada

Supplemental data for this article can be accessed online at https://doi.org/10.1080/2050571X.2024.2374160.

Bosker, 2021; Chandrasekaran, Van Engen, Xie, Beevers, & Maddox, 2015; Cooke et al., 2013; Holmes, Domingo, & Johnsrude, 2018a; Holmes, To, & Johnsrude, 2021, Irsik et al., 2022a; 2022b).

The sentences that are used to assess speech intelligibility can be categorized into closed-set sentences and open-set sentences. Closed-set sentences follow a strict grammatical structure (e.g., 'Name verb number adjective noun') and are made of a limited set of words (Bolia, Nelson, Ericson, & Simpson, 2000; Gustafsson & Arlinger, 1994; Kidd, Best, & Mason, 2008). Speech-intelligibility scoring for closed-set sentences is relatively easy for a human scorer or can be automated in experimental procedures, for example, by asking participants to select keywords from a list presented on a computer screen (Holmes et al., 2018a; Holmes et al., 2021; Vigouroux & Miller, 2007). Closed-set sentences, however, have disadvantages because they vary little and are repetitive. A listener can learn the template structure and form expectations about the sentences to facilitate speech intelligibility, but such expectations cannot be formed in real-life situations. Closed-set sentences are thus not very ecologically valid (Sommers, Kirk, & Pisoni, 1997).

Open-set sentences are sentences that do not follow a strict grammatical structure (although strictness may vary), can be of varying length, and are created from an open-ended vocabulary (Bilger et al., 1984; Clopper, Pisoni, & Tierney, 2006; Gilbert, Tamati, & Pisoni, 2013; IEEE, 1969; McHenry & Parle, 2006; Nilsson et al., 1994; O'Neill, Parke, Kreft, & Oxenham, 2020). Open-set sentences are thus more ecologically valid than closed-set sentences. However, scoring of speech intelligibility for open-set sentences can be challenging, especially for responses made by typing the sentence heard, because scoring cannot be automated simply by sequentially comparing individual words of the original sentence with the words from the response. Scoring typically involves manual processing of participant responses by a human, because participants may report sentences only partially (e.g., only the last few words) or words may be slightly out of order. For typed responses, participants may also make spelling mistakes, past/present tense mistakes, or plural/singular mistakes. Human scorers can correct some of these mistakes, and they are often not considered errors of speech intelligibility per se. However, correcting mistakes and preparing participant responses manually is time consuming and thus costly (Borrie et al., 2019; Bosker, 2021). Manual processing of typed responses by a human can take 30 min per participant (own experience; see also Borrie et al., 2019; Bosker, 2021; depending on the number of sentences used in an experiment). Especially for studies conducted online, for which often a high number of participants is recorded (e.g., N = 200), manual processing of 30 min amounts to 100 h (or 2.5 weeks of full-time work; Borrie et al., 2019; Herrmann, 2023; Irsik et al., 2022a, 2022b). Automated approaches that avoid manual processing would be valuable for studies with many participants. Automated approaches would also be useful for studies using adaptive experimental procedures, for example, where intelligibility scores on previous trials determine the level of speech masking on subsequent trials.

Apart from a few in-house computer programs to automatically score participant transcripts that were not detailed and are not publicly available (Allison & Hustad, 2014; Wild, Vorperian Houri, Kent Ray, Bolt Daniel, & Austin, 2018), two works have suggested automated speech-intelligibility scoring approaches (Borrie et al., 2019; Bosker, 2021). Borrie et al.'s approach – called Autoscore – counts the number of words in a participant's response that match the words in a target transcript. The approach can be used with different rules as to what counts as an error. A human-made file with common spelling errors can be applied to correct some mistakes. Autoscore has been shown to perform well and the provided R code (and online web-based implementation) enables researchers and clinicians to use it (Borrie et al., 2019). However, Autoscore is not readily available in languages other than English (although scoring rules can be created), and spelling lists may need to be manually adapted for different studies. Moreover, research experiments are increasingly built using, for example, Python programming (i.e., PsychoPy; Peirce, 2007; Peirce et al., 2019), enabling real-time adjustment of speech conditions based on the participant's response. Autoscore, given its R code, cannot be easily used in this regard.

Bosker's approach aimed to remedy some of these drawbacks by using an approximate string matching procedure (Bosker, 2021). Python programming code was made publicly available. However, despite reporting a high correlation between the intelligibility scores from the automated procedure and the scores by a human, the method by Bosker overestimates intelligibility by about 15% in cases when participants start guessing what they heard, for example, under unfavorable conditions (Figure S1; also commented on in the original work; Bosker, 2021). This overestimation makes the approach not very useful in practice. Hence, a different procedure that can be used with different languages, estimates intelligibility accurately, does not need manual processing by a human, and can be used in Python may be useful to the community.

Natural Language Processing (NLP) models have drastically advanced over the past few years. NLP models provide latent representations of words and sentences (Cer et al., 2018; Devlin, Chang, Lee, & Toutanova, 2019; Mikolov, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014;

Radford et al., 2019), and models have been developed to fulfill a variety of tasks and analyses such as text completion, searches, clustering, classification, and comparisons. Models are trained on text data from different sources, such as Wikipedia, news webpages, question-answer webpages, discussion forums, or broad web scrapes, and may be further augmented with supervised data (Devlin et al., 2019; Radford et al., 2019). Many trained NLP models are publicly available, require only minimal Python code, and are provided for different languages.

Critically for the current study, NLP models can be used to perform text-similarity analyses (Yuan, Neubig, & Liu, 2021; Zhang, Kishore, Wu, Weinberger, & Artzi, 2020). To this end, a word or sentence is mapped onto a high-dimensional numerical vector – called embedding – that captures the semantic space of a word (word embedding) or sentence (sentence embedding) (Cer et al., 2018; Devlin et al., 2019; Mikolov et al., 2013; Pennington et al., 2014). The vectors of two words or sentences that are semantically similar correlate higher (e.g., 'shoe' vs 'sock') compared to the vectors of two words or sentences that are semantically less similar (e.g., 'shoe' vs 'table'). Such comparisons resemble the analysis of speech intelligibility, where a human scorer quantifies the similarity between the original speech segment and the speech segment produced by a participant. For example, Google's USE (Universal Sentence Encoder; Cer et al., 2018) and OpenAI's ADA2 (https://platform.openai.com/docs/) map a longer speech segment onto an embedding vector, whereas Google's BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) and OpenAI's GPT2 (Generative Pre-training Transformer; Radford et al., 2019) map a word (or 'wordpieces') onto an embedding vector. The current study explores whether speech-intelligibility scoring could be automated using sentence and word embeddings from NLP models, and the extent to which NLP-based intelligibility scoring mirrors human intelligibility scoring.

One area in which automated speech-intelligibility scoring may be useful for research and clinical purposes is speech-in-noise perception. For NLP-based intelligibility scoring to be useful, it should capture well-known effects of speech-in-noise perception to the same extent human intelligibility scoring captures them. For example, speech is more intelligible when it is masked by a masker with slow amplitude-envelope fluctuations (e.g., 4 Hz) compared to when speech is masked by a masker with a flat, unmodulated amplitude envelope (Cooke, 2006; Dubno, Horwitz, & Ahlstrom, 2002; Festen & Plomp, 1990; Herrmann, 2023; Irsik et al., 2022b; Li & Loizou, 2007; Miller & Licklider, 1950). This speech-intelligibility benefit for a modulated (i.e., fluctuating) relative to an unmodulated masker is thought to result from processing the speech fragments that are revealed when speech-masking is transiently released. The effect is thus referred to as 'release from masking' or 'masking release' (Bacon, Opie, & Montoya, 1998; Gustafsson & Arlinger, 1994; Irsik et al., 2022b). Moreover, for older adults, speech in noise is typically less intelligible and the intelligibility benefit from a modulated relative to an unmodulated masker is typically reduced (Bacon et al., 1998; Dubno et al., 2002, 2003; George, Festen, & Houtgast, 2006; Gustafsson & Arlinger, 1994; Herrmann, 2023; Irsik et al., 2022b; Lorenzi, Husson, Ardoint, & Debruille, 2006; Summers & Molis, 2004), at least for short, disconnected sentences (Irsik et al., 2022b). The reduced benefit from masking release in older adults is thought to result from age-related changes in temporal processing (Dubno, Horwitz, & Ahlstrom, 2003; George et al., 2006; Gnansia, Jourdes, & Lorenzi, 2008; Moore, 2008) and cognitive factors (Irsik et al., 2022b).

The current study uses existing data from younger and older adults (Herrmann, 2023) to investigate whether NLP-based intelligibility scoring captures the known masking-release benefit as well as the age-related decline in speech intelligibility and masking-release benefit. Speech-intelligibility scores are calculated using two sentence-embedding models (OpenAI's ADA2, Google's USE) and two word-embedding models (Google's BERT, OpenAI's GPT2). Analyses and results for ADA2 and BERT are detailed in this article, whereas only a summary is provided for the other models, because results were highly redundant among the four NLP models, with a slightly better performance of ADA2 and BERT over USE and GPT2, respectively.

Methods and materials

Participants

Two speech-intelligibility datasets of a previous study (Herrmann, 2023; Experiment 1 and 2) were used to investigate whether state-of-the art, artificial-intelligence (AI) based tools can be used to automate speech-intelligibility scoring. The experimental procedures from which the two datasets were derived were similar, but the speaker gender differed (details are described below). The two datasets enable separate analyses of the automated speech-intelligibility scoring procedure and thus provide insights about some degree of generalizability of the approach.

Dataset 1 (Experiment 1 in Herrmann, 2023) comprised usable data from 42 younger adults between 20–38 years (mean age: 29.3 years, 26 male, 16 female) and 35 older adults between 54–72 years (mean age: 61.3 years, 12 male, 23 female). Dataset 2 (Experiment 2 in Herrmann, 2023) comprised useable data from 35 younger adults between 24–38 years (mean age: 32.4 years, 21 male, 13 female, 1 nonbinary) and 32 older adults between 59–74 years (mean age: 64.9 years, 13 male, 19 female). Estimated audiometric pure-tone average thresholds based on a digit-in-noise perception task (Smits, Goverts, & Festen, 2013; Smits, Kapteyn, & Houtgast, 2004) were about 7 dB HL for younger and 17 dB HL for older adults (Herrmann, 2023).

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.

Experimental setup

Experiments were conducted online in an internet browser. Custom-written JavaScript/html scripts with jsPsych JavaScript libraries were used to implement experimental procedures (Version 7.2.1; de Leeuw, 2015). Scripts were stored at an online repository (https://gitlab.pavlovia.org) and hosted via Pavlovia (https://pavlovia.org/). Participants used a link to the Pavlovia platform, provided on the recruitment platform, to perform the experimental tasks. No specifications as to the type/brand of equipment participants should use (e.g., computer, screen, operating system, etc.) were provided, but participants were asked to use headphones. Participants set their computer volume to a comfortable level using a reference sound and all auditory stimuli were presented at this comfortable listening level (Herrmann, 2023).

