International Journal of Speech Technology
https://doi.org/10.1007/s10772-023-10027-y

The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults

Björn Herrmann[1,2]

Received: 22 July 2022 / Accepted: 19 February 2023

Abstract
Artificial intelligence (AI) based synthesized speech has become almost human-like, ubiquitous in everyday live (e.g., smart phones, grocery self-checkouts), and relatively easy to synthesize. This opens opportunities to use AI speech in research and clinical areas, such as hearing sciences, audiology, and speech pathology, where recordings of speech materials by voice actors can be time- and cost-intensive. However, much research thus far has focused on technological developments towards more human-like voices evaluated by younger adults. How older adults perceive AI speech is unclear. Using Google's Wavenet text-to-speech synthesizer, the current study explores whether AI speech can be used to investigate common speech-in-noise perception phenomena in younger and older adults. Speech intelligibility was recorded for human speech and synthesized speech masked by a modulated or an unmodulated multi-talker babble noise. For both human and AI speech, speech intelligibility was better for the modulated than the unmodulated masker (masking release), and this masking-release benefit was reduced in older adults. Release from masking effects were comparable between human and AI speech, suggesting that modern AI speech could be useful for hearing and speech research. The data further suggest that older adults recognize the presentation of AI speech less frequently, rate AI speech as more natural, and are less able to discriminate between human and AI speech compared to younger adults. Research on speech perception in older adults may thus especially benefit from modern AI-based synthesized speech because, to them, AI speech feels much like spoken by a human.

Keywords Aging · Speech intelligibility · Text-to-speech synthesis · Artificial intelligence

1 Introduction

Advances in artificial intelligence (AI) lead to an increasing number of encounters with computer-generated (AI) speech in our everyday lives, for example, in smart homes and phones (e.g., Amazon's Alexa, Google's Assistant, Apple' Siri), automated phone services, train announcements, and grocery self-checkouts (Ammari et al., 2019; Bentley et al., 2018; Buteau & Lee, 2021; Richter, 2020). Moreover, the tools to generate and manipulate AI-based synthesized speech become increasingly accessible to lay persons. This availability of tools to leverage AI speech could facilitate new advances in research and clinical disciplines, including hearing science, audiology, and speech pathology, in which creating and norming human-spoken speech materials for diagnostic and rehabilitative purposes is an ongoing challenge, and could potentially be substituted by AI speech synthesis. However, research thus far has focused on technology development towards human-like voices and the degree to which AI speech is perceived by young, normal-hearing listeners (e.g., Aoki et al., 2022; Cohn & Zellou, 2020; Cohn et al., 2020; Govender et al., 2019a, 2019b). How older adults perceive and experience AI speech is unclear and topic of the current study.

Research on synthesized speech is not new (Cooke et al., 2013; Drager et al., 2006; Greene et al., 1986; Li & Loizou, 2007; Raitio et al., 2014; Simpson & Hart, 1977), but only in recent years, with the advent of AI (e.g., van den Oord et al., 2016), has synthesized speech become less robotic, more human-like, and applicable in everyday life (Buteau & Lee, 2021; Richter, 2020). This advancement towards human-like, synthesized speech is also reflected in the development progression of the standard rating scale to evaluate the

✉ Björn Herrmann
bherrmann@research.baycrest.org

[1] Rotman Research Institute, Baycrest Health Sciences, 3560 Bathurst St, North York, ON M6A 2E1, Canada

[2] Department of Psychology, University of Toronto, Toronto, ON M5S 1A1, Canada

Published online: 13 March 2023

experience of listeners with synthesized speech—the Mean Opinion Score (MOS; Salza et al., 1996). The main focus of evaluation has originally been on the degree to which synthesized speech is intelligible (Salza et al., 1996). With high speech intelligibility becoming the norm, the scale has been gradually adapted (Polkosky & Lewis, 2003) to more and more focus on the degree to which synthesized speech resembles speech spoken by a human (MOS-X2; Mean Opinion Score Expanded Version 2; Lewis, 2018).

Google and Amazon, among others, provide access to their AI-based text-to-speech synthesizers, for example, through minimal computer programming (e.g., Python) that interfaces with the company's Application Programming Interfaces (APIs; e.g., Google Wavenet, https://cloud.google.com/text-to-speech/docs/wavenet; van den Oord et al., 2016). AI speech can be generated in different languages (e.g., > 30 for Google Wavenet), with different accents and speaker genders. State-of-the-art AI-based text-to-speech synthesizers further enable user manipulations of acoustic parameters, such as the overall speech rate and frequency, or emphasis, pauses, and volume of specific sections of speech (e.g., using Speech Synthesis Markup Language [SSML]; Taylor & Isard, 1997; see https://www.w3.org/TR/speech-synthesis11/). The relatively easy access to modern, AI-based text-to-speech synthesizers and the increasingly human-like nature of AI speech offer a solution for research and clinical applications that otherwise require time- and cost-intensive recording and norming of human-spoken speech materials. For example, global access to hearing-loss diagnosis and rehabilitation requires spoken speech materials in different languages. However, recordings from native-speaking voice actors and norming by hearing care providers can be challenging in different languages, and AI speech may provide an alternative. Even in research laboratories, synthesizing instead of recording speech materials may be time- and cost-effective. While AI speech could provide a cost-effective way to obtain speech materials, AI speech will only be useful if there are little to no meaningful differences in how human and AI speech is perceived.

A few recent studies have compared perception of human speech to AI-based synthesized speech in a variety of contexts (Aoki et al., 2022; Cohn & Zellou, 2020; Cohn et al., 2020). For example, studies have investigated the experienced valence of human vs AI speech (Cohn et al., 2020) or how humans produce speech when talking to voice-AI systems compared to humans (Cohn & Zellou, 2021; Cohn et al., 2021, 2022; Zellou et al., 2021). Most relevant for the current work, a few studies have investigated whether speech intelligibility differs between synthesized speech and human-spoken speech (Aoki et al., 2022; Cohn & Zellou, 2020; Govender et al., 2019a, 2019b). Intelligibility for speech masked by background sound has been suggested to be lower for synthesized speech compared to human-spoken speech (Aoki et al., 2022; Cooke et al., 2013; Simantiraki et al., 2018). However, the most recent AI speech synthesizers have not been available at the time some of the works were conducted, thus limiting this conclusion somewhat (Cooke et al., 2013; Simantiraki et al., 2018). Moreover, only a small number of masking levels (N = 1–3) have often been used in previous work (e.g., Aoki et al., 2022; Cohn & Zellou, 2020; Cooke et al., 2013; Govender et al., 2019a, 2019b). Presentation of a low number of masking levels can limit interpretations about whether AI speech is less intelligible than human speech, because intelligibility differences can be related to differences in speech-perception thresholds or speech-perception variability (i.e., the steepness of the relation between masking levels and intelligibility). Measuring the whole psychometric intelligibility continuum from very low to very high speech intelligibility (Irsik et al., 2022; Liu & Jin, 2019; MacPherson & Akeroyd, 2014; Masalski et al., 2021; Ross et al., 2021; Wingfield et al., 2000) enables a more detailed characterization of potential differences between AI and human speech. Finally, overall differences between computer-generated speech and human speech are also hard to interpret because speech materials typically vary on a number of acoustic parameters that may be unspecific to the human-vs-AI-speech contrast but could also be present for different human voices. Comparisons between AI and human speech may be most fruitful in interaction designs, where a difference between two conditions is compared between AI and human speech.

One area that lends itself well to interaction designs, has real-world relevance, and would potentially benefit from AI speech for research and clinical purposes is speech-in-noise perception. More than 25% of people aged 55 years or older live with some form of hearing impairment that makes speech comprehension in the presence of background sound challenging (Cruickshanks et al., 1998; Feder et al., 2015; Goman & Lin, 2016; Pichora-Fuller et al., 2016). Auditory speech materials are crucial for investigations of how individuals perceive speech in the presence of background noise and how speech-in-noise perception changes across the lifespan. Previous work has shown that speech masked by a masker whose amplitude envelope fluctuates at a slow modulation frequency (e.g., 4 Hz) is more intelligible than speech masked by a masker with a flat, unmodulated amplitude envelope (Cooke, 2006; Dubno et al., 2002; Festen & Plomp, 1990; Irsik et al., 2022; Li & Loizou, 2007; Miller & Licklider, 1950). The difference in speech intelligibility between these two conditions is referred to as "release from masking" or "masking release" (Bacon et al., 1998; Gustafsson & Arlinger, 1994; Irsik et al., 2022). It is further well established 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; Irsik et al., 2022;

Lorenzi et al., 2006; Summers & Molis, 2004), at least for short, disconnected sentences (Irsik et al., 2022). This reduced masking-release benefit for older adults is thought to be due to an age-related decline in temporal processing (Dubno et al., 2003; George et al., 2006; Gnansia et al., 2008; Moore, 2008), although cognitive and motivational factors may contribute as well (Irsik et al., 2022). Capitalizing on the known masking-release effect and the age-related reduction in masking-release benefit provides a fruitful avenue to investigate whether AI speech differs from human-spoken speech, because interactions involving speech type (AI vs human speech) are tested.

Perception of synthesized speech has mostly been investigated in young, normal-hearing adults (Aoki et al., 2022; Cohn & Zellou, 2020; Cohn et al., 2020; Cooke et al., 2013; Govender et al., 2019a, 2019b). However, older adults may be as much exposed to modern AI-based synthesized speech in their everyday lives as younger adults are. Older adults engage with automated phone services and grocery self-checkouts, listen to train announcement and navigation assistance, and use smart phones and homes. Moreover, hospital and other medical services, services in long-term care settings, and support systems for aging in place increasingly explore interactive technology for older adults that relies on AI speech (Kim, 2021; Milne-Ives et al., 2020; O'Brien et al., 2020). The strong research focus on AI-speech perception in younger adults could limit age-informed technological developments that could facilitate inclusive technology, and more knowledge about AI-speech perception in older adults is needed.

The current study has two goals. The first goal is to investigate whether known age-related changes in speech-in-noise intelligibility are similarly present for human vs AI speech (Experiments 1 and 2). To this end, younger and older adults listen to human-spoken and AI-based synthesized sentences masked by multi-talker background babble. The amplitude envelope of the masking babble is either flat (i.e., unmodulated) or amplitude modulated. The research is designed as an interaction paradigm, such that speech intelligibility is compared between human-spoken and AI speech as the difference between speech masked by modulated vs unmodulated maskers (release from masking). Moreover, release from masking is expected to be larger in younger compared to older adults (Bacon et al., 1998; George et al., 2006; Gustafsson & Arlinger, 1994; Irsik et al., 2022). The current study explores whether this interaction differs between human and AI speech. The second goal of the current study is to investigate directly whether younger and older adults differ in their perception of AI speech (Experiment 3). To this end, younger and older adults listen to sentences spoken by 10 different human speakers (5 male, 5 female) and synthesized sentences using 10 AI voices (5 male, 5 female), and decide whether sentences were computer-generated or spoken by a human. This allows us to explore the sensitivity of younger and older adults to discriminate human from AI speech.