Sentence materials and experimental procedures

Participants listened to 128 sentences from the Harvard sentence lists 1–15 (IEEE, 1969). Half of the sentences were spoken by a human, whereas the other half were computer-generated using Google's AI-based Wavenet text-to-speech synthesizer (https://cloud.google.com/text-to-speech/docs/wavenet; van den Oord et al., 2016) to investigate differences in intelligibility between human and AI speech in the previous work (Herrmann, 2023). For Dataset 1, sentences were spoken by a female native English speaker or computer-generated using a female voice. For Dataset 2, sentences were spoken by a male native English speaker or computer-generated using a male voice. Speech-intelligibility was very similar for human-spoken and computer-generated speech (Herrmann, 2023). Hence, sentences for different speech types were collapsed during analysis.

Sentences were presented either under clear conditions or in twelve-talker babble from the Revised Speech in Noise test (R-SPIN; Bilger, 1984). The babble noise masker was either unmodulated (i.e., relatively flat amplitude envelope) or sinusoidally amplitude modulated at a rate of 4 Hz (100% depth) to investigate speech intelligibility benefits associated with masking release (Dubno et al., 2002; Gustafsson & Arlinger, 1994; Irsik et al., 2022b; Summers & Molis, 2004). For sentences masked by background babble, the signal-to-noise ratio (SNR) between the speech signal and the background babble was manipulated by adjusting the level of the sentence relative to the babble masker for 7 different SNR levels (Dataset 1: −9, −6.67, −4.33, −2, + 0.33, + 2.67, +5 dB; Dataset 2: −11, −8.33, −5.67, −3, −0.33, 2.33, + 5 dB). All sentence/babble mixtures were normalized relative to the same root-mean square amplitude (RMS).

In four blocks, participants listened to four sentences for each of the two speech types (human, Wavenet), two masker types (unmodulated, modulated), and eight SNR levels. Speech types, masker types, and SNR levels were distributed such that each participant listened to each of the 128 sentences only once. To ensure intelligibility results are not confounded by specific sentences, 32 versions were generated across which the assignment of a sentence to different speech types, masker types, and SNR levels was systematically varied. Each participant was randomly assigned to one of the versions at the beginning of the experimental session.

For each trial, a fixation cross was presented on the computer screen while concurrently a sentence played. An input box occurred subsequently on the screen and participants were prompted as follows: 'Please type the words exactly as you heard them (even if you only understood parts of what was said)'. After they typed in their response, a 0.4 s blank screen was presented before the next trial started. Participants performed a brief 12-trial training block prior to the four experimental blocks.

Manual data processing for human speech-intelligibility scoring

Throughout the manuscript, the terms 'human scoring', 'human intelligibility scoring', or related wordings are used to refer to the processing pipeline described in this section, involving manual processing of participant responses. In detail, responses for which participants indicated that they did not understand what was said, such as 'nothing', 'gibberish', 'unintelligible', 'not understood', and alike were set to a no response '' using computer code. Responses made by participants were then processed manually such that different or omitted words were counted as errors, whereas words with minor misspellings, minor word-order errors, incorrect grammatical number (e.g., singular vs. plural), and incorrect grammatical tense (e.g., past vs present tense) were corrected to match the original sentence (Borrie et al., 2019; Bosker, 2021; Herrmann, 2023). Corrections took about 20–30 min per participant (cf. Borrie et al., 2019). After manual corrections were made, an automatized procedure using custom MATLAB scripts was employed. Every word per response was coded 0 or 1 depending on whether the word matched the corresponding word in the original sentence, and the proportion of correctly reported words was calculated. The proportion of correct words was averaged across sentences, separately for each SNR level and masker type (Datasets 1 and 2 were analyzed independently). Henceforth, the term 'human intelligibility score' is used to refer to the proportion of correct words so that matching terminology can be used for the outcome of the automated procedure using natural language models described below (i.e., NLP-based intelligibility score).

Speech-intelligibility scoring using natural language models

In the current study, OpenAI's ADA2 and Google's BERT were used to investigate whether sentence embeddings and word embeddings, respectively, can be used for intelligibility scoring (see also related approaches for evaluations of text generation; Yuan et al., 2021; Zhang et al., 2020). Google's Universal Sentence Encoder (USE; sentence embedding) and OpenAI's GPT2 (word embedding) as well as different versions of the models were also explored. However, because the results were largely similar among models, with USE and GPT2 performing slightly worse, only the results from ADA2 and BERT are presented in detail (for USE and GPT2 results, see Figure S3). Notes about the results for USE and GPT2 are provided alongside the results for ADA2 and BERT, and a summary table is provided at the end of the results section. Python code for all four approaches is publicly available at https://osf.io/bxysw/.

Preprocessing of sentences

Responses by participants were minimally preprocessed using custom MATLAB scripts (Bosker, 2021). Manual steps were avoided to ensure that the speech-intelligibility scoring approach is fully automated. As for the human intelligibility scoring, responses for which participants indicated that they did not understand what was said were set to a no response '' using computer code. Numbers provided as digits by the participant were converted to the corresponding word. Four word combinations were also corrected that led to some variability across participants (e.g., 'treetop' to 'tree top' or 'halfway' to 'half way'), although it appeared to not make a noticeable difference for the analyses. In addition, punctuations, such as periods, commas, and semicolons were removed, because they are interpreted by NLP models (Clark, Khandelwal, Levy, & Manning, 2019; Rogers, Kovaleva, & Rumshisky, 2020) but not during human scoring. A response made by a participant that matched exactly the original sentence apart from a comma can reduce the NLP intelligibility score, but not the human score. Letters of the original sentences and the response transcripts were converted to lowercase.

OpenAI's ADA2

OpenAI made publicly available a large language model for text comparisons, called 'text-embedding-ada-002' (https://platform.openai.com/docs/guides/embeddings; December 2022) and henceforth referred to as ADA2. ADA2 represents text input as a 1536-dimensional embedding vector. Embedding vectors were calculated for each original sentence and each response transcript. An intelligibility score was calculated as the Spearman correlation between the embedding vector for the original sentence and the embedding vector for the corresponding response transcript (Figure 1A). Note that Pearson correlation or cosine similarity lead to almost identical results (Figure S2, supplementary materials; $r > 0.999$; Zhelezniak, Savkov, Shen, & Hammerla, 2019). Spearman correlation is less sensitive to outliers, if there were any, than Pearson's correlation or cosine similarity (Zhelezniak et al., 2019). Moreover, correlation values for embedding vectors are in practice between 0 and 1 (rather than between − 1 and 1), because NLP embedding vectors are typically not negatively related to each other. As a result, the correlation values closely match the range for the proportion of correctly reported words (i.e., 0–1) and could thus be interpreted similarly. Henceforth, the term 'NLP intelligibility score' is used to refer to the Spearman correlation value that describes the relation between the embedding vector for the original sentence and the embedding vector for the corresponding response transcript.

Critically, a few embedding dimensions of OpenAI's ADA2 have very high embedding scores for all sentences (Figure 2). As a result, embedding vectors for seemingly unrelated sentences correlate relatively highly (e.g., a Spearman correlation of 0.457 for sentences: 'Smoky fires lack flame and heat.' and 'Rice is often served in round bowls.'). Since the correlation value is used as the intelligibility score, this bias would lead to high intelligibility scores, which appears to be particularly prominent for low SNRs (Figure S4), when participants provide more incorrect responses. This bias could affect the estimation of SNR thresholds and slopes (Figure S4). To account for this issue, embeddings were normalized by subtracting the mean embedding vector across the 128 original sentences from the raw embedding vector of each sentence and each response transcript (Mu &

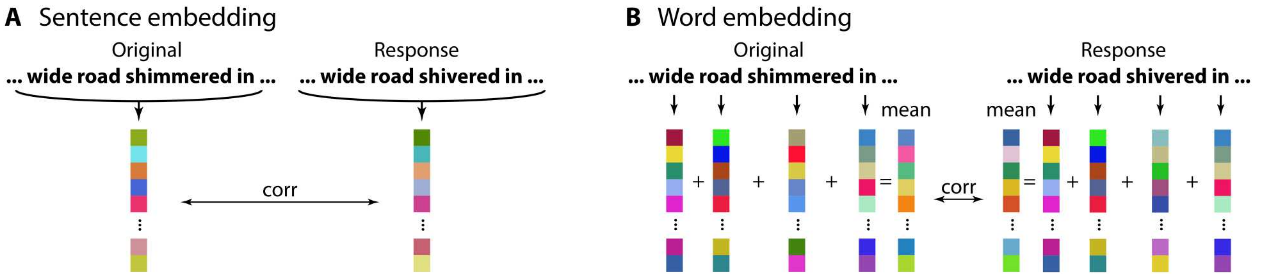
A Sentence embedding
Original
... wide road shimmered in ...
Response
... wide road shivered in ...
corr
B Word embedding
Original
... wide road shimmered in ...
mean
Response
... wide road shivered in ...
mean
corr

Figure 1. Schematic of sentence-embedding and word-embedding approaches to quantify speech intelligibility. A: Shows schematically the approach using sentence embeddings. A sentence is mapped onto a high-dimensional vector of real numbers (i.e., the embedding) using a large language model (Natural Language Processing [NLP] model). A vector provides a high-dimensional, semantic representation of a sentence. To obtain scores that quantify speech intelligibility, one numerical vector is calculated for the original sentence and one vector for the response made by a participant. The intelligibility score is then calculated as the Spearman correlation between the two vectors. The score can be interpreted similar to the proportion of correct words in a sentence. **B:** Shows schematically the approach using word embeddings. Each word of a sentence is mapped onto a high-dimensional vector of real numbers (i.e., the embedding) using a large language model. Vectors for individual words are averaged, separately for the original sentence and the response transcript, and the intelligibility score is then calculated as the Spearman correlation between the two averaged vectors.

Viswanath, 2018). Mean-subtraction has been shown to effectively remove the bias introduced by a few high embedding scores that are consistent across text inputs (Mu & Viswanath, 2018). Normalization was calculated prior to calculating the Spearman correlation. For the two unrelated, example sentences above, the Spearman correlation between the two normalized embedding vectors was 0.0004. Out of the four NLP models explored here, only Google's USE suffers little from this bias. For each participant, an intelligibility score (Spearman correlation) was calculated for each of the 128 sentences and scores were averaged across sentences, separately for each SNR and masker condition.

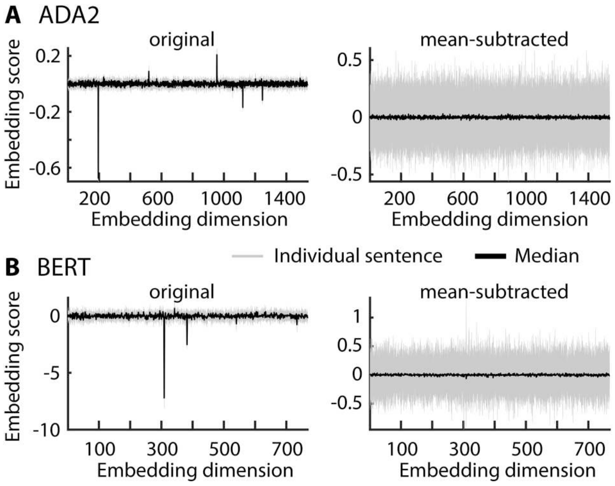

Figure 2. Embedding vectors for sentences used in the current study. A: Embeddings for OpenAI's ADA2. Both columns show embedding scores as a function of embedding dimensions for each individual sentence (light gray) and the median embedding score across sentences (black). The left column shows the original embedding scores derived from ADA2. Noteworthy are the 7 dimensions that exhibit very large scores across all sentences. The right column shows the normalized embedding scores, for which the mean embedding across sentences was subtracted from each individual sentence, effectively removing the large embedding scores that can bias analyses (Mu & Viswanath, 2018). **B:** Same as for panel A, using Google's Bidirectional Encoder Representations from Transformers (BERT) word embeddings (average scores across individual words per sentence are shown). As for ADA2, a few embedding dimensions of BERT show very high embedding scores that can be removed through mean-normalization (right column).

Google's bidirectional encoder representations from transformers (BERT)

The Bidirectional Encoder Representations from Transformers (BERT) is an NLP model developed by Google that is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context text (Devlin et al., 2019; Rogers et al., 2020). For BERT, the text input is tokenized into 'wordpieces' (Devlin et al., 2019; Rogers et al., 2020; Wu et al., 2016). For each token, BERT provides an embedding vector for each of the base model's 12 layers. This contrasts with ADA2, which provides one embedding vector for a longer string of text. Embeddings of the 12 layers were averaged. Initial examinations indicated that BERT intelligibility scores for the last 1/3 of layers differed more substantially from human intelligibility scores than BERT intelligibility scores calculated for the averaged embeddings across all 12 layers (or the average across the first 1/3 of layers). Later layers are thought to capture more context-specific representations (Ethayarajh, 2019; Rogers et al., 2020), which may impair performance for the current task to quantify speech intelligibility.

Embedding vectors for each token (i.e., word or 'wordpiece') were separately averaged for the original sentence and the participant's response transcript. An intelligibility score was calculated as the Spearman correlation between the averaged embedding vector (across tokens) for the original sentence and the averaged embedding vector (across tokens) for the response transcript (Figure 1B). Averaging embedding vectors is a simple approach that leads to good performance. Other approaches of combining token representations could be explored in future work (Rogers et al., 2020; Tanaka, Shinnou, Cao, Bai, & Ma, 2020; Toshniwal et al., 2020). As for ADA2, some of the embedding dimensions of BERT have very high embedding scores for all words/sentences (Figure 2), leading to biased intelligibility scores for low SNRs (Figure S4). Hence, prior to calculating the Spearman correlation, embeddings (averaged across tokens) were normalized by subtracting the mean embedding vector across the 128 original sentences from the embedding vector of each sentence and response (Mu & Viswanath, 2018). For each participant, an intelligibility score (Spearman correlation) was calculated for each of the 128 sentences and scores were averaged across sentences, separately for each SNR and masker condition.

Autoscore

Intelligibility scores using the Autoscore approach were also calculated using the R scripts provided along with the original publication (default settings; Borrie et al., 2019). The results for the Autoscore analyses are provided in Figure S5 in the supplementary information since the main focus of the current article was on NLP-based speech intelligibility. Autoscore results mirror the results for NLP models (Figure S5). The Fuzzy string-based approach (Bosker, 2021) was not explored in detail because it overestimates intelligibility for low SNRs (Figure S1).

Statistical analysis of intelligibility scores

Correlations between the 128 human intelligibility scores and the 128 NLP intelligibility scores per person were about 0.95 for both younger and older adults (Figure S6). However, correlations do not capture systematic differences across SNR levels between scoring types, for example, a substantial overestimation of intelligibility scores for the Fuzzy algorithm (Bosker, 2021; Figure S1) and NLP models for which embedding scores were non-normalized (Figure S4).

In order to investigate whether intelligibility scores differ between scoring types (human, NLP) at each of the different SNR levels, scores were averaged across masker types (modulated, unmodulated). A t-test was used to compare intelligibility scores between human and NLP scoring types (testing the human minus NLP difference against zero), separately for SNR levels and age groups (younger, older). False discovery rate (FDR; Benjamini & Hochberg, 1995; Genovese, Lazar, & Nichols, 2002) was used to correct for multiple comparisons across SNR levels.

To investigate whether the release-from-masking effect (i.e., the difference between modulated and unmodulated maskers) differs between scoring types (human, NLP) at any of the SNR levels, intelligibility scores for the modulated masker were subtracted from the intelligibility scores for the unmodulated masker, separately for SNR levels and age groups. The difference between masker types was then compared between scoring types (human, NLP) using a paired-samples t-test and FDR-thresholding (Benjamini & Hochberg, 1995; Genovese et al., 2002).

While contrasting human and NLP scoring types for each SNR level provides insights into where the two approaches diverge, the more common approach to analyze speech-intelligibility data for different SNR levels is to use psychometric function fits (Herrmann, 2023; Irsik et al., 2022b; Pichora-Fuller, 2008; Wu et al., 2012). To this end, for each participant, a logistic function was fit to the intelligibility scores, separately for each Masker Type (modulated, unmodulated) and Scoring Type (human, NLP) as a function of SNR level (excluding the clear condition), using the following equation:

$$y = \frac{1}{(1 + e^{-r(x-x_0)})}$$

where r is the slope, x_0 is the inflection point or the speech-reception threshold associated with 50% speech intelligibility, and x refers to the SNR levels (Dataset 1: $-9, -6.67, -4.33, -2, +0.33, +2.67, +5$ dB SNR; Dataset 2: $-11, -8.33, -5.67, -3, -0.33, 2.33, +$ 5 dB SNR). Slopes and thresholds were analyzed in repeated-measures analyses of variance (rmANOVAs) with Masker Type (unmodulated, modulated) and Scoring Type (human, NLP) as within-participants factors and Age Group (younger, older) as between-participants factor. Interactions were resolved using paired and independent samples t-tests. Analyses were conducted separately for Dataset 1 and Dataset 2.

All data analyses described were carried out using MATLAB (MathWorks), Python, and JASP software (JASP, 2023; version 0.16.4.0). Effect sizes for rmANOVAs and t-tests are reported using omega squared (ω^2) and Cohen's d (d), respectively.

Results for dataset 1

Analysis of intelligibility scores for each SNR

Figure 3 suggests that intelligibility scores were fairly similar between NLP and human scoring (left column). Correlations between scoring types were about 0.95 (Figure S6), but there were small systematic differences. On average, NLP intelligibility for ADA2

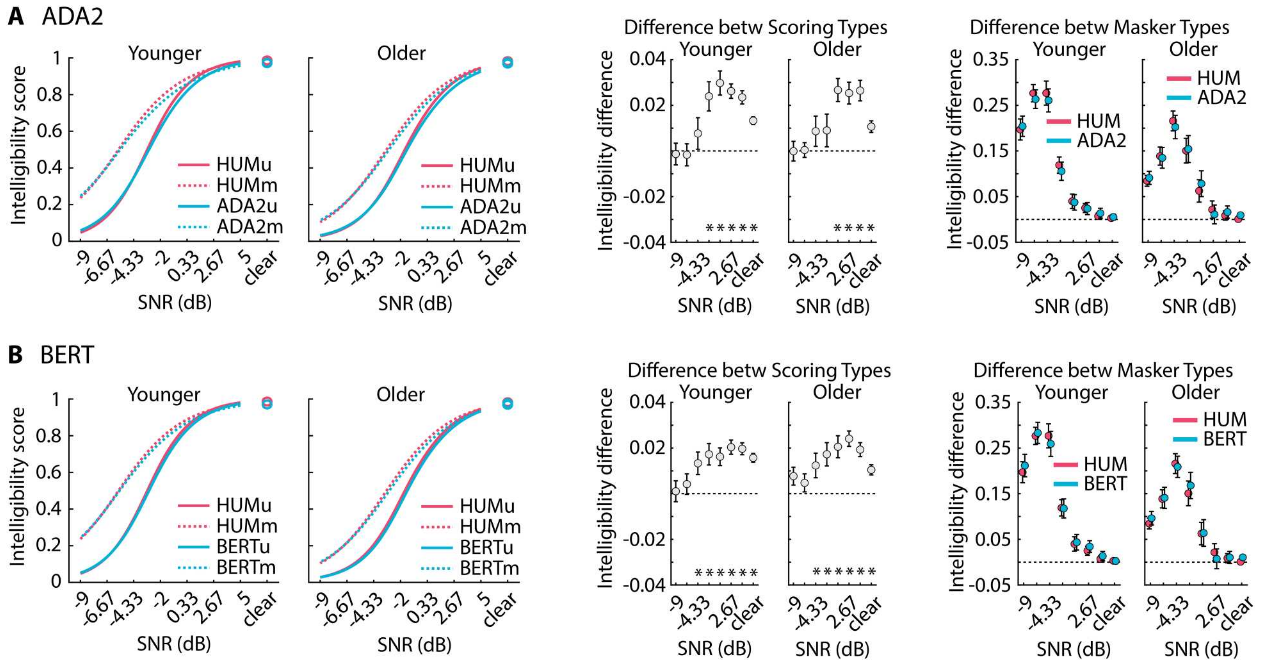

Figure 3. Intelligibility scores for sentence-in-noise listening using human and NLP scoring – Dataset 1. A: Data for OpenAI's ADA2. Left: Predicted intelligibility scores resulting from logistic function fits. Intelligibility scores are shown for sentences scored by a human (HUM) or ADA2 and for sentences presented in modulated (m) or unmodulated (u) background babble. Middle: Difference in intelligibility scores between scoring types (HUM minus ADA2). An asterisk indicates a significant difference between scoring types (i.e., from zero, $p \leq 0.05$, FDR-thresholded; Benjamini & Hochberg, 1995; Genovese et al., 2002). Right: Difference in intelligibility scores between masker types (modulated minus unmodulated). There were no differences between scoring types for the Masker Type effect (FDR-thresholded). **B:** Same as for panel A using Google's BERT.

and BERT was lower than human intelligibility by about 2% for several SNRs, especially those SNRs at which speech was more intelligible ($p_{FDR} \leq 0.05$; Figure 3, middle column). Critically, there was no difference in the effect of Masker Type (modulated minus unmodulated) between NLP scoring and human scoring at any of the SNR levels (Figure 3, right column). Results for USE and GPT2 were similar but the underestimation was slightly greater, up to 4%, for SNRs at which speech was more intelligible (Figure S3).