2 General methods

Data for the three experiments of the current study are publically available at https://osf.io/4jufv/.

2.1 Participants

In each of the three experiments discussed below, participants were recruited on Amazon Mechanical Turk (MTurk) using the Cloud Research interface (formerly TurkPrime; Litman et al., 2017), restricting access to the study to individuals who indicated on MTurk/Cloud Research that they were born in the United States of America and that their native language is English. Participants who indicated in the demographic questionnaire that their native language was not English were excluded. Participants who self-reported having a history of neurological disease or using a hearing aid were also excluded. Demographic details about participants and the number of excluded participants are provided below for each individual experiment. Each participant provided informed consent before completing the experiment and received $8.5 USD (Experiments 1 and 2) or $7.5 USD (Experiment 3) at the same hourly rate for their participation. New participants were recruited for each experiment such that participants in any one of the experiments were barred from participating in any of the other experiments. The study was conducted in accordance with the Declaration of Helsinki and the Canadian Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans, and approved by the Research Ethics Board at the Baycrest Centre for Geriatric Care.

2.2 Experimental setup

All three experiments were conducted online in an internet browser. Experimental scripts were written in JavaScript/html, using jsPsych JavaScript libraries for precise stimulus control (Version 7.2.1; de Leeuw, 2015). The experiment code was stored at an online repository (https://gitlab.pavlovia.org) and hosted via Pavlovia (https://pavlovia.org/). Participants recruited via Amazon Mechanical Turk (MTurk) using the Cloud Research interface (Litman et al., 2017) were provided with a link to the Pavlovia 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.

2.3 Auditory setup

At the beginning of the experiment, participants completed a volume check to set their computer volume to a comfortable level. All participants reported using either in-ear phones or over-the-ear headphones. In addition to explicitly asking participants what means of sound delivery they used, they also completed a headphone-check procedure (Woods et al., 2017). During the headphone-check procedure, participants performed a three-interval forced choice task, in which they determined which of three consecutive 200-Hz sine tones was the quietest. The three tones differed such that one was presented at the comfortable listening level, one at –6 dB relative to the other two tones, and one at the comfortable listening level with a 180° phase difference between the left and right headphone channels (anti-phase tone). This task is relatively easy over headphones, but difficult over loudspeakers, because the pressure waves generated from an anti-phase tone interfere (Woods et al., 2017). If they were listening through loudspeakers, they would likely erroneously select the anti-phase tone as the quietest tone. Participants performed six trials, during which the soft tone occurred equally often in the three intervals (two times each). Data from participants who made more than two errors were excluded from data analysis. The number of data sets that were excluded is provided for each experiment below.

2.4 Hearing assessment

Hearing was assessed for each participant using self-report (subjective) ratings of general hearing abilities and hearing problems. Each participant rated the question "How would you rate your general hearing abilities?" and the statement "I often experience hearing problems." on an 11-point scale, ranging from 0 to 10. For the question concerning general hearing abilities, the endpoints of the scale referred to as "very poor" and "very good", respectively. For the statement concerning hearing problems, the endpoints referred to "strongly disagree" and "strongly agree", respectively. Age group differences in self-reported hearing was analyzed using Wilcoxon's rank sum test.

Objective hearing abilities were assessed using an online implementation of the digits-in-noise test (Smits et al., 2004, 2013). Digits-in-noise thresholds have been shown to correlate well with audiometrically-measured pure-tone average thresholds (0.5–4 kHz; $r > 0.7$; De Sousa et al., 2020; Koole et al., 2016; Potgieter et al., 2018). Digit-in-noise tests have been used and validated for testing in-lab (De Sousa et al., 2020; Koole et al., 2016; Smits et al., 2013) and remote testing, such as by telephone (Smits & Houtgast, 2005; Smits et al., 2004, 2006) or a smart phone application (Brown et al., 2019; Potgieter et al., 2018). In the current study, participants listened to digit triplets (e.g., 2-5-3) masked with 12-talker babble noise (Bilger, 1984). After the noise and digit triplet ended, participants typed the digits they heard in the order the digits were presented. The signal-to-noise ratio (SNR) was manipulated by varying the level of the spoken digits relative to the level of the babble noise. Digit triplets were presented at 29 SNRs (range: −18 dB to +15.6 dB; step size: 1.2 dB). One-hundred digit triplets were pre-generated for each of the 29 SNRs by randomly selecting three different digits ranging from 1 to 9 (onset-to-onset interval: 0.85 s). The babble masker had a duration of 3 s and the sequence of digits started 0.5 s after babble onset. At the beginning of an experimental session, twenty-nine digit triplets (one per SNR) were randomly selected for each participant. Each participant completed two practice trials with high SNRs followed by the 29 test trials. For data analysis, a trial was only considered correct if all three digits were typed in the order they were presented. A logistic function was fit to the data, and the resulting threshold was used as a dependent measure. An independent samples t-test was used to compare digit-in-noise thresholds between age groups.

3 Experiment 1

Experiment 1 aimed to explore whether known age-related changes in speech-in-noise perception are similarly present for human vs AI speech. To this end, speech-in-noise intelligibility thresholds were estimated for younger and older adults who listened to speech spoken by a human female voice or to computer-generated speech using a female voice of Google's Wavenet text-to-speech synthesizer. Both types of speech were masked by 12-talker background babble that was either modulated in amplitude at 4 Hz or unmodulated.

3.1 Methods

3.1.1 Participants

Forty-two younger (mean age: 29.3 years, age range: 20–38 years) and 35 older adults (mean age: 61.3 years, age range: 54–72 years) participated in Experiment 1. Out of the 42 younger adults, 26 identified as male and 16 as female. Out of the 35 older adults, 12 identified as male and 23 as female.

Data from additional participants were excluded if they met one or more of the following criteria: using a hearing aid ($N = 2$), having a history of a neurological disease ($N = 2$), incorrectly responding to more than two trials in the three-interval forced choice headphone-check ($N = 13$), reported that they were distracted during the experiment ($N = 6$), or reported being a non-native English speaker ($N = 0$). Participants who passed these criteria but had an average performance in the speech-in-noise intelligibility task that

was lower than one standard deviation from the group mean (separately for younger and older adults) were also excluded (N = 5). Exclusion of datasets from participants who showed low performance removed people with outlier speech-in-noise thresholds, but none of the statistical results reported below were qualitatively affected by inclusion or exclusion of these datasets. The current more conservative approach to data inclusion was chosen because data quality can be a concern in online studies (Chmielewski & Kucker, 2020; Eyal et al., 2021; Kennedy et al., 2020). Nevertheless, with appropriate data quality checks and strict data exclusion criteria (e.g., based on low performance), online research has generally been shown to lead to similar results as those from studies conducted in-person in the laboratory (Agley et al., 2022; Berinsky et al., 2014; Buchanan & Scofield, 2018; Chmielewski & Kucker, 2020; Gosling et al., 2004; Thomas & Clifford, 2017). Overall, datasets from 24 participants were excluded.

3.1.2 Materials and procedure for speech-in-noise intelligibility task

Materials for the speech-in-noise intelligibility task were 128 sentences from the Harvard sentence lists 1–15 (IEEE, 1969). Sentences were spoken by a human female speaker and computer-generated using Google's Wavenet text-to-speech synthesizer (https://cloud.google.com/text-to-speech/docs/wavenet; van den Oord et al., 2016).

The original speech of the human female speaker was recorded at a 44.1 kHz sampling frequency, whereas speech of the Google Wavenet voices are synthesized at 24 kHz. Human speech recordings were downsampled to 24 kHz using MATLAB's 'resample.m' function to match the sampling frequency of the Wavenet speech. For each sentence, Praat software (Boersma, 2001) was used to calculate the sentence duration and estimate the median fundamental frequency across time (F0; prosodic contour). Duration and median F0 estimates were used to adapt Google Wavenet parameters.

Transcripts of the 128 sentences were fed into Google's text-to-speech synthesizer using custom Python scripts (see Google's code example: https://cloud.google.com/text-to-speech/docs/create-audio) that interfaced with the Google Cloud text-to-speech API (Application Programming Interface). The Google Wavenet female voice "en-US-Wavenet-H" was used (https://cloud.google.com/text-to-speech/docs/voices). Speech of this voice is synthesized with a US-English accent and was subjectively selected to sound closest, compared to the other Wavenet female voices, to the human female voice. Synthesized sentences ended with a silent period of a few hundred milliseconds, which was removed using custom MATLAB scripts. The parameters 'speaking_rate' and 'pitch' were used to manipulate the duration and median F0 of the 128 Wavenet sentences, such that average duration and median F0 across the sets of 128 sentences did not differ statistically between human speech and Wavenet speech ($p > 0.6$; Fig. 1C). The variance of median F0s across sentences was smaller for Wavenet than human speech (Levene test: $p = 5.8 \times 10^{-5}$; there was no difference for duration variance: $p = 0.163$). Duration and median F0 of human-spoken sentence and Wavenet sentences significantly correlated (duration: $r = 0.764$, $p = 1 \times 10^{-25}$; median F0: $r = 0.427$, $p = 5 \times 10^{-7}$). Sample waveforms and corresponding cochleograms using a simple auditory-periphery model (McDermott & Simoncelli, 2011) are displayed in Figs. 1A and B, respectively.

Sentences were presented either under clear conditions or twelve-talker babble from the Revised Speech in Noise test (R-SPIN; Bilger, 1984) was added to the sentences as a masker. Babble noise 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., 2022; 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: $-9, -6.67, -4.33, -2, +0.33, +2.67, +5$ dB. All sentence/babble mixtures were normalized relative to the same root-mean square amplitude (RMS). All speech materials were saved as wav-file for presentation online.

In four blocks of stimulation, participants listened to four sentences for each of the two speech types (human, Wavenet), two masker types (unmodulated, modulated), and eight SNR levels (clear, $-9, -6.67, -4.33, -2, +0.33, +2.67, +5$ dB). Hence, each participant listened to 128 sentences throughout the experiment (32 sentences per each of the four blocks). 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.

On each trial during the experiment, a fixation cross was presented on the computer screen while concurrently a sentence was presented auditorily. After each sentence, an input box occurred on the screen and participants were asked "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.