A qualitative assessment indicated that the ~2% underestimation of NLP scoring relative to human scoring for moderate to high SNRs resulted from spelling mistakes and errors associated with grammatical tense (present vs past tense) and number (singular vs plural). For example, a participant's response that incorrectly contains one word in its plural instead of its singular form leads to a minor reduction in the Spearman correlation between the NLP embedding vectors of the original sentence and the response transcript. Manual processing of the participant's response corrects this, leading to ~2% higher intelligibility scores on average. This seems to matter less when more word errors occur as SNR decreases. The NLP underestimation could possibly be reduced by correcting spelling mistakes at the pre-processing stage using programming code before calculating NLP intelligibility scores. Errors related to grammatical tense or gender are likely harder to correct through computer code without some template matching procedure. Hence, the underestimation of NLP models may be challenging to remove entirely. Nevertheless, participants did make an error in these cases, human scorers only tend to ignore such errors (although different scoring rules may be used in different studies; Borrie et al., 2019).

Slope analysis

A rmANOVA was calculated to investigate whether slopes differ between scoring types, masker types, and age groups. For ADA2, slopes were shallower (i.e., smaller values) for ADA2 scoring than human scoring (effect of Scoring Type: $F_{1,75} = 47.531$, $p = 1.5 \cdot 10^{-9}$, $\omega^2 = 0.005$) and shallower for modulated compared to unmodulated maskers (effect of Masker Type: $F_{1,75} = 32.397$, $p = 2.3 \cdot 10^{-7}$, $\omega^2 = 0.175$). The Scoring Type × Age Group interaction was significant ($F_{1,75} = 4.648$, $p = 0.034$, $\omega^2 = 0.003$), showing that the difference between Scoring Types was smaller for older compared to younger adults. No other effects or interactions were significant ($ps > 0.05$) Figure 4.

The analysis of BERT revealed, again, that slopes were shallower for BERT than human scoring (effect of Scoring Type: $F_{1,75} = 15.352$, $p = 2 \cdot 10^{-4}$, $\omega^2 = 0.008$) and shallower for modulated compared to unmodulated maskers (effect of Masker Type: $F_{1,75} =$

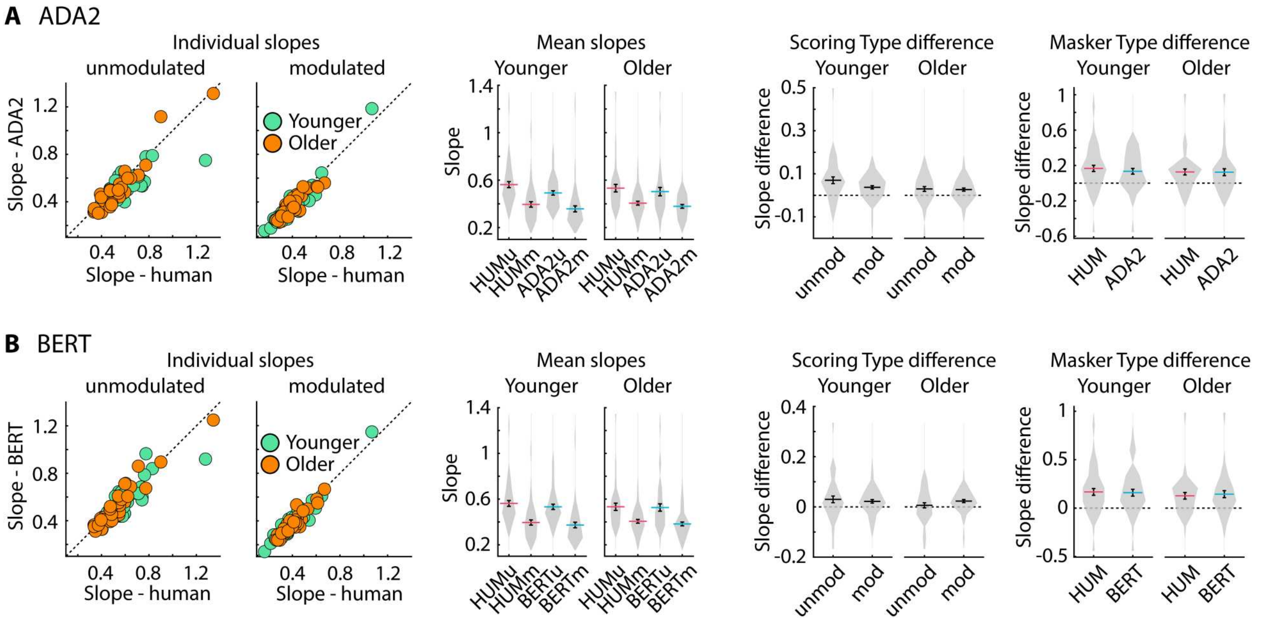

Figure 4. Slopes from logistic function fits to intelligibility scores – Dataset 1. A: Data for OpenAI's ADA2. First column: Scatter plots showing slopes for individual participants. Second column: Mean slopes and violin plots (histograms) are shown as colored horizontal line and gray shading, respectively. Third column: Difference in slopes between scoring types (HUM minus ADA2). Mean slopes (black line) and violin plots (gray shading) are shown. Fourth column: Difference in slopes between masker types (unmodulated minus modulated). Mean slopes (colored line) and violin plots (gray shading) are shown. Error bars reflect the standard error of the mean. **B:** Same as for panel A using Google's BERT.

36.997, $p = 4.6 \cdot 10^{-8}$, $\omega^2 = 0.195$). No other effects or interactions were significant (ps > 0.15).

In sum, the results across models (including USE and GPT2) show shallower slopes for NLP scoring compared to human scoring. This slope difference may not be surprising given the ~2% underestimation of NLP speech intelligibility scores for moderate to high SNRs.

Threshold analysis

For ADA2, thresholds were lower (i.e., better) for human scoring than ADA2 scoring by about 0.1–0.2 dB SNR (effect of Scoring Type: $F_{1,75} = 26.116$, $p = 2.4 \cdot 10^{-6}$, $\omega^2 = 0.005$), for modulated compared to unmodulated maskers (effect of Masker Type: $F_{1,75} = 239.250$, $p = 4.9 \cdot 10^{-25}$, $\omega^2 = 0.363$), and for younger compared to older adults (effect of Age Group: $F_{1,75} = 51.920$, $p = 3.8 \cdot 10^{-10}$, $\omega^2 = 0.251$). The difference between the modulated and the unmodulated masker was smaller in older compared to younger adults ($F_{1,75} = 10.807$, $p = 0.002$, $\omega^2 = 0.023$), indicating an age-related reduction in the masking-release benefit (c.f. Bacon et al., 1998; Gustafsson & Arlinger, 1994; Herrmann, 2023; Irsik et al., 2022b). No other effects or interactions were significant (ps > 0.1) Figure 5.

For BERT, thresholds were lower for human scoring than BERT scoring by about 0.1–0.2 dB SNR (effect of Scoring Type: $F_{1,75} = 35.993$, $p = 6.5 \cdot 10^{-8}$, $\omega^2 = 0.004$), for modulated compared to unmodulated maskers (effect of Masker Type: $F_{1,75} = 221.043$, $p = 4.6 \cdot 10^{-24}$, $\omega^2 = 0.359$), and for younger compared to older adults (effect of Age Group: $F_{1,75} = 52.636$, $p = 3.1 \cdot 10^{-10}$, $\omega^2 = 0.254$). The difference between the modulated and the unmodulated masker was smaller in older compared to younger adults ($F_{1,75} = 11.278$, $p = 0.001$, $\omega^2 = 0.025$), again, indicating an age-related reduction in masking-release benefit. No other effects or interactions were significant (ps > 0.1).

In sum, threshold data show that NLP scoring results in an overall 0.1–0.2 dB SNR increase (i.e., worsening) in the estimated speech-reception thresholds for ADA2 and BERT (0.3-0.5 dB SNR for USE; 0.2-0.3 dB SNR for GPT2). This threshold difference was relatively small compared to the effects of Age Group (1.6–2.6 dB SNR) and Masker Type (1.6–2.6 dB SNR). The threshold difference related to scoring types did not significantly differ between masker types and age groups, because interactions involving the factor Scoring Type were not significant (also for USE and GPT2). NLP scoring appeared as sensitive as human scoring to the known masker-type and age-group effects as well as to the known reduction in the release-from-masking in older compared to younger adults.

Results for dataset 2

The results for Dataset 2 largely mirror those for Dataset 1.

Analysis of intelligibility scores for each SNR

Figure 6 indicates that intelligibility scores were fairly similar between human and NLP scoring, and