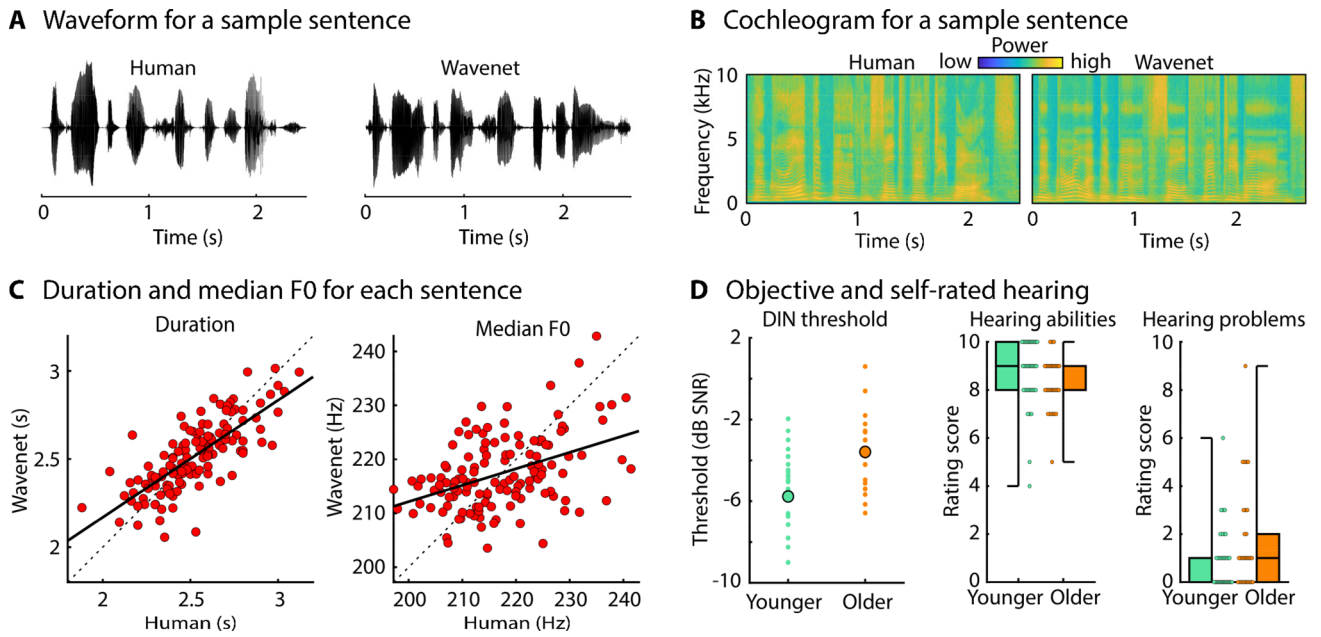

Fig. 1 Stimulus representations and hearing abilities of participants in Experiment 1. **A** Sample waveforms for the same sentence spoken by a human female speaker and synthesized using Google Wavenet. **B** Cochleograms for the same sentences as in panel A. **C** Duration and median fundamental frequency (F0) for the 128 human and 128 Wavenet sentences. Dots reflect values for individual sentences. The dashed 45-degree line indicates hypothetical data points if human and Wavenet sentences would have the same values. The black solid line reflects the true correlation between human and Wavenet sentences. **D** Objective and subjective hearing abilities. Digit-in-noise thresholds (left; big dots = mean across participants), self-rated general hearing abilities (middle), and self-rated hearing problems (right). Small dots reflect data from individual participants

To familiarize participants with the procedure, they performed a brief 12-trial training block prior to the four experimental blocks. Training trials included human and Wavenet speech masked by different masker types (unmodulated, modulated) at 0 dB and 5 dB SNR.

For the data analysis, the proportion of correctly reported words was calculated separately for each speech type, masker type, and SNR level. Different or omitted words were counted as errors, but minor misspellings and incorrect grammatical number (e.g., singular vs. plural) were not. A logistic function was fit to the proportion of correctly reported words separately for each speech type and masker type as a function of SNR level using the following equation:

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

where r is the slope, x_0 is the inflection point or the speech-reception threshold associated with 50% proportion of correctly reported words, and x refers to the SNR levels (-9, -6.67, -4.33, -2, $+0.33$, $+2.67$, $+5$ dB; the clear condition was not used for this analysis). To examine differences in masking release as a function of age, the threshold (x_0) from the logistic function fit was calculated, separately for each speech type and masker type. Thresholds were analyzed using repeated-measures analyses of variance (rmANOVAs) with Speech Type (human, Wavenet) and Masker Type (unmodulated, modulated) as within-participants factors and Age Group (younger, older) as between-participants factor. Interactions were resolved using paired and independent samples t-tests.

3.1.3 Assessment of naturalness of wavenet speech

After participants listened to the four blocks of human and Wavenet-generated speech, participants were asked "Did you notice that one of the two voices was computer generated?" and provided a binary "yes" or "no" response. Seven participants did not provide a response to this question. A chi-square test was used to compare the proportion of "yes" responses between age groups.

In an additional ~ 5 min block following the speech-in-noise intelligibility procedures, participants rated their experiences of individual sentences using the MOS-X2 scale (Mean Opinion Score Expanded Version 2; Lewis, 2018). The MOS-X2 is a four-item scale that assesses a person's experience listening to computer-generated speech (Lewis, 2018). The four items assess 'intelligibility', 'naturalness', 'prosody', and 'social impression', and are rated on an 11-point scale ranging from 0 to 10, where 10 reflects

the highest score. Participants listened to fifteen randomly selected sentences from the list of 128 sentences. Five sentences each were from the human female speaker, Wavenet-synthesized speech, and Google's Standard synthesized speech. The latter was included as an additional comparison, and duration and median F0 were adjusted to match the sentences of the human speaker (similar to the procedure that was carried out for Wavenet sentences). Google's Standard synthesized speech generally sounds more robotic than Wavenet speech. All fifteen sentences were presented under clear conditions without added background noise. After each sentence, participants rated each of the four MOS-X2 items. For the data analysis, rating scores were averaged across the four items and then across sentences, separately for each of the three speech types. A rmANOVA was calculated for the MOS-X2 scores, using Speech Type (human, Wavenet, Standard) as a within-participants factor and Age Group (younger, older) as a between-participants factor. Interactions were resolved using paired and independent samples t-tests.

3.2 Results

3.2.1 Hearing

Older adults had higher (i.e., worse) digit-in-noise perception thresholds compared to younger adults ($t_{75} = 5.708$, $p = 2.2 \times 10^{-7}$, $d = 1.306$; Fig. 1D). Previous work has provided the regression coefficients that describe the linear relation between audiometric pure-tone average thresholds (PTAs) and digit-in-noise thresholds ($\beta 1 = 4.94$, $\beta 0 = 33.84$; for slope and intercept, respectively; Smits et al., 2004). Using Smits et al.'s regression coefficients, the mean estimated PTA for younger and older adults was 5.3 and 16.1 dB HL, respectively, which is consistent with PTAs measured in the lab for younger and older adults recruited from the general public (Herrmann et al., 2022; Irsik et al., 2021; Presacco et al., 2016). Wilcoxon rank sum tests also showed that older adults self-reported having lower general hearing abilities than younger adults ($p = 0.022$), whereas no age-group difference was found for self-reported hearing problems ($p = 0.136$; Fig. 1D).

3.2.2 Speech-in-noise intelligibility task

For the speech-in-noise intelligibility task, the proportions of correctly reported words are depicted in Fig. 2A. Analysis of speech-in-noise thresholds showed lower (i.e., better) thresholds for younger compared to older adults (effect of Age Group: $F_{1,75} = 52.38$, $p = 3.3 \times 10^{-10}$, $\omega^2 = 0.253$). Speech-in-noise thresholds were also lower for Wavenet than human speech (effect of Speech Type: $F_{1,75} = 38.014$, $p = 3.2 \times 10^{-8}$, $\omega^2 = 0.037$), and this effect was larger in older compared to younger adults (Speech Type × Age Group interaction: $F_{1,75} = 5.714$, $p = 0.019$, $\omega^2 = 0.005$). Speech-in-noise thresholds were lower for modulated compared to unmodulated maskers (masking release; effect of Masker Type: $F_{1,75} = 222.066$, $p = 4 \times 10^{-24}$, $\omega^2 = 0.353$). Consistent with previous work (Dubno et al., 2002; Gustafsson & Arlinger, 1994; Irsik et al., 2022; Summers & Molis, 2004), the effect of Masker Type – that is, the release from masking – was larger for younger compared to older adults (Masker Type × Age Group interaction: $F_{1,75} = 11.733$, $p = 0.001$, $\omega^2 = 0.026$), but both age groups showed a release-from-masking benefit (younger: $t_{41} = 11.544$, $p = 1 \times 10^{-14}$, $d = 1.781$; older: $t_{34} = 10.631$, $p = 2.4 \times 10^{-12}$, $d = 1.797$; Fig. 2B). The Speech Type factor did not modulate the release-from-masking effect (Speech Type × Masker Type interaction: $F_{1,75} = 0.210$, $p = 0.648$, $\omega^2 < 0.001$) nor the release-from-masking difference between age groups (Speech Type × Masker Type × Age Group interaction: $F_{1,75} = 0.005$, $p = 0.942$, $\omega^2 < 0.001$).

Slopes of the psychometric function were steeper for Wavenet than human speech (effect of Speech Type: $F_{1,75} = 32.92$, $p = 1.9 \times 10^{-7}$, $\omega^2 = 0.132$) and for unmodulated than modulated maskers (effect of Masker Type: $F_{1,75} = 31.84$, $p = 2.8 \times 10^{-7}$, $\omega^2 = 0.15$), suggesting that, in both cases, speech intelligibility was less variable across sentences. No other effects or interactions were significant (all $p > 0.15$).

3.2.3 Subjective experience of naturalness of speech materials

After the speech-in-noise intelligibility task, participants indicated whether they noticed that one of the two voices was computer generated. 83% of younger adults, but only 26% of older adults responded 'yes' ($\chi^2 = 23.03$, $p = 1.6 \times 10^{-6}$), suggesting that few older adults noticed that computer-generated speech was presented (see Fig. 3).

Participants rated their experience listening to sentences of different Speech Types using the MOS-X2 scale. MOS-X2 scores were lower for Wavenet ($t_{76} = 11.586$, $p = 1.8 \times 10^{-18}$, $d = 1.32$) and Standard speech ($t_{76} = 14.531$, $p = 1.2 \times 10^{-23}$, $d = 1.656$) compared to human speech, and lower for Standard compared to Wavenet speech ($t_{76} = 7.918$, $p = 1.6 \times 10^{-11}$, $d = 0.902$; effect of Speech Type: $F_{2,150} = 181.487$, $p = 8.9 \times 10^{-41}$, $\omega^2 = 0.459$). MOS-X2 scores were larger for older compared to younger adults (effect of Age Group: $F_{1,75} = 8.747$, $p = 0.004$, $\omega^2 = 0.048$), but only for Wavenet ($t_{75} = 2.746$, $p = 0.008$, $d = 0.629$) and Standard speech ($t_{75} = 4.012$, $p = 1.4 \times 10^{-4}$, $d = 0.918$), whereas MOS-X2 scores for human speech were slightly lower for older compared to younger adults ($t_{75} = -2.114$, $p = 0.038$, $d = 0.484$; Speech Type × Age Group interaction: $F_{2,150} = 17.679$, $p = 1.3 \times 10^{-7}$, $\omega^2 = 0.073$; Fig. 3).