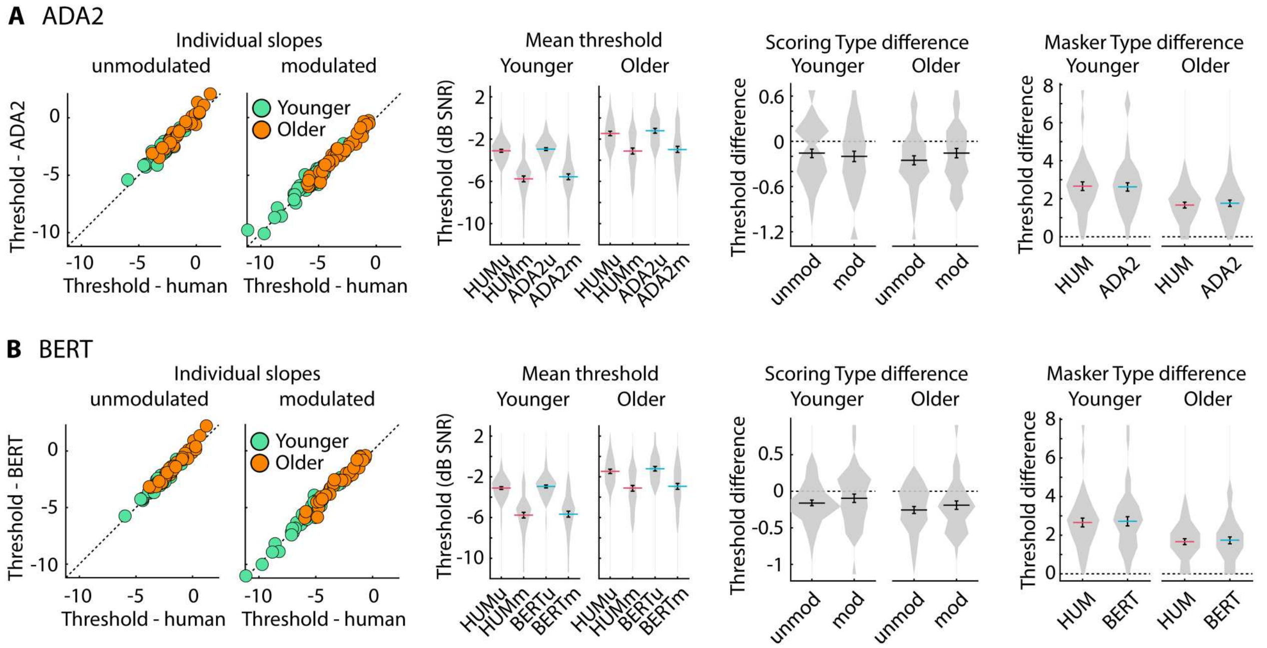

Figure 5. Thresholds from logistic function fits to intelligibility scores – Dataset 1. A: Data for OpenAI's ADA2. First column: Scatter plots showing thresholds for individual participants. Second column: Mean thresholds and violin plots (histograms) are shown as colored horizontal line and gray shading, respectively. Third column: Difference in thresholds between scoring types (HUM minus ADA2). Mean slopes (black line) and violin plots (gray shading) are shown. Fourth column: Difference in thresholds between masker types (unmodulated minus modulated). Mean slopes (colored line) and violin plots (gray shading) are shown. Error bars reflect the standard error of the mean. **B:** Same as for panel A using Google's BERT.

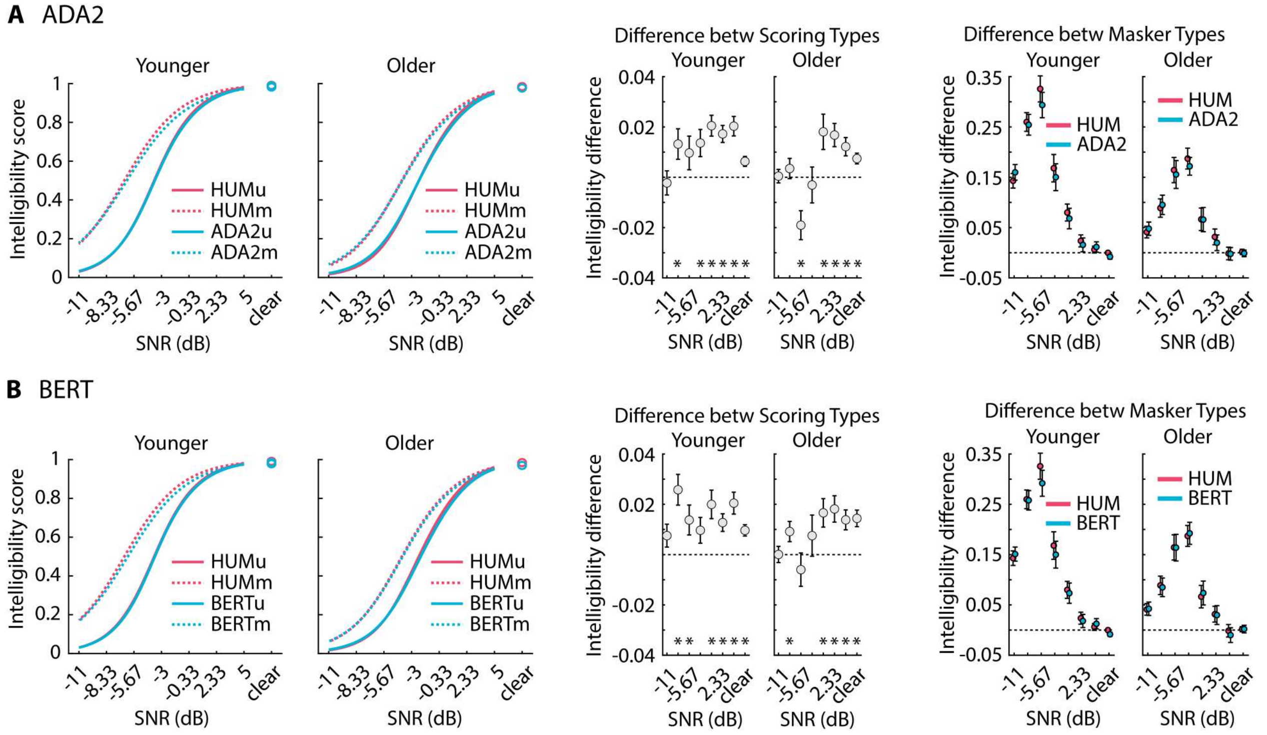

Figure 6. Intelligibility scores for sentence-in-noise listening using human and NLP scoring – Dataset 2. A: Data for OpenAI's ADA2. Left: Predicted intelligibility scores resulting from logistic function fits. Intelligibility scores are shown for sentences scored by a human (HUM) or ADA2 and for sentences presented in modulated (m) or unmodulated (u) background babble. Middle: Difference in intelligibility scores between scoring types (HUM minus ADA2). An asterisk indicates a significant difference between scoring types (i.e., from zero, $p \leq 0.05$, FDR-thresholded; Benjamini & Hochberg, 1995; Genovese et al., 2002). Right: Difference in intelligibility scores between masker types (modulated minus unmodulated). There were no differences between scoring types for the Masker Type effect (FDR-thresholded). **B:** Same as for panel A using Google's BERT.

correlations between scoring types were about 0.95 (Figure S6). Again, both NLP approaches underestimated intelligibility by ∼2% for moderate to highly intelligible SNR levels ($p_{FDR} \leq 0.05$; Figure 6, middle column). There was no difference in the effect of Masker Type (modulated minus unmodulated) between NLP and human scoring at any of the SNR levels (Figure 6, right column). Underestimation for USE and GPT2 was up to 4% and the Masker Type effect for USE and GPT2 was smaller at a few levels compared to human scoring.

Slope analysis

For ADA2, slopes were shallower (i.e., smaller values) for ADA2 than human scoring (effect of Scoring Type: $F_{1,65} = 21.344$, $p = 1.9 \cdot 10^{-5}$, $\omega^2 = 0.024$) and for modulated compared to unmodulated maskers (effect of Masker Type: $F_{1,65} = 34.900$, $p = 1.4 \cdot 10^{-7}$, $\omega^2 = 0.175$). No other effects or interactions were significant (ps > 0.1).

For BERT, shallower slopes were observed for BERT than human scoring (effect of Scoring Type: $F_{1,75} = 4.092$, $p = 0.047$, $\omega^2 = 0.004$) and for modulated compared to unmodulated maskers (effect of Masker Type: $F_{1,75} = 33.138$, $p = 2.5 \cdot 10^{-7}$, $\omega^2 = 0.190$). The Scoring Type × Masker Type × Age Group interaction was also significant ($F_{1,75} = 4.111$, $p = 0.047$, $\omega^2 = 0.003$), because the difference in slopes between BERT and human intelligibility for unmodulated maskers was numerically greater for older adults, whereas for modulated maskers, it was numerically greater for younger adults; but none of the effects were significant (ps > 0.05; Figure 7B, third column). The effect of Age Group and the other interactions were not significant (ps > 0.05).

In sum, slopes were shallower for NLP scoring compared to human scoring (as for Dataset 1), which is expected given the ∼2% underestimation of NLP speech intelligibility scores for moderate to high SNRs. Results were similar for USE and GPT2.

Threshold analysis

For ADA2, thresholds were lower for human than ADA2 scoring by about 0.05–0.25 dB SNR (effect of Scoring Type: $F_{1,65} = 7.820$, $p = 0.007$, $\omega^2 = 0.001$), for modulated compared to unmodulated maskers (effect of Masker Type: $F_{1,65} = 277.836$, $p = 3.7 \cdot 10^{-25}$, $\omega^2 = 0.378$), and for younger compared to older adults (effect of Age Group: $F_{1,65} = 39.370$, $p = 3.2 \cdot 10^{-8}$, $\omega^2 = 0.225$). The threshold difference between the modulated and the unmodulated masker was smaller in older compared to younger adults ($F_{1,65} = 22.828$, $p = 1.1 \cdot 10^{-5}$, $\omega^2 = 0.046$), showing the age-related reduction in masking-release benefit. The other interactions were not significant (ps > 0.05; Figure 8A).

For BERT, thresholds were lower for human scoring than BERT scoring by about 0.1–0.35 dB SNR (effect of Scoring Type: $F_{1,65} = 16.974$, $p = 1.1 \cdot 10^{-4}$, $\omega^2 = 0.004$), for modulated compared to unmodulated maskers (effect of Masker Type: $F_{1,65} = 298.056$, $p = 5.7 \cdot 10^{-27}$,

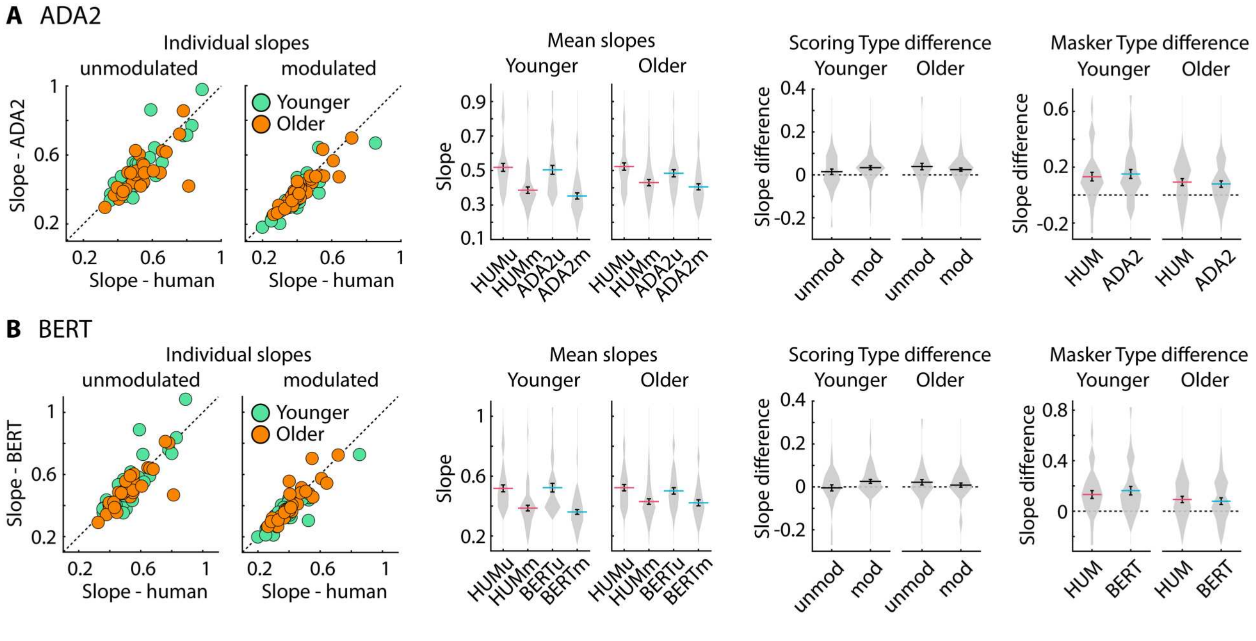

Figure 7. **Slopes from logistic function fits to intelligibility scores – Dataset 2. A:** Data for OpenAI's ADA2. First column: Scatter plots showing slopes for individual participants. Second column: Mean slopes and violin plots (histograms) are shown as colored horizontal line and gray shading, respectively. Third column: Difference in slopes between scoring types (HUM minus ADA2). Mean slopes (black line) and violin plots (gray shading) are shown. Fourth column: Difference in slopes between masker types (unmodulated minus modulated). Mean slopes (colored line) and violin plots (gray shading) are shown. Error bars reflect the standard error of the mean. **B:** Same as for panel A using Google's BERT.