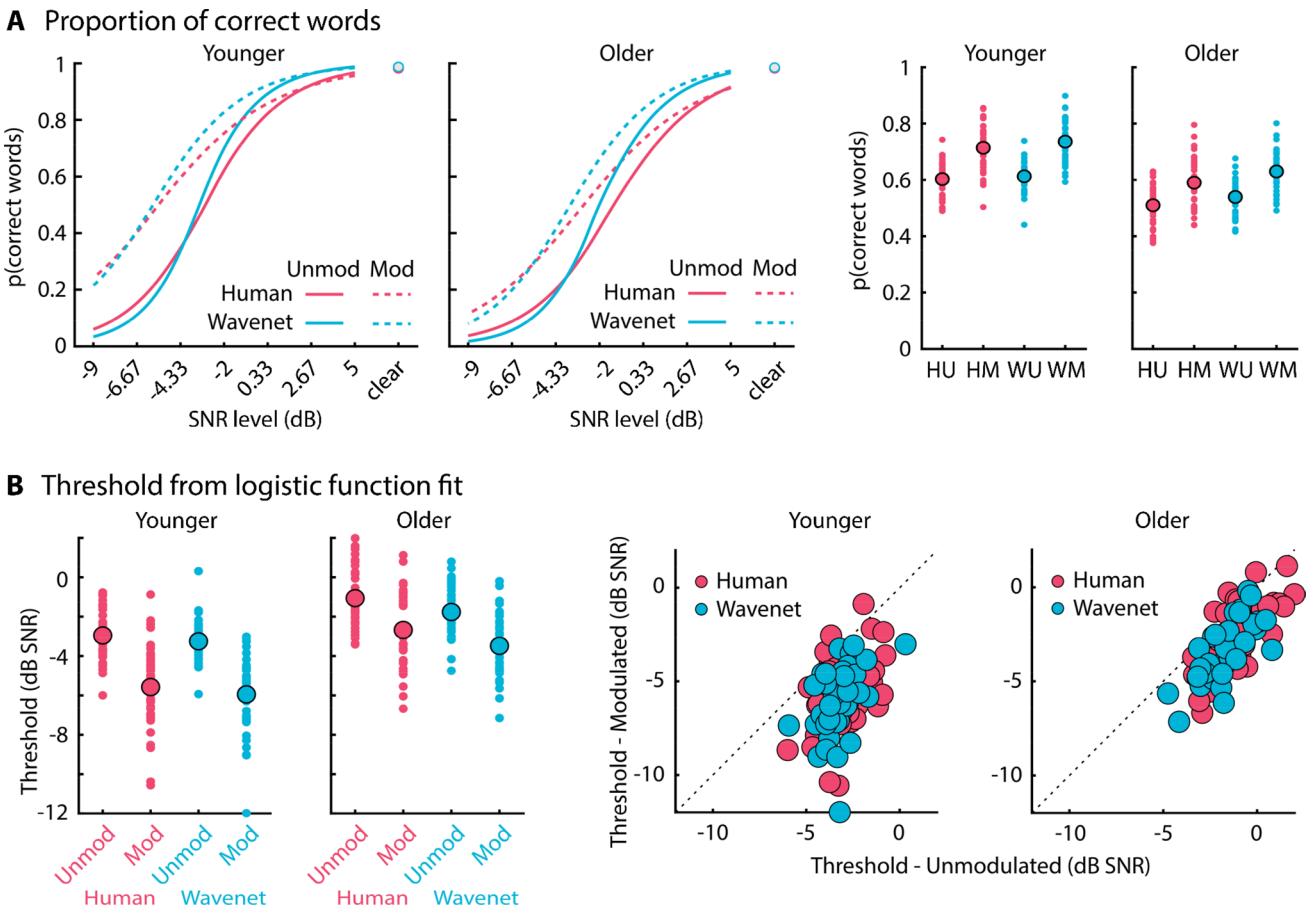

Fig. 2 Intelligibility results and speech-in-noise thresholds for Experiment 1. **A** Predicted proportion correct word reports for different signal-to-noise ratios (SNRs) from logistic function fits (left) and mean proportion correct word reports (right). **B** Speech-in-noise thresholds estimated from logistic function fits. The dashed line on the right-hand plots indicates the threshold if there were no difference between thresholds for modulated and unmodulated maskers. Unmod – unmodulated masker, Mod—modulated masker

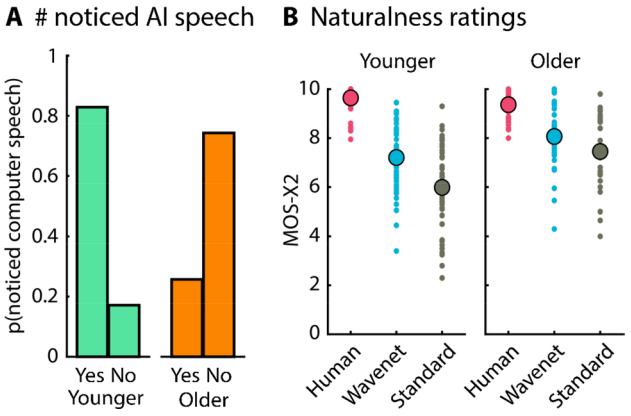

Fig. 3 Recognition and naturalness ratings of synthesized AI speech. **A** Proportion of participants indicating after the speech-in-noise task whether they noticed that one of the two voices was computer generated. **B** Naturalness ratings (MOS-X2) of sentences uttered by a human vs synthesized by a computer. Synthesized voices were Google's Wavenet and Standard voices

3.3 Summary

Experiment 1 shows for both human-spoken and AI-based synthesized speech that speech intelligibility for younger and older adults is better for modulated compared to unmodulated maskers (release from masking; Dubno et al., 2002; Irsik et al., 2022; Miller & Licklider, 1950) and that the intelligibility benefit for modulated maskers is reduced for older compared to younger adults (Bacon et al., 1998; George et al., 2006; Gustafsson & Arlinger, 1994; Irsik et al., 2022). The release from masking effect and the reduced masking release for older adults did not statistically differ between human and AI speech. These results suggest that AI speech could be used to investigate speech-in-noise perception phenomena. Interestingly, only 26% of older adults recognized that computer-generated (synthesized) speech was used, as opposed to the 83% of younger adults, and older adults rated synthesized speech as more natural and human-like compared to younger adults. In Experiment 1, a female voice was used to investigate speech-in-noise intelligibility

and experiences with AI speech. Many voice-AI assistive systems use female voices by default, and it may thus be that individuals have less experience with male compared to female AI speech. Hence, younger and older adults might notice less well that computer-generated speech is used for a male compared to a female voice. Experiment 2 was designed to generalize the release-from-masking effects and reduced masking release for older adults to male AI speech.

4 Experiment 2

Experiment 2 mirrored procedures of Experiment 1, but speech recordings from a male human speaker and synthesized speech for a male voice were used in Experiment 2.

4.1 Methods

4.1.1 Participants

Thirty-five younger (mean age: 32.4 years, age range: 24–38 years) and 32 older adults (mean age: 64.9 years, age range: 59–74 years) participated in Experiment 2. Out of the 35 younger adults, 21 identified as male, 13 as female, and 1 as nonbinary. Out of the 32 older adults, 13 identified as male and 19 as female.

Data from additional participants were excluded if they met one or more of the following exclusion criteria: using a hearing aid (N = 1), having a history of a neurological disease (N = 3), failing the three-interval forced choice headphone-check (N = 22), reported that they were distracted during the experiment (N = 5), or reported being a non-native English speaker (N = 0). Participants who passed these criteria but showed an average speech-in-noise intelligibility performance that was lower than one standard deviation from the group mean (separately for younger and older adults) were also excluded (N = 12). One additional participant's data were also excluded because the person showed a very high digit-in-noise threshold (9.5 dB SNR), despite good performance in the speech-in-noise intelligibility task. Similar to Experiment 1, exclusion of participants with low overall performance removed people with outlier speech-in-noise thresholds, but none of the statistical results reported below were qualitatively affected by inclusion or exclusion of these datasets. Overall, 41 datasets were excluded.

4.1.2 Materials and procedures for speech-in-noise intelligibility task

Procedures for the speech-in-noise intelligibility task were the same as in Experiment 1, with the exception that male speech was used and the signal-to-noise ratios at which speech was masked were slightly altered to capture a wide range of intelligibility performance levels.

Speech recordings of the Harvard 128 sentences made by a human male speaker were taken from a publically available resource (https://depts.washington.edu/phonlab/projects/uwnu.php; McCloy et al., 2018; Panfili et al., 2017). Speech recordings of the 'PNM055' speaker were used in Experiment 2. The Google Wavenet male voice was "en-US-Wavenet-D" (https://cloud.google.com/text-to-speech/docs/voices). As for Experiment 1, the parameters 'speaking_rate' and 'pitch' were used to manipulate the duration and median F0 of the 128 Wavenet sentences, such that average duration and median F0 across the sets of 128 sentences did not differ statistically between human speech and Wavenet speech ($p > 0.7$; Fig. 1C). The variance of duration across sentences was smaller for Wavenet than human speech (Levene test: $p = 0.009$; there was no difference for the variance of median F0: $p = 0.509$). Duration and median F0 of human and Wavenet sentences correlated (duration: $r = 0.738$, $p = 2.6 \times 10^{-23}$; median F0: $r = 0.372$, $p = 1.6 \times 10^{-5}$). Sample waveforms and corresponding cochleograms (McDermott & Simoncelli, 2011) are displayed in Fig. 4A and B. Sentences were presented either under clear conditions or twelve-talker babble (Bilger, 1984) was added to the sentences using seven SNR levels: $-11, -8.33, -5.67, -3, -0.33, 2.33, +5$ dB (levels differ slightly from levels used in Experiment 1, because pilot tested showed that intelligibility of the male voices was better relative to the female voices used in Experiment 1). Babble noise was either unmodulated or sinusoidally amplitude modulated at a rate of 4 Hz.

4.1.3 Assessment of naturalness of Wavenet speech

Procedures to assess the experience with Wavenet speech were similar to those in Experiment 1.

In Experiment 2, participants were additionally asked about their general experience with voice-AI systems. Experience was assessed using the following question: "Do you have experience using a voice-AI (artificial intelligence) system, such as Amazon Alexa, Google Assistant, or Apple Siri?". Participants provided a self-assessment on a 11-point rating scale ranging from 0 to 10, where the end points referred to "no experience" and "high experience", respectively. Age group differences in self-reported voice-AI experience was analyzed using Wilcoxon's rank sum test.

4.2 Results

4.2.1 Hearing

Older adults showed higher (i.e., worse) digit-in-noise thresholds compare to younger adults ($t_{65} = 3.88$, $p = 2.5 \times 10^{-4}$, $d = 0.949$). The digit-in-noise thresholds

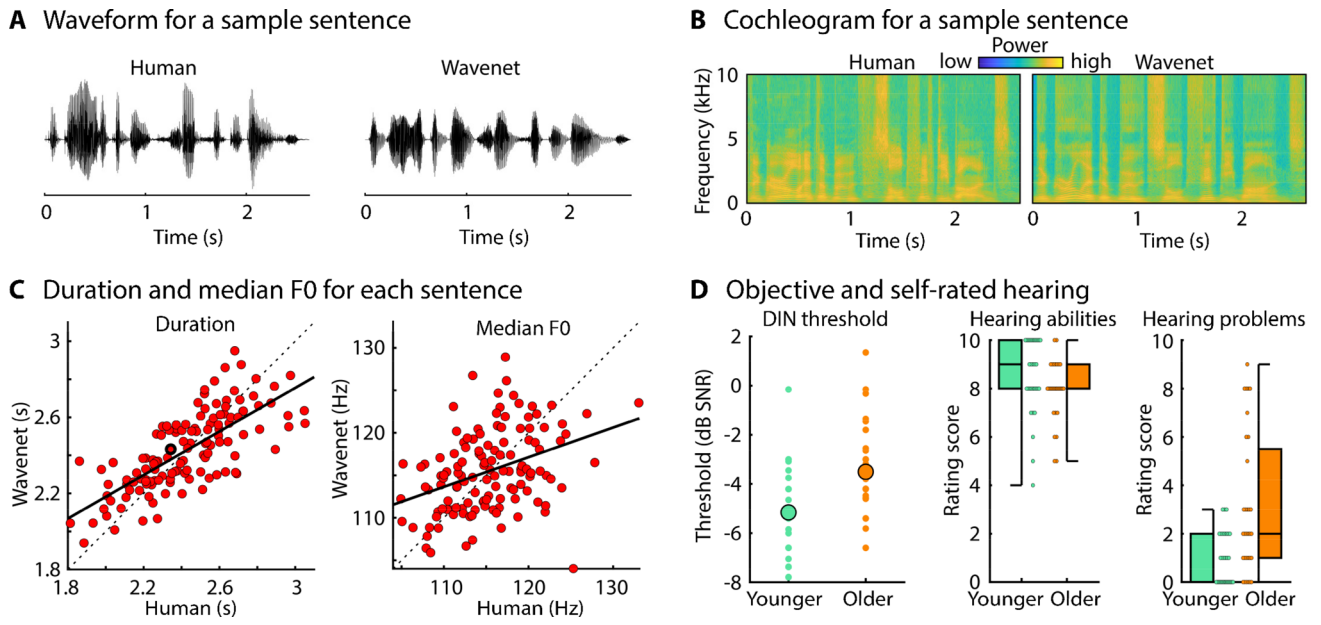

Fig. 4 Stimulus representations and hearing abilities of participants in Experiment 2. **A** Sample waveforms for the same sentence spoken by a human female speaker and synthesized using Google Wavenet. **B** Cochleograms for the same sentences as in panel A. **C** Duration and median fundamental frequency (F0) for the 128 human and 128 Wavenet sentences. Dots reflect values for individual sentences. The dashed 45-degree line indicates hypothetical data points if human and Wavenet sentences would have the same values. The black solid line reflects the true correlation between human and Wavenet sentences. **D** Objective and subjective hearing abilities. Digit-in-noise threshold (left; big dots = mean across participants), self-rated general hearing abilities (middle), and self-rated hearing problems (right). Small dots reflect data from individual participants

correspond to an approximate average PTA of 8.4 dB HL for younger adults and a PTA of 16.5 dB HL for older adults (Smits et al., 2004). Older adults also self-rated their general hearing abilities to be lower ($p = 0.021$) and their hearing problems to be greater ($p = 4.6 \times 10^{-4}$) compared to younger adults (Fig. 4D).

4.2.2 Speech-in-noise intelligibility task

Proportions of correctly reported words are depicted in Fig. 5A. Speech-in-noise perception thresholds were lower (i.e., better) for younger compared to older adults (effect of Age Group: $F_{1,65} = 39.314$, $p = 3.3 \times 10^{-8}$, $\omega^2 = 0.225$). Thresholds were also lower for human than Wavenet speech (effect of Speech Type: $F_{1,65} = 93.393$, $p = 3.4 \times 10^{-14}$, $\omega^2 = 0.148$) and lower for modulated than unmodulated maskers (release from masking; effect of Masker Type: $F_{1,65} = 297.897$, $p = 5.8 \times 10^{-26}$, $\omega^2 = 0.396$). Release from masking was significant for both human ($t_{66} = 11.668$, $p = 1.1 \times 10^{-17}$, $d = 1.425$) and Wavenet speech ($t_{66} = 12.546$, $p = 3.9 \times 10^{-19}$, $d = 1.533$), but tended to be larger for human compared to Wavenet speech (Speech Type × Masker Type interaction: $F_{1,65} = 3.328$, $p = 0.073$, $\omega^2 = 0.004$). Intelligibility benefit from masking release was smaller for older compared to younger adults (Masker Type × Age Group interaction: $F_{1,65} = 27.334$, $p = 1.9 \times 10^{-6}$, $\omega^2 = 0.055$), although both younger ($t_{34} = 15.792$, $p = 3.2 \times 10^{-17}$, $d = 2.669$) and older adults ($t_{31} = 8.619$, $p = 9.835 \times 10^{-10}$, $d = 1.524$) showed lower thresholds for modulated compared to unmodulated maskers (masking release). Speech Type did not significantly modulate the difference in masking release between younger and older adults (Speech Type × Masker Type × Age Group interaction: $F_{1,65} = 0.032$, $p = 0.859$, $\omega^2 < 0.001$).

Slopes of the psychometric function were steeper for Wavenet than human speech (effect of Speech Type: $F_{1,65} = 22.203$, $p = 1.3 \times 10^{-5}$, $\omega^2 = 0.133$) and for unmodulated than modulated maskers (effect of Masker Type: $F_{1,65} = 16.165$, $p = 1.5 \times 10^{-4}$, $\omega^2 = 0.112$), suggesting that, in both cases, speech intelligibility was less variable. No other effects or interactions were significant (all $p > 0.15$).

4.2.3 Subjective experience of naturalness of speech materials

After the speech-in-noise intelligibility task, 71% of younger adults but only 13% of older adults responded 'yes' to the question whether they noticed that one of the two voices during the speech-intelligibility task was computer-generated ($\chi^2 = 23.646$, $p = 1.2 \times 10^{-6}$; Fig. 6A). The percentage of people noticing the computer-generated speech appeared about 12% lower for the male voice used in Experiment 2

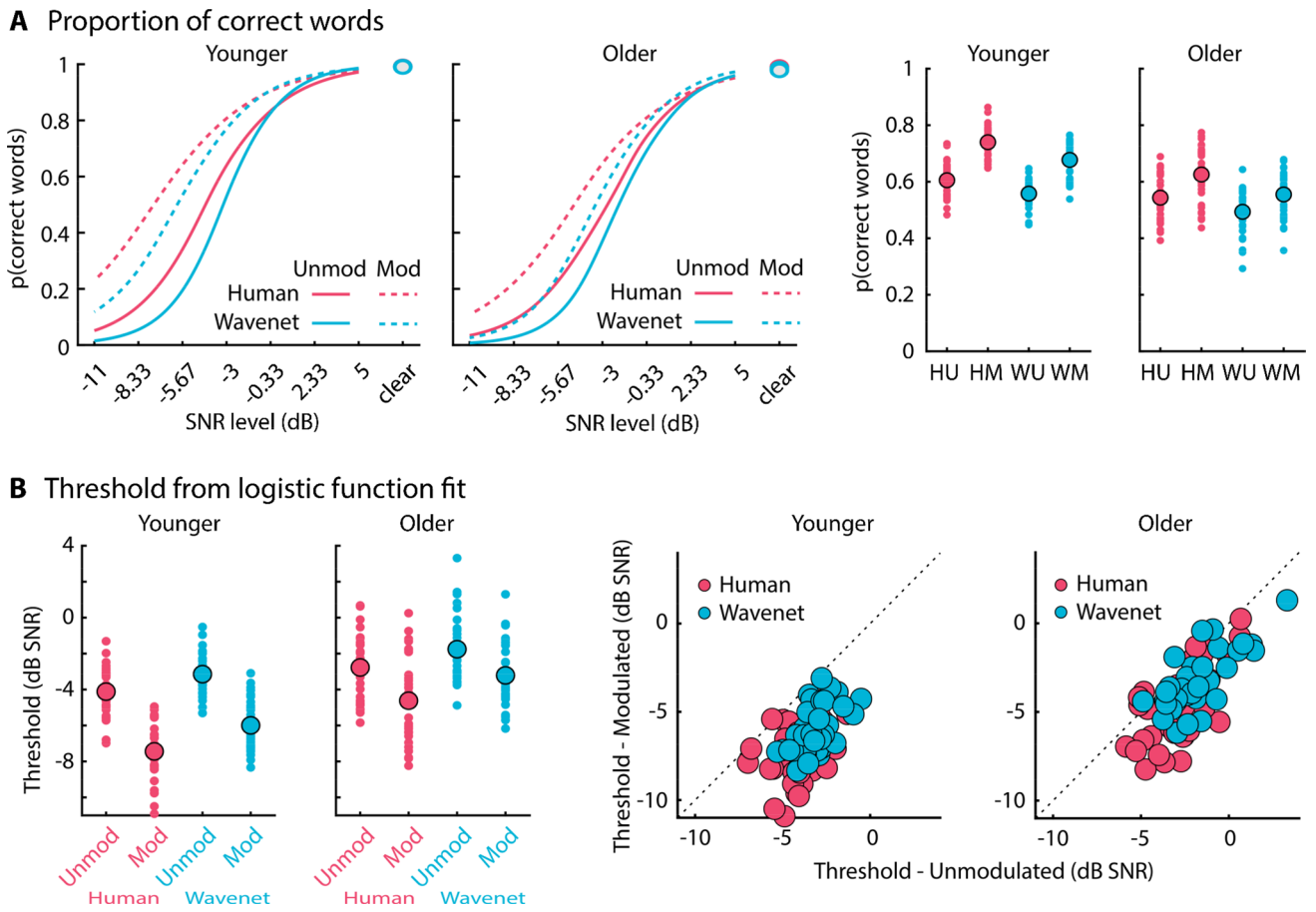

Fig. 5 Intelligibility results and speech-in-noise thresholds for Experiment 2. **A** Predicted proportion correct word reports for different signal-to-noise ratios (SNRs) from logistic function fits (left) and mean proportion correct word reports across all levels (right). **B** Speech-in-noise thresholds estimated from logistic function fits. The dashed line on the right-hand plots indicates the threshold if there were no difference between thresholds for modulated and unmodulated maskers. Unmod – unmodulated masker, Mod –modulated masker