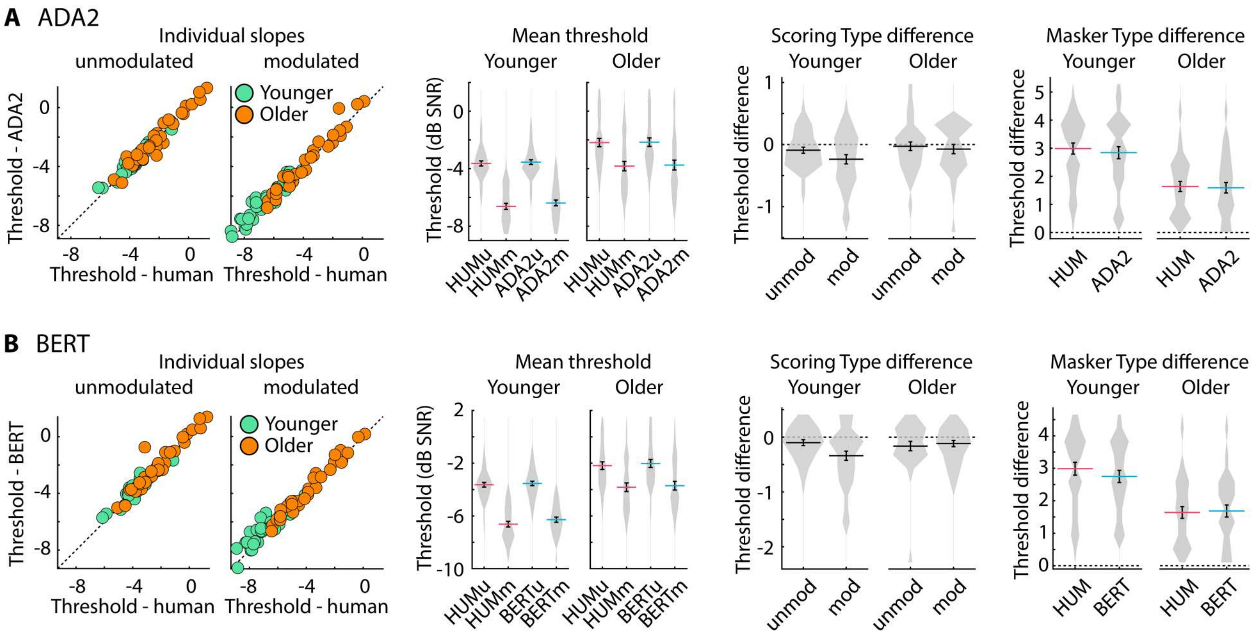

Figure 8. Thresholds from logistic function fits to intelligibility scores – Dataset 2. A: Data for OpenAI's ADA2. First column: Scatter plots showing thresholds for individual participants. Second column: Mean thresholds and violin plots (histograms) are shown as colored horizontal line and gray shading, respectively. Third column: Difference in thresholds between scoring types (HUM minus ADA2). Mean slopes (black line) and violin plots (gray shading) are shown. Fourth column: Difference in thresholds between masker types (unmodulated minus modulated). Mean slopes (colored line) and violin plots (gray shading) are shown. Error bars reflect the standard error of the mean. **B:** Same as for panel A using Google's BERT.

$\omega^2 = 0.382$), and for younger compared to older adults (effect of Age Group: $F_{1,65} = 40.238$, $p = 2.5 \cdot 10^{-8}$, $\omega^2 = 0.229$). The difference between the modulated and the unmodulated masker was smaller in older compared to younger adults ($F_{1,65} = 21.132$, $p = 2 \cdot 10^{-5}$, $\omega^2 = 0.040$). The Scoring Type × Masker Type × Age Group interaction was also significant ($F_{1,65} = 7.767$, $p = 0.007$, $\omega^2 < 0.001$; Figure 8B), while the other interactions were not (ps > 0.05). The Masker Type effect (unmodulated minus modulated) was smaller for BERT than human scoring for younger adults ($t_{34} = 3.892$, $p = 4.4 \cdot 10^{-4}$, $d = 0.099$), but not for older adults ($t_{31} = 0.559$, $p = 0.580$, $d = 0.099$).

In sum, results for Dataset 2 largely replicate results for Dataset 1, particularly for ADA2. Threshold data show that NLP scoring results in a ~0.2 dB SNR increase (i.e., worsening) in the overall estimated speech-reception thresholds. The threshold difference related to scoring types did not differ between masker types and age groups for ADA2, whereas it was smaller for BERT for younger, but not older adults. NLP scoring appeared as sensitive as human scoring to the known masker-type and age-group effects as well as to the known reduction in the release-from-masking in older compared to younger adults. ADA2 perhaps captures these effects best out of the models examined.

Results for USE and GPT2 were largely similar, although for both some of the interactions with Scoring Type were significant, making the picture more complicated for these two NLP models. Results for different versions of the models were also very similar. Moreover, the Autoscore approach based on one-to-one word comparisons (Borrie et al., 2019) also performed highly similarly to ADA2 and BERT (Figure S5; with the same biases). A summary of the results for all models (in different versions) and both datasets is provided in Table 1.

Discussion

Scoring speech-intelligibility data produced by a participant often requires manual processing steps carried out by a human that can be time-consuming and thus costly. The current study investigated the extent to which Natural Language Processing (NLP) models can be used to automate speech-intelligibility scoring, and thus provide an alternative approach to human intelligibility scoring. State-of-the-art NLP models that focus on word or sentence representations were applied to two datasets from an experiment where younger and older adults listened to speech masked by modulated or unmodulated background noise. The results show that human intelligibility scores correlated highly with NLP intelligibility scores (~0.95). NLP intelligibility scores were about 2% lower for moderate to high signal-to-noise ratios and estimated speech-reception thresholds higher by about 0.1-0.2 dB compared to human intelligibility scores. Critically, NLP intelligibility scoring captured known threshold differences between age groups (younger, older) and masker types (modulated,

Table 1. Summary of results for different models. For all results reported here, embedding vectors were mean-normalized to account for the high embedding scores at a few embedding dimensions that would bias the results (Figure 2). The three exceptions are the two USE models, for which no normalization was used because the bias for USE was minimal, and Autoscore. The first two rows show the results for the two models (bold) reported in more detail throughout the main text. In all four data columns, results for Datasets 1 and 2 are shown in the left and right sub-columns, respectively. The column 'Underestimation' shows the degree to which the model underestimated speech intelligibility relative to human scoring for moderate to high SNRs. The asterisk indicates that the model also showed an overestimation at low SNRs (~1-2%). The column 'Threshold difference' displays the overall threshold change for a model relative to human scoring. A larger value means a higher SNR for NLP/Autoscore than human scoring. The columns 'Interaction with Scoring Type' indicate which interactions were significant: 1 – Age Group × Soring Type, 2 – Masker Type × Soring Type, 3 – Age Group × Masker Type × Soring Type, the dash indicates none of the three interactions was significant. All models showed the known threshold effects of Age Group, Masker Type, and Age Group × Masker Type interaction, replicating previous work (Bacon et al., 1998; Dubno et al., 2003; Gustafsson & Arlinger, 1994; Irsik et al., 2022b). All models also showed a significantly shallower slope for modulated compared to unmodulated maskers and a shallower slope for NLP/Autoscore intelligibility scoring than for human intelligibility scoring, which is due to the underestimation of intelligibility at moderate to high SNRs. These effects are thus not listed in the table. ADA2 – OpenAI's ADA2, BERT – Bidirectional Encoder Representations from Transformers (base, large model), GPT2 – Generative Pre-Training Transformer (standard, large model), USE – Universal Sentence Encoder (model 4, model 5), Autoscore (Borrie et al., 2019).

Models	Underestimation (%)		Threshold difference (dB SNR)		Slope: Interaction with Scoring Type		Threshold: Interaction with Scoring Type	
ADA2	3	2	0.19	0.11	1	–	–	–
BERT (base) all layers	2	2	0.17	0.18	–	3	–	3
BERT (base) first 1/3 of layers	2*	2	0	0.02	–	–	2	–
BERT (base) last 1/3 of layers	5	5	0.55	0.54	–	–	–	2,3
BERT (large) all layers	4	4	0.42	0.44	–	–	–	2,3
BERT (large) first 1/3 of layers	2	2	0.08	0.09	–	–	2	3
BERT (large) last 1/3 of layers	8	8	0.99	0.99	1	–	2	1,2,3
GPT2	4	4	0.26	0.24	2	–	–	2
GPT2-large	4	4	0.27	0.37	2	–	–	2
USE4	4	3	0.43	0.35	–	–	–	1,2
USE5	3	3	0.33	0.23	3	–	–	2
USE4 non-normalized	2*	2*	0.04	0.07	–	3	–	–
USE5 non-normalized	2*	2*	0.1	0.13	1,3	–	–	–
Autoscore	2*	3*	0.11	0.17	2	–	–	–

unmodulated) as well as the reduced masker-type effect for older adults. For the best NLP model (OpenAI's ADA2), these effects did not differ from human scoring, whereas for the other NLP models there were minor differences. The current results show that NLP models provide an alternative to human intelligibility scoring.