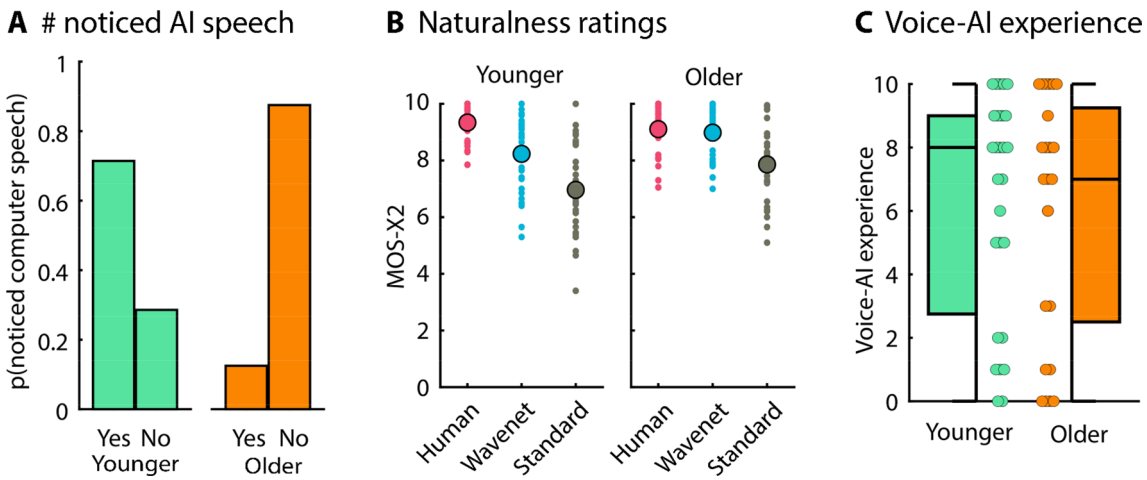

Fig. 6 Recognition and naturalness ratings of synthesized speech. **A** Proportion of participants indicating after the speech-in-noise task whether they noticed that one of the two voices was computer-generated. **B** Naturalness ratings of sentences uttered by a human vs synthesized (computer-generated). Synthesized speech was generated using Google's Wavenet and Standard voices. **C** Self-rated experience with voice-AI systems

compared to the female voice using in Experiment 1, but this was not significant ($\chi^2 = 1.658$, $p = 0.198$).

Analysis of the MOS-X2 scores showed, similar to Experiment 1, that human speech was rated higher than Google's Wavenet ($t_{66} = 4.655$, $p = 1.6 \times 10^{-5}$, $d = 0.569$) and Standard speech ($t_{66} = 10.57$, $p = 7.8 \cdot 10^{-16}$, $d = 1.291$), and that Wavenet speech was rated higher than Standard speech ($t_{66} = 9.803$, $p = 1.7 \times 10^{-14}$, $d = 1.198$; effect of Speech Type: $F_{2,130} = 90.451$, $p = 2.4 \times 10^{-25}$, $\omega^2 = 0.306$). MOS-X2 scores were larger for older compared to younger adults for Wavenet ($t_{65} = 2.721$, $p = 0.008$, $d = 0.666$) and Standard speech ($t_{65} = 2.599$, $p = 0.012$, $d = 0.636$), but not for human speech ($t_{65} = -1.229$, $p = 0.224$, $d = 0.301$; Speech Type × Age Group interaction: $F_{2,130} = 10.028$, $p = 8.9 \cdot 10^{-5}$, $\omega^2 = 0.043$; effect of Age Group: $F_{1,65} = 4.275$, $p = 0.043$, $\omega^2 = 0.024$; Fig. 6B).

Naturalness ratings appeared somewhat higher for the male voice used in Experiment 2 (Fig. 6B) compared to the female voice used in Experiment 1 (Fig. 3B). To examine this difference directly, the between-participants factor Voice Gender (female, male) was added to the rmANOVA analyzing MOS-X2 scores. As before, the effects of Speech Type ($F_{2,280} = 253.164$, $p = 1.6 \times 10^{-63}$, $\omega^2 = 0.381$) and Age Group ($F_{1,140} = 12.518$, $p = 5.5 \times 10^{-4}$, $\omega^2 = 0.039$), and the Speech Type × Age Group interaction were significant ($F_{2,280} = 26.537$, $p = 2.8 \times 10^{-11}$, $\omega^2 = 0.059$), mirroring the differences described above. Critically, MOS-X2 scores for male Wavenet and Standard sentences were greater than MOS-X2 scores for the female voice (Wavenet: $t_{142} = 4.525$, $p = 1.3 \times 10^{-5}$, $d = 0.756$; Standard: $t_{142} = 2.717$, $p = 0.007$, $d = 0.454$), whereas MOS-X2 scores for sentences spoken by the male human speaker were lower compared to the female speaker ($t_{142} = -2.532$, $p = 0.012$, $d = 0.423$; Speech Type × Voice Gender interaction: $F_{2,280} = 20.686$, $p = 2.2 \times 10^{-9}$, $\omega^2 = 0.046$; effect of Voice Gender: $F_{1,140} = 7.987$, $p = 0.005$, $\omega^2 = 0.024$). These effects did not differ between younger and older adults (Speech Type × Voice Gender × Age Group interaction: $F_{2,280} = 1.232$, $p = 0.293$, $\omega^2 < 0.001$).

Self-rated experience with voice-AI systems, such as Apple's Siri or Amazon's Alexa, did not significantly differ between age groups ($p = 0.774$; Fig. 6C).

4.3 Summary

Results of Experiment 2 mirrored those of Experiment 1 in that, for both human and Wavenet speech, intelligibility was reduced for the modulated compared to the unmodulated masker (masking release) and this masking-release effect was smaller for older compared to younger adults. A female voice was used in Experiment 1 and a male voice was used in Experiment 2. Hence, the data demonstrate that common speech-in-noise perception effects can be observed using state-of-the-art AI-based synthesized speech and that the results generalize across different AI voices and genders.

Experiment 2, similar to Experiment 1, showed that older adults recognized synthesized speech less frequently (13%) compared to younger adults (71%), and that older adults rated synthesized speech as being more natural and human-like than younger adults. However, these results are based on subjective reports and ratings. Whether reduced recognition of AI speech is the result of a response bias or whether older adults are indeed unable to discriminate between human and AI speech cannot be answered based on the data from Experiments 1 and 2. Experiment 3 was designed to answer this question using sentences spoken by 10 different human speakers and synthesized for 10 different voices.

5 Experiment 3

Experiment 3 aimed to investigate the perceptual sensitivity of younger and older adults to discriminate between human-spoken and AI-based synthesized speech.

5.1 Methods

5.1.1 Participants

Thirty-three younger adults (mean age: 32.2 years, age range: 24–38 years) and 34 older adults (mean age: 65.8 years, age range: 59–78 years) participated in Experiment 3. Of the younger adults, 23 identified as male, 9 as female, and one as genderqueer. Of the older adults, 16 identified as male or man, 17 as female or woman, and one person did not provide a meaningful response regarding their gender.

Data from additional participants were excluded if they met one or more of the following exclusion criteria: using a hearing aid (N = 3), having a history of a neurological disease (N = 5), failing the three-interval forced choice headphone-check (N = 20), or reported being a non-native English speaker (N = 1). In Experiment 3, all sentences were presented with minimal background masking (i.e., at + 10 dB SNR) or under clear conditions, and were thus highly intelligible. Hence, participants who passed the exclusion criteria mentioned above but showed an average speech-in-noise intelligibility performance lower than 85% were also excluded, assuming they were unable or unwilling to perform the task (additional N = 9). Data from one additional participant were also excluded because the person used the same response category on every trial for their rating of whether a sentence was AI speech or human speech. Overall, data from 37 participants were excluded.

5.1.2 Materials and procedures for speech-in-noise intelligibility and speech-type discrimination tasks

Eighty sentences from the Harvard sentence lists 1–15 were selected for Experiment 3. Sentence recordings from 5 male and 5 female human speakers were taken from a publically available resource (https://depts.washington.edu/phonlab/projects/uwnu.php; McCloy et al., 2018; Panfili et al., 2017). Speaker IDs for female voices were PNF133, PNF136, PNF140, PNF142, and PNF144, and speaker IDs for male voices were PNM055, PNM077, PNM078, PNM079, and PNM083 (McCloy et al., 2018; Panfili et al., 2017). As for Experiments 1 and 2, each of the 80 sentences per speaker were resampled from a sampling frequency of 44.1 kHz to 24 kHz. Praat software (Boersma, 2001) was used to calculate each sentence's duration and median F0 (Fig. 7).

For Wavenet voices, five female and five male voices with an US accent ("en-US-Wavenet") were used. IDs for female voices were C, E, F, G, and H. IDs for male voices were A, B, D, I, and J (https://cloud.google.com/text-to-speech/docs/voices). As for Experiments 1 and 2, the parameters 'speaking_rate' and 'pitch' were used to manipulate the duration and median F0, such that for each Wavenet voice the average duration and median F0 across the sets of 80 sentences did not differ statistically between human and Wavenet speech (for all; $p > 0.2$; with one exception: The median F0 for Wavenet male voice D was larger than the median F0 for PNM078; $p < 0.01$; the fundamental frequency of PNM078's voice was very low). Hence, to reduce potential differences between human and Wavenet speech, each Wavenet voice was paired with one human voice of the same gender to adjust Wavenet parameters (Fig. 7).

Each of the 80 sentences was randomly assigned to one of the 20 voices (10 human, 10 Wavenet) at the beginning of the experimental session. Throughout the experiment, each participant thus listened to four sentences for each of the 20 voices. Sentences were presented in four blocks and each block contained one sentence for each of the 20 voices. On each trial, a fixation cross was presented on the computer screen while a sentence was presented auditorily. After each sentence, an input box occurred on the screen and participants were asked "Please type the words exactly as you heard them (even if you only understood parts of what was said)". The speech-intelligibility task was used to determine whether participants listened to the sentences. After they typed in their response, participants were asked "Was the voice of the sentence a human or a computer-generated voice?". Participants selected one out of six response

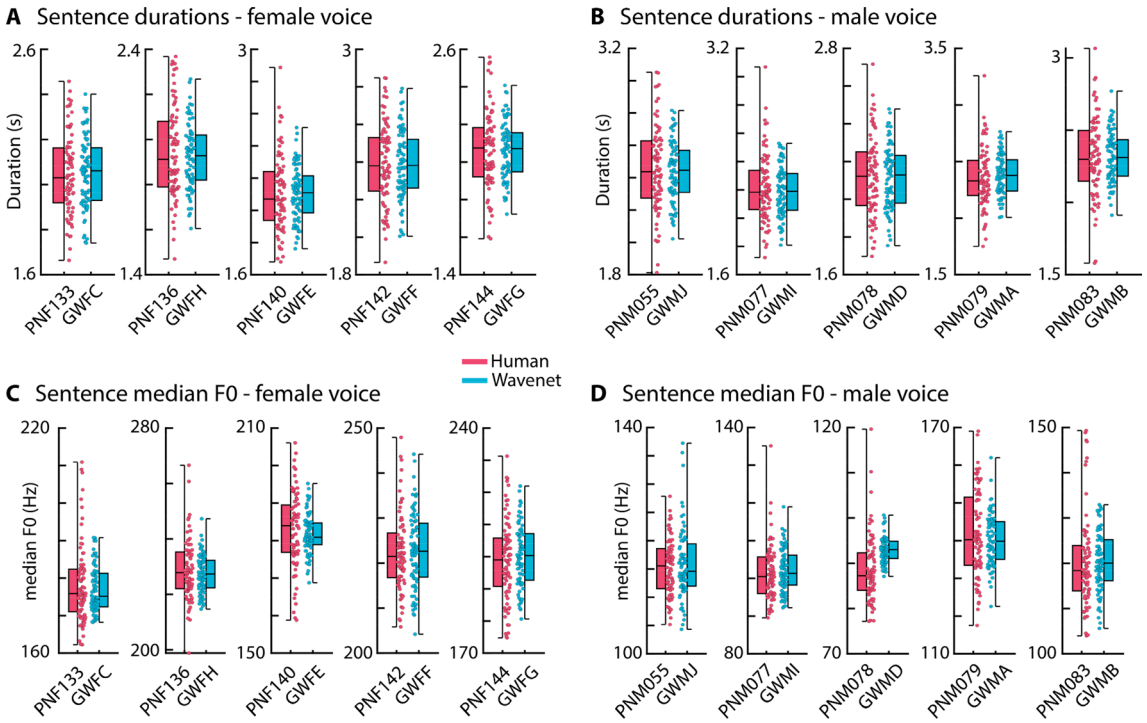

Fig. 7 Duration and median F0 for human-spoken and Wavenet-synthesized sentences used in Experiment 3. **A** Sentence durations for female voices. **B** Sentence durations for male voices. **C** Sentence median F0s for female voices. **D** Sentence median F0s for male voices. For Human voices, the abbreviations refer to the labels of original stimulus recordings (McCloy et al., 2018; Panfili et al., 2017). For Wavenet voices, the first two letters 'GW' of the voice abbreviation refers to Google Wavenet, the third letter 'F' vs 'M' refers to female and male, respectively, and the fourth letter refers to the specific Wavenet voice. Boxplots are shown along with data points for each individual sentence (dots)

categories presented on the computer screen: “Human voice (very confident)", "Human voice (somewhat confident)", "Human voice (not confident)", "Computer voice (not confident)", "Computer voice (somewhat confident)", and "Computer voice (very confident)". A 0.4 s blank screen was presented before the next trial started.

For 17 younger and 18 older adults, all sentences were presented in 12-talker background babble at + 10 dB SNR. This minimal masking was chosen to mask potential subtle differences in the quality of the recordings between human and Wavenet speech, while enabling high speech intelligibility for younger and older adults (Holder et al., 2018). To avoid that any age differences emerge because older adults find listening to speech in + 10 dB SNR background noise more effortful than younger adults despite high speech intelligibility, sentences were presented under clear conditions for additional 16 younger and 16 older adults. Critically, the speech-type discrimination effects reported below were present for both subgroups individually (speech conditions: clear vs + 10 dB SNR). To avoid presenting similar results twice, data sets were pooled and jointly analyzed. A between-participants factor was used in the analysis to explicitly account for differences between speech conditions (clear, + 10 dB SNR).

5.1.3 Data analysis

For the analysis of speech-intelligibility data, the proportion of correctly reported words was calculated for each sentence. Different or omitted words were counted as errors, but minor misspellings and incorrect grammatical number (e.g., singular vs. plural) were not. Data from participants with an overall proportion of correctly reported words lower than 0.85 were excluded from further analysis. An rmANOVA was calculated using the proportion of correctly reported words as dependent measure. The rmANOVA included the within-participants factor Speech Type (human, Wavenet) and the between-participant factors Age Group (younger, older) and Speech Condition (clear, + 10 dB SNR).

The analysis of the performance in the speech-type discrimination task focused on correctly reported sentences in order to ensure that only sentences to which participants had listened attentively are included in the analysis. To analyze the proportion of “human voice” responses, the 6-point response and confidence scale was converted such that “Human voice (very confident)” was coded as 1, “Computer voice (very confident)” as 0, and the other four responses categories equally spaced in-between. An rmANOVA was calculated for the proportion of ‘human voice’ responses using the within-participants factor Speech Type (human, Wavenet) and the between-participant factors Age Group (younger, older) and Speech Condition (clear, + 10 dB SNR).

Perceptual sensitivity (d-prime) and response bias (c) were also calculated for each participant (Macmillan & Creelman, 2004). D-prime values were calculated using custom MATLAB scripts. A response was considered a hit when the sentence was spoken by a human and the participant responded with “Human voice”. A response was considered a false alarm when the sentence was Wavenet-synthesized and the participant responded with “Human voice”. D-prime cannot be calculated when hit rates or false alarm rates are zero or one (Macmillan & Creelman, 2004). To handle these cases, the following corrections were calculated. When the hit rate was zero, it was set to $\frac{1}{2\times n}$ where n refers to the number of sentences spoken by a human. When the hit rate was one, it was set to $1 - \frac{1}{2\times n}$. When the false alarm rate was zero, it was set to $\frac{1}{2\times m}$ where m refers to the number of Wavenet sentences (Macmillan & Creelman, 2004). Data from participants with a false alarm rate of one were removed (N = 1; selecting the same response category on all trials). Age-group differences were assessed using an independent-samples t-test.

The standard voice gender of many assistive voice-AI and smart systems is female and participants may thus have more experience and familiarity with female compared to male AI voices. In order to explore whether the ability to discriminate between human and Wavenet speech depends on the gender of the voice, perceptual sensitivity and bias were also calculated separately for different voice genders. D-prime and bias c were separately analyzed in a rmANOVA, including the within-participants factor Voice Gender (male, female), and the between-participants factors Age Group (younger, older) and Speech Condition (clear, + 10 dB SNR).

5.2 Results

5.2.1 Hearing

Older adults showed higher (i.e., worse) digit-in-noise thresholds compared to younger adults ($t_{65} = 5.069$, $p = 3.5 \times 10^{-6}$, $d = 1.239$). The digit-in-noise thresholds correspond to an approximate average PTA of 6.8 dB HL for younger adults and a PTA of 19 dB HL for older adults (Smits et al., 2004). Self-rated general hearing abilities and hearing problems did not differ between age groups ($p > 0.3$; Fig. 8A). Self-reported experience with voice-AI systems did also not differ between younger and older adults ($p = 0.263$; Fig. 8B).

5.2.2 Speech-in-noise intelligibility

The proportion of correct word reports was greater for Wavenet compared to human speech ($F_{1,63} = 12.019$, $p = 9.5 \times 10^{-4}$, $\omega^2 = 0.015$; mean proportion: 0.966 and 0.958, respectively) and greater for clear speech compared

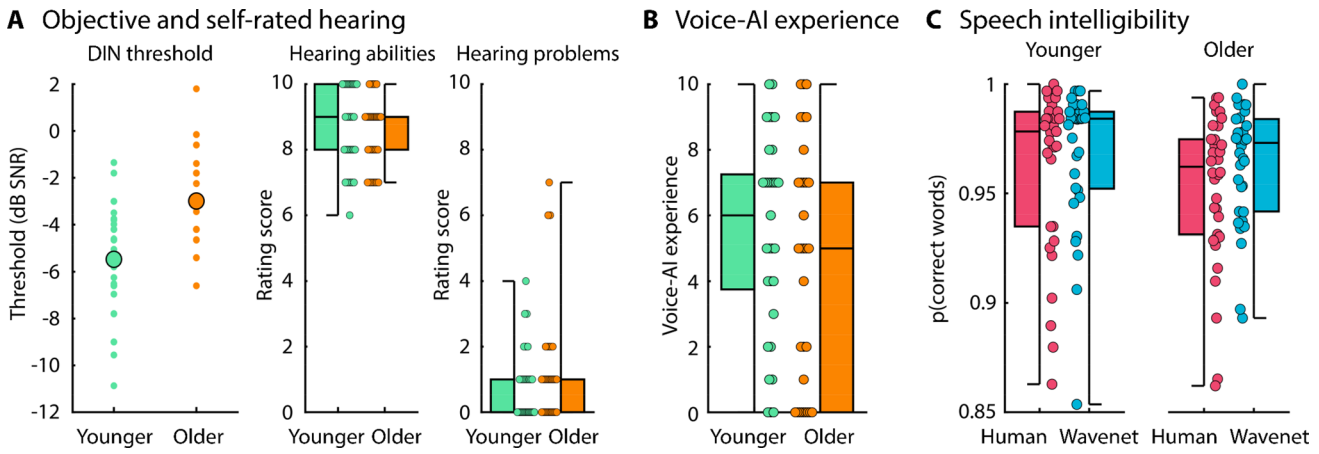

Fig. 8 Hearing assessments, voice-AI experience, and speech intelligibility in Experiment 3. **A** Objective and subjective hearing abilities. Digit-in-noise thresholds (left; large dots = mean across participants), self-rated general hearing abilities (middle), and self-rated hearing problems (right). **B** Self-rated experience with voice-AI systems. **C** Proportion correct word reports in the speech-intelligibility task. Small dots reflect data from individual participants

to speech masked by + 10 dB SNR ($F_{1,63} = 6.274$, $p = 0.015$, $\omega^2 = 0.04$; mean proportion: 0.972 and 0.953, respectively). No other effects or interactions were significant (for all $p > 0.1$; Fig. 8C).

5.2.3 Speech-type discrimination performance

The analysis of the proportion of "human voice" responses showed that participants responded "human voice" more frequently when the voice was a human voice compared to a Wavenet voice (effect of Speech Type: $F_{1,63} = 289.889$, $p = 3 \times 10^{-25}$, $\omega^2 = 0.697$). Older adults categorized speech overall more frequently as originating from a "human voice" than younger adults (effect of Age Group: $F_{1,63} = 5.078$, $p = 0.028$, $\omega^2 = 0.031$). The Speech Type × Age Group interaction was also significant ($F_{1,63} = 24.532$, $p = 5.8 \times 10^{-6}$, $\omega^2 = 0.158$). Older adults categorized speech spoken by a human less frequently as "human voice" ($t_{65} = -2.189$, $p = 0.032$, $d = 0.535$) and Wavenet speech more frequently as "human voice" ($t_{65} = 4.812$, $p = 9.3 \times 10^{-6}$, $d = 1.176$) compared to younger adults (Fig. 9A). None of the other effects and interactions were significant (for all $p > 0.4$).