NLP speech intelligibility is sensitive to common speech-in-noise phenomena

To investigate how well NLP models capture common speech-in-noise phenomena, participants listened to speech masked either by a modulated or an unmodulated masker. All NLP models tested here captured the previously reported phenomena that speech is more intelligible when it is masked by a modulated compared to an unmodulated masker (Bacon et al., 1998; Cooke, 2006; Dubno et al., 2002; Festen & Plomp, 1990; Gustafsson & Arlinger, 1994; Herrmann, 2023; Irsik et al., 2022b; Li & Loizou, 2007; Miller & Licklider, 1950), that older adults find masked speech less intelligible (Dupuis & Pichora-Fuller, 2014; Gustafsson & Arlinger, 1994; Irsik et al., 2022b; Presacco, Simon, & Anderson, 2019), and that older adults benefit less from a modulated relative to an unmodulated masker for speech intelligibility (Bacon et al., 1998; Dubno et al., 2002, 2003; George et al., 2006; Gustafsson & Arlinger, 1994; Herrmann, 2023; Irsik et al., 2022b; Lorenzi et al., 2006; Summers & Molis, 2004). The latter is thought to indicate that older adults are less able to capitalize on the speech glimpses released from the modulated masker (Dubno et al., 2003; George et al., 2006; Gnansia et al., 2008; Moore, 2008), although there is some indication that this might only be the case for short, disconnected sentences but not for engaging spoken stories (Irsik et al., 2022b). For OpenAI's ADA2 (and the non-normalized USE models; Table 1), these effects did not significantly differ from human scoring, whereas the other models show minor differences that were not fully consistent across models.

The current data demonstrate that intelligibility scores calculated with modern NLP models provide an alternative to human intelligibility scoring. Particularly for online experiments, where often several hundreds of participants are recorded, NLP scoring may provide a cost- and time-efficient way to obtain intelligibility scores. Moreover, experiments that comprise online feedback or adaptive procedures based on a participant's performance may also benefit from NLP intelligibility scores and the relatively simple Python code.

Potential future avenues to improve NLP speech intelligibility

Averaged NLP speech-intelligibility scores were relatively similar compared to human intelligibility scores with a few minor differences. The main difference was the ~2% underestimation of speech intelligibility by NLP models (ADA2, BERT) relative to human scoring for moderate to high SNRs, leading to shallower slopes and 0.1-0.2 dB SNR higher thresholds of the logistic function fits for NLP models. As discussed above, the difference appears to arise from small grammar and spelling errors made by participants. Human scorers manually correct such errors (Borrie et al., 2019), whereas NLP models are sensitive to them, leading to reduced intelligibility scores. The underestimation observed for NLP models may thus, in part, be related to human error correction, rather than the NLP models per se. Moreover, although human scores are generally highly accurate (Borrie et al., 2019; Hustad Katherine, 2006; Stilp, Kiefte, Alexander, & Kluender, 2010), humans also make occasional errors, such as scoring an incorrect word as correct or a correct word as incorrect (Borrie et al., 2019). Consistent with the NLP models, Autoscore also underestimated speech intelligibility by ~2% (Figure S5). Critically, speech-reception thresholds differed only minimally (0.1-0.2 dB) from human scoring, and this change in thresholds appeared consistent across conditions and age groups (particularly for ADA2).

MATLAB, Python, and other programming software comprise functions for spelling correction that could possibly be used to reduce the underestimation. However, spelling-correction functions sometimes change a real word to a different real word. Automated spelling correction would therefore need to be paired with a large external corpus to first categorize whether a word is a real word and then correct spelling errors only for non-words prior to NLP intelligibility calculations. A list of common misspellings of keywords could be another way of automating the correction of spelling mistakes (Borrie et al., 2019; Burleson & Souza, 2022). Grammatical gender and tense errors or writing errors that end up as a real word (e.g., 'faced' vs 'phased') are harder to correct automatically using computer programming code, and NLP underestimation may thus persist even when spelling errors are accounted for.

The procedures to calculate speech intelligibility highlight a more general issue for typed responses. Participants convert what they hear into a typed text response (Aoki et al., 2022; Herrmann, 2023; Chandrasekaran et al., 2015; Cooke et al., 2013; Holmes et al., 2018a; Holmes et al., 2021, Irsik et al., 2022a; 2022b). Mistakes of grammatical gender, grammatical tense, or writing, which may be corrected by a human scorer, could result from a participant hearing the speech incorrectly and typing the response as heard, or may be the result of correctly heard speech but an erroneously typed response. A human scorer correcting such mistakes assumes that the speech was heard correctly, but erroneously typed (or they consider such errors to be minor and not to contribute to speech intelligibility). The procedures used to obtain NLP intelligibility scores do not correct such mistakes and thus assume the response was typed in line with what the listener heard. The small differences between NLP and human scores may thus be related to assumption differences associated with the two scoring types, and are consistent with the recognized challenge to identify the exact level of speech intelligibility (Borrie et al., 2019; Bosker, 2021; Miller, 2013).

Another issue is that the resolution/discreteness of speech intelligibility values is lower when intelligibility is scored by a human compared to when it is calculated using an NLP model. That is, a sentence with 8 words corresponds to a relatively discrete resolution of 12.5% correctly reported words (1/8 * 100), whereas NLP speech-intelligibility scores theoretically have an infinite resolution depending on the participant's response (because intelligibility is calculated as the Spearman correlation between two high-dimensional vectors). The difference in score resolution/discreteness may additionally contribute to differences between scoring types. Nevertheless, NLP intelligibility scores could perhaps be used with fewer sentences than human scoring because of this higher resolution, but this needs to be explored in more detail in future studies.

Practical considerations when using NLP models to estimate speech intelligibility

The current study used data recorded online, where people typed what they heard in a text field using their computer keyboard. Typed responses are common in speech-intelligibility research (Aoki et al., 2022; Herrmann, 2023; Chandrasekaran et al., 2015; Cooke et al., 2013; Holmes et al., 2018a; Holmes et al., 2021, Irsik et al., 2022a; 2022b), but it is not the only approach that is frequently used. Researchers and clinicians may also score verbal responses in real time, while the participant or patient is in the lab or clinic, or record verbal responses for scoring at a later time (Dupuis & Pichora-Fuller, 2014; Gustafsson & Arlinger, 1994; Miller, 2013; Winn & Teece, 2021). Although NLP speech-intelligibility scoring may currently be most practical for written participant/patient responses, verbal responses could be transcribed to written text using modern AI-based speech-to-text converters prior to NLP intelligibility scoring. For example, OpenAI's Whisper (https://openai.com/index/whisper/; Radford et al., 2022) provides a powerful AI-based speech-to-text conversion for many different languages that requires only minimal Python coding and could be paired with the current intelligibility scoring approach to automate intelligibility scoring for verbal responses.

The strength of NLP models is that 1) a large amount of data can be scored with relatively high accuracy in a short period of time, 2) scoring can be performed for different languages, and 3) scoring can be automated for real-time use. Nevertheless, NLP scoring also has downsides. Paraphrases of a sentence using words that are not in the sentence would result in a low score by a human, whereas the NLP score may be higher because of remaining semantic similarities. The data from over 140 participants of the current study (including different genders and age groups) suggest that this may not be an issue, given the close match between the NLP and the human scoring (Figures 3 and 6), but there will be circumstances or populations for which human scoring is preferred to avoid scores to be potentially influenced by paraphrases. In addition, NLP intelligibility scores are not word identification scores. An NLP score reflects the (Spearman) correlation between the embedding vector of the original sentence and the vector of response sentence. Hence, in a strict sense, the 2% reduction for NLP relative to human intelligibility scoring does not mean that the NLP models scored 2 out of 100 words incorrectly (relative to a human). In practice, however, the correlation values range essentially from 0 to 1 (rather than from – 1–1), because NLP embedding vectors are typically not negatively related to each other. The correlation values (NLP scores) therefore closely match the range for the proportion of correctly reported words (see Figures 3 and 6) and could thus be interpreted similarly to proportion of correctly reported words in many, albeit likely not all contexts.

The current study suggests that OpenAI's ADA2 estimates speech intelligibility best out of the NLP models tested here (ADA2, USE, GPT2, BERT), using the human intelligibility scores as a reference point. The suggestion that ADA2 was slightly superior compared to the other models is based on the observation that all other NLP models, with the exception of the non-normalized USE models, showed some interaction with Scoring Type (human, NLP) in the rmANOVA for thresholds. ADA2 also showed a relatively small (0.1-0.2 dB SNR) speech-reception threshold difference relative to human scoring. It is noteworthy though that differences between models were minor and that BERT also performed well. Natural language processing models are evolving quite rapidly, and future models may perform even better.

While ADA2 appears to perform best in the current study, other criteria may also be considered when selecting an NLP model for speech intelligibility scoring, such as the complexity of the programming code (i.e., ease of implementation), computational time, and costs. Using OpenAI's ADA2 requires an account with OpenAI and an application programming interface (API) key, but the programming code to obtain embedding vectors is minimal. ADA2 calculations took about 1–2 min per participant (128 original sentences + 128 responses) on a Desktop computer (3.2 GHz Intel Core i7-8700 processor with 64 GB of RAM). Each processed token is associated with costs (linked to the OpenAI account), but costs are very minimal, amounting to about $0.2 CAD (i.e., 0.14€, $0.15 USD) for the current study, including running many iterations to examine the model's capacity to score intelligibility.

BERT, GPT2, and USE are not associated with direct costs. Programming code for USE and GPT2 is also minimal, whereas the code for BERT is a bit longer (see Python code https://osf.io/bxysw/). Calculations of speech intelligibility using GPT2 or USE were fast, taking about 30 s per participant. Calculation of BERT speech-intelligibility scores took about 10 min per participant. BERT, GPT2, and USE models can be downloaded to a local folder on the user's computer (https://huggingface.co/), which is currently not possible for ADA2. Storing and using an NLP model locally enables offline speech-intelligibility calculations and ensures that the model is available in the future should replications or recalculations be required.

Finally, the development and training of NLP models require massive computational infrastructure that can be associated with a large carbon footprint (Luccioni, Viguier, & Ligozat, 2023; Rillig, Ågerstrand, Bi, Gould, & Sauerland, 2023). Although the current approach uses NLP models that are already trained, and the carbon footprint is thus much smaller, the size of the carbon footprint could be another consideration when choosing an NLP model. Because the models tested here perform relatively equally, a user could choose a model with a lower carbon footprint without risking a decline in performance.

Conclusions

The current study investigated whether state-of-the-art Natural Language Processing (NLP) models could be used to automate speech-intelligibility scoring, which typically requires manual processing by a human and is thus costly. The results show that NLP models (OpenAI's ADA2 and GPT2, and Google's USE and BERT) provide very similar intelligibility scores compared to intelligibility scores by a human, although NLP models minimally underestimated intelligibility (~2%) for moderately to highly favorable speech-clarity conditions. Critically, NLP models captured known age-group (younger, older) and masker-type (modulated, unmodulated) differences as well as the reduced masker-type effect for older adults. The effects did not differ between human scoring and OpenAI's ADA2, whereas other NLP models showed minor differences. The current results show that NLP models provide an alternative to human intelligibility scoring.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Canada Research Chairs: [Grant Number CRC-2019-00156, 232733]; Natural Sciences and Engineering Research Council of Canada: [Grant Number RGPIN-2021-02602].