Perceptual sensitivity (d-prime) and bias (c) were also calculated to distinguish between a person's ability to discriminate human from AI speech and a person's bias towards selecting one response category over another. Younger compared to older adults were better able to discriminate between human and AI speech, as indicated by larger d-prime values ($F_{1,63} = 27.075$, $p = 2.3 \times 10^{-6}$, $\omega^2 = 0.284$; Fig. 9B left; no effect of Speech Condition or Age Group × Speech Condition interaction, $p > 0.4$). There were no effects related to bias ($p > 0.1$), but both age groups were biased towards responding "human voice" (t-test against zero; younger: $t_{32} = 5.351$, $p = 7.2 \times 10^{-6}$, $d = 0.932$; older: $t_{33} = 6.084$, $p = 7.5 \times 10^{-7}$, $d = 1.043$; Fig. 9B left). A regression aiming to predict d-prime from digit-in-noise perception thresholds (additionally including predictors Speech Condition and Age Group) did not reveal a significant relationship ($t_{63} = -1.093$, $p = 0.279$).

Analyses thus far have focused on an aggregate of responses across different voices and voice genders. The right-hand side of Fig. 9A shows the proportion of "human voice" responses for each human and AI voice. Response patterns appear largely consistent across voices. However, the data suggest potential differences associated with voice gender. Hence, perceptual sensitivity and bias were analyzed in separate rmANOVAs with the additional within-participants factor Voice Gender. D-prime was larger for older compared to younger adults (effect of Age Group: $F_{1,63} = 30.962$, $p = 5.8 \times 10^{-7}$, $\omega^2 = 0.190$) and for female compared to male voices (effect of Voice Gender: $F_{1,63} = 72.403$, $p = 4.6 \times 10^{-12}$, $\omega^2 = 0.177$). The Age Group × Voice Gender interaction was also significant ($F_{1,63} = 4.044$, $p = 0.049$, $\omega^2 = 0.009$). D-prime was larger for younger adults compared to older adults for both female ($t_{65} = 5.794$, $p = 2.2 \times 10^{-7}$, $d = 1.416$) and male voices ($t_{65} = 4.353$, $p = 4.9 \cdot 10^{-5}$, $d = 1.064$), but the difference was larger for female voices. There was no effect or interaction involving Speech Condition ($p > 0.6$).

Analysis of bias showed that individuals were more biased towards reporting "human voice" for male than female voices (effect of Voice Gender: $F_{1,63} = 18.752$, $p = 5.4 \times 10^{-5}$, $\omega^2 = 0.068$) and older adults seemed somewhat more biased (effect of Age Group: $F_{1,63} = 3.070$, $p = 0.085$, $\omega^2 = 0.016$). These effects were specified by the Age Group × Voice Gender interaction ($F_{1,63} = 19.356$,

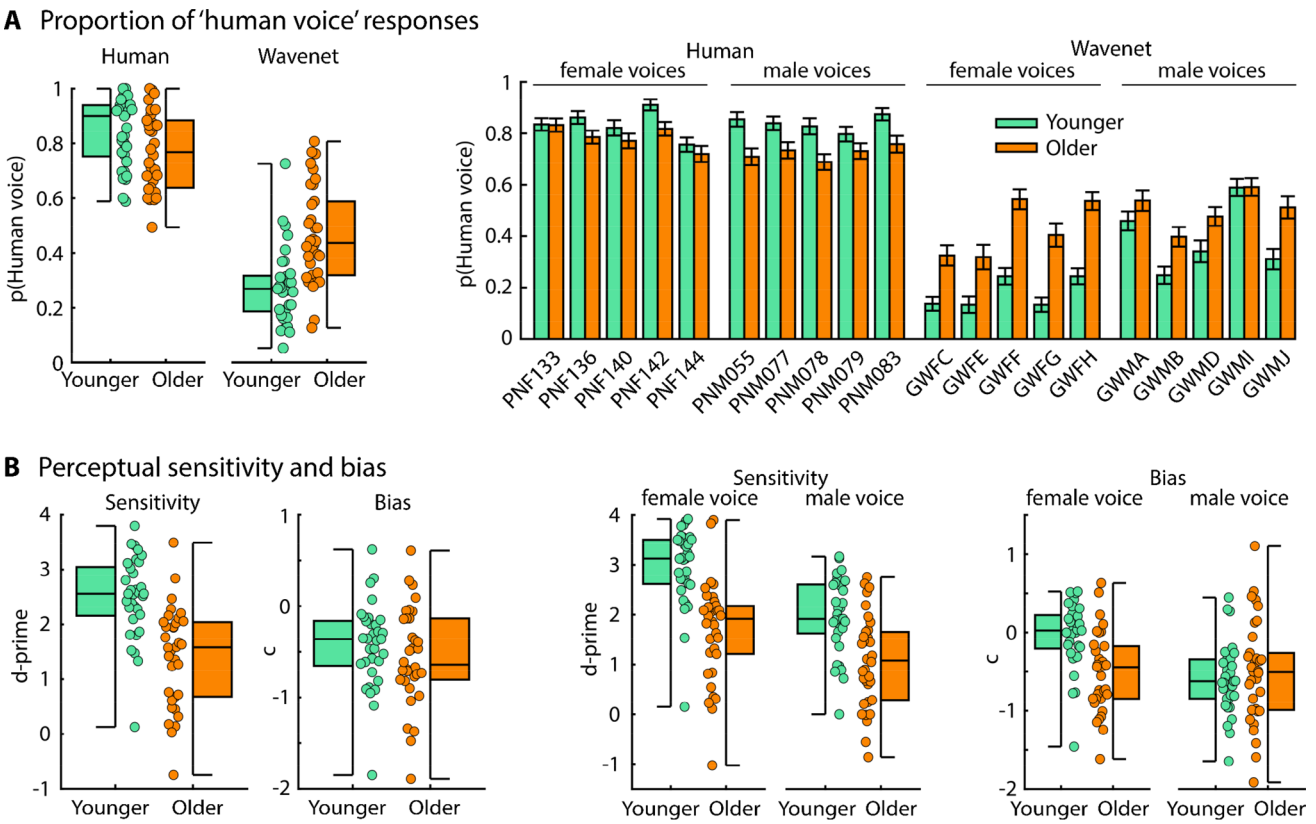

Fig. 9 Performance in human vs AI-speech discrimination. **A** Proportion of 'human voice' responses for human and Wavenet voices. Left: Averaged responses across the 10 human voices and 10 Wavenet voices. Boxplots and data from each individual (dots) are displayed. Right: Responses separately for the 10 human voices (5 male, 5 female) and the 10 Wavenet voices (5 male, 5 female). Error bars reflect the standard error of the mean. For human voices, abbreviations refer to the labels of original stimulus recordings (McCloy et al., 2018; Panfili et al., 2017). For Wavenet voices, the first two letters 'GW' of the abbreviation refers to Google Wavenet, the third letter 'F' vs 'M' refers to female and male, respectively, and the fourth letter refers to the specific Wavenet voice (https://cloud.google.com/text-to-speech/docs/voices). **B** Perceptual sensitivity and bias for younger and older adults across all voices (left) and separately for female vs male voices (right). Boxplots and data from each individual (dots) are displayed

$p = 4.3 \times 10^{-5}$, $\omega^2 = 0.070$), showing that older adults were more biased towards responding "human voice" compared to younger adults for female voices ($t_{65} = 4.185$, $p = 8.7 \times 10^{-5}$, $d = 1.023$), but not for male voices ($t_{65} = -0.619$, $p = 0.538$, $d = 0.151$). In fact, younger adults did not show a bias for female voices (t-test against zero: $t_{32} = 0.474$, $p = 0.639$, $d = 0.082$), only for male voices ($t_{32} = 7.763$, $p = 7.5 \times 10^{-9}$, $d = 1.351$), whereas older adults were biased for both voices (female: $t_{33} = 5.890$, $p = 1.3 \times 10^{-6}$, $d = 1.010$; male: $t_{33} = 4.574$, $p = 6.4 \times 10^{-5}$, $d = 0.784$).

5.3 Summary

Experiment 3 shows that older adults are less able to discriminate between human and AI speech than younger adults. This effect was present for both female and male voices, but the age-group difference was slightly larger for female compared to male voices. Discrimination between human and AI speech was also generally lower for male compared to female voices.

6 General discussion

The current behavioral study investigated how younger and older adults perceive modern AI-based synthesized speech. Specifically, the study explored whether AI-based synthesized speech can be used to investigate common speech-in-noise perception phenomena, here masking release. Speech intelligibility for human speech and synthesized speech was recorded in younger and older adults for speech masked by a modulated or an unmodulated multi-talker babble noise. For both human and AI speech, speech intelligibility was better for modulated compared to unmodulated maskers (i.e., masking release) and older adults showed a reduced masking-release benefit. These results fit with previous work using human-recorded speech and suggest that modern

AI-based synthesized speech could be used for hearing and speech research. The data further suggest that older adults recognize the presentation of AI speech less frequently, rate AI speech as more natural, and are less able to discriminate between human and AI speech compared to younger adults. Hence, research on speech perception in older adults may especially benefit from modern AI-based synthesized speech because they experience it much like spoken by a human.

6.1 Hearing abilities

In all three experiments of the current study, the older-adult group had worse (i.e., higher) digit-in-noise thresholds (Figs. 1, 4, and 8). This is also reflected in overall lower speech-intelligibility performance when older, compared to younger adults, listened to human and Wavenet sentences in Experiments 1 and 2 (Figs. 2 and 5). Estimations of the audiometric pure-tone average thresholds (PTA) based on the digit-in-noise thresholds (Smits et al., 2004, 2013) suggest that older adults in the current study had on average about 10 dB HL higher PTAs, with thresholds for the younger-adult group at around 7 dB HL and thresholds for the older-adult group at around 17 dB HL. This is consistent with in-lab PTA measurements involving community-dwelling older adults (Henry et al., 2017; Herrmann et al., 2018, 2019; Presacco et al., 2016) and, more generally, with the known decline in auditory sensitivity associated with aging (Moore, 2007; Plack, 2014). The sample of older adults in the current study thus had some degree of hearing impairment, consistent with normal aging (individuals with hearing aids were excluded).

6.2 Release from masking is present for human speech and AI-based synthesized speech

Experiment 1 and 2 aimed to investigate whether common speech-in-noise perception phenomena – here release from masking – can be observed using modern AI-based synthesized speech. Consistent with previous work, Experiments 1 and 2 show for both human and AI speech that speech intelligibility was higher for modulated compared to unmodulated maskers (i.e., masking release; cf. Bacon et al., 1998; Cooke, 2006; Dubno et al., 2002; Festen & Plomp, 1990; Gustafsson & Arlinger, 1994; Irsik et al., 2022; Li & Loizou, 2007; Miller & Licklider, 1950), and that this masking release was greater for younger compared to older adults (cf. Bacon et al., 1998; Dubno et al., 2002, 2003; George et al., 2006; Gustafsson & Arlinger, 1994; Irsik et al., 2022; Lorenzi et al., 2006; Summers & Molis, 2004). The magnitude of the masking-release effects did not differ between human and AI speech in Experiment 1, and only the masking-release effect, but not the reduction