References

Allison, K. M., & Hustad, K. C. (2014). Impact of sentence length and phonetic complexity on intelligibility of 5-year-old children with cerebral palsy. *International Journal of Speech-Language Pathology*, *16*, 396–407.

Aoki, N. B., Cohn, M., & Zellou, G. (2022). The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise. *JASA Express Letters*, *2*, 045204.

Bacon, S. P., Opie, J. M., & Montoya, D. Y. (1998). The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds. *Journal of Speech, Language, and Hearing Research*, *41*, 549–563.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*, 289–300.

Bilger, R. C. (1984). *Manual for the clinical use of the revised SPIN test*. Champaign, IL, USA: The University of Illinois.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech, Language, and Hearing Research*, *27*, 32–48.

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, *107*, 1065–1066.

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*, 392–399.

Bosker, H. R. (2021). Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies. *Behavior Research Methods*, *53*, 1945–1953.

Burleson, A. M., & Souza, P. E. (2022). Cognitive and linguistic abilities and perceptual restoration of missing speech: Evidence from online assessment. *Frontiers in Psychology*, *13*, 1059192.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., St John, R., … Kurzweil, R. (2018). Universal sentence encoder. *arXiv*. doi:10.48550/arXiv.1803.11175

Chandrasekaran, B., Van Engen, K., Xie, Z., Beevers, C. G., & Maddox, W. T. (2015). Influence of depressive symptoms on speech perception in adverse listening conditions. *Cognition and Emotion*, *29*, 900–909.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.), *ACL workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 276–286). Florence, Italy: Association for Computational Linguistics.

Clopper, C. G., Pisoni, D. B., & Tierney, A. T. (2006). Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology*, *17*, 331–349.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*, 1562–1573.

Cooke, M., Mayo, C., & Valentini-Botinhao, C. (2013). Intelligibility-enhancing speech modifications: The hurricane challenge. In *Proceedings of interspeech* (pp. 3552–3556). Lyon, France. doi:10.21437/Interspeech.2013-764

de Leeuw, J. R. (2015). Jspsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, 1–12.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. doi:10.48550/arXiv.1810.04805

Dubno, J. R., Horwitz, A. R., & Ahlstrom, J. B. (2002). Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing. *The Journal of the Acoustical Society of America*, *111*, 2897–2907.

Dubno, J. R., Horwitz, A. R., & Ahlstrom, J. B. (2003). Recovery from prior stimulation: Masking of speech by interrupted noise for younger and older adults with normal hearing. *The Journal of the Acoustical Society of America*, *113*, 2084–2094.

Dupuis, K., & Pichora-Fuller, M. K. (2014). Intelligibility of emotional speech in younger and older adults. *Ear & Hearing*, *35*, 695–707.

Ethayarajh, K. (2019). *How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 55–65), Hong Kong, People's Republic of China: Association for Computational Linguistics.

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*, 153–162.

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, *88*, 1725–1736.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, *15*, 870–878.

George, E. L. J., Festen, J. M., & Houtgast, T. (2006). Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *120*, 2295–2311.

Gilbert, J. L., Tamati, T. N., & Pisoni, D. B. (2013). Development, reliability, and validity of PRESTO: A new high-variability sentence recognition test. *Journal of the American Academy of Audiology*, *24*, 026–036.

Gnansia, D., Jourdes, V., & Lorenzi, C. (2008). Effect of masker modulation depth on speech masking release. *Hearing Research*, *239*, 60–68.

Gustafsson, HÅ, & Arlinger, S. D. (1994). Masking of speech by amplitude-modulated noise. *The Journal of the Acoustical Society of America*, *95*, 518–529.

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.

Hirsh, I. J., Reynolds, E. G., & Joseph, M. (1954). Intelligibility of different speech materials. *The Journal of the Acoustical Society of America*, *26*, 530–538.

Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018a). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science*, *29*, 1575–1583.

Holmes, E., Folkeard, P., Johnsrude, I. S., & Scollie, S. (2018b). Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. *International Journal of Audiology*, *57*, 483–492.

Holmes, E., To, G., & Johnsrude, I. S. (2021). How long does it take for a voice to become familiar? Speech intelligibility and voice recognition are differentially sensitive to voice training. *Psychological Science*, *32*, 903–915.

Hustad Katherine, C. (2006). A closer look at transcription intelligibility for speakers with dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners. *American Journal of Speech-Language Pathology*, *15*, 268–277.

IEEE. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*, 225–246.

Irsik, V. C., Johnsrude, I. S., & Herrmann, B. (2022a). Neural activity during story listening is synchronized across individuals despite acoustic masking. *Journal of Cognitive Neuroscience*, *34*, 933–950.

Irsik, V. C., Johnsrude, I. S., & Herrmann, B. (2022b). Age-related deficits in dip-listening evident for isolated sentences but not for spoken stories. *Scientific Reports*, *12*, 5898.

JASP. (2023). JASP [Computer software]. In: https://jasp-stats.org/.

Kidd, G., Jr., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America*, *124*, 3793–3802.

Li, N., & Loizou, P. C. (2007). Factors influencing glimpsing of speech in noise. *The Journal of the Acoustical Society of America*, *122*, 1165–1172.

Lorenzi, C., Husson, M., Ardoint, M., & Debruille, X. (2006). Speech masking release in listeners with flat hearing loss: Effects of masker fluctuation rate on identification scores and phonetic feature reception. *International Journal of Audiology*, *45*, 487–495.

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*, 1–15.

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*, 953–978.

McHenry, M. A., & Parle, A. M. (2006). Construction of a set of unpredictable sentences for intelligibility testing. *Journal of Medical Speech – Language Pathology*, *14*, 269.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. doi:10.48550/arXiv.1301.3781

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders*, *48*, 601–612.

Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, *22*, 167–173.

Moore, B. C. J. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology*, *9*, 399–406.

Mu, J., & Viswanath, P. (2018). *All-but-the-top: Simple and effective postprocessing for word representations*. 6th International Conference on Learning Representations, Vancouver, Canada.

Nielsen, J. B., & Dau, T. (2009). Development of a danish speech intelligibility test. *International Journal of Audiology*, *48*, 729–741.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*, 1085–1099.

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., … Kramer, S. E. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, *351*, 68–79.

O'Neill, E. R., Parke, M. N., Kreft, H. A., & Oxenham, A. J. (2020). Development and validation of sentences without semantic context to complement the basic English lexicon sentences. *Journal of Speech, Language, and Hearing Research*, *63*, 3847–3854.

Parmar, B. J., Rajasingam, S. L., Bizley, J. K., & Vickers, D. A. (2022). Factors affecting the use of speech testing in adult audiology. *American Journal of Audiology*, *31*, 528–540.

Peirce, J. W. (2007). Psychopy – psychophysics software in python. *Journal of Neuroscience Methods*, *162*, 8–13.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195–203.

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) (pp. 1532–1543).

Pichora-Fuller, M. K. (2008). Use of supportive context by younger and older adult listeners: Balancing bottom-up and topdown information processing. *International Journal of Audiology*, *47*, 72–82.

Presacco, A., Simon, J. Z., & Anderson, S. (2019). Speech-in-noise representation in the aging midbrain and cortex: Effects of hearing loss. *PLoS One*, *14*, e0213899.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv*. doi: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://d4mucfpksywvcloudfrontnet/better-language-models/language-modelspdf

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*, 3464–3466.

Ritz, H., Wild, C. J., & Johnsrude, I. S. (2022). Parametric cognitive load reveals hidden costs in the neural processing of perfectly intelligible degraded speech. *The Journal of Neuroscience*, *42*, 4619–4628.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about How BERT works.

Transactions of the Association for Computational Linguistics, *8*, 842–866.

Smits, C., Goverts, S. T., & Festen, J. M. (2013). The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *The Journal of the Acoustical Society of America*, *133*, 1693–1706.

Smits, C., Kapteyn, T. S., & Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology*, *43*, 15–28.

Sommers, M. S., Kirk, K. I., & Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. *I: The Effects of Response Format. Ear Hear*, *18*, 89–99.

Stilp, C. E., Kiefte, M., Alexander, J. M., & Kluender, K. R. (2010). Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences. *The Journal of the Acoustical Society of America*, *128*, 2112–2126.

Summers, V., & Molis, M. R. (2004). Speech recognition in fluctuating and continuous maskers. *Journal of Speech, Language, and Hearing Research*, *47*, 245–256.

Tanaka, H., Shinnou, H., Cao, R., Bai, J., & Ma, W. (2020). Document classification by word embeddings of BERT. In L.-M. Nguyen, X.-H. Phan, K. Hasida, & S. Tojo (Eds.), *In: Computational linguistics* (pp. 145–154). Singapore: Springer Singapore.

Toshniwal, S., Shi, H., Shi, B., Gao, L., Livescu, K., & Gimpel, K. (2020). A cross-task analysis of text span representations. *arXiv*. doi:10.48550/arXiv.2006.03866

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., … Kavukcuoglu, K. (2016). *Wavenet: A generative model for Raw audio*. Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), (p. 125).

Vigouroux, J., & Miller, N. (2007). Intelligibility testing: Issues in closed versus open format scoring. *Newcastle and Durham Working Papers in Linguistics*, *12*, 83–95.

Wild, A., Vorperian Houri, K., Kent Ray, D., Bolt Daniel, M., & Austin, D. (2018). Single-word speech intelligibility in children and adults with down syndrome. *American Journal of Speech-Language Pathology*, *27*, 222–236.

Wilson, R. H. (2003). Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. *Journal of the American Academy of Audiology*, *14*, 453–470.

Winn, M. B., & Teece, K. H. (2021). Listening effort is not the same as speech intelligibility score. *Trends in Hearing*, *25*, 1–26.

Wu, C., Cao, S., Zhou, F., Wang, C., Wu, X., & Li, L. (2012). Masking of speech in people with first-episode schizophrenia and people with chronic schizophrenia. *Schizophrenia Research*, *134*, 33–41.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*. doi:10.48550/arXiv.1609.08144

Yuan, W., Neubig, G., & Liu, P. (2021). BARTScore: Evaluating generated text as text generation. *arXiv*. doi:10.48550/arXiv.2106.11520

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *arXiv*. doi:10.48550/arXiv.1904.09675

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). Minneapolis, Minnesota: Association for Computational Linguistics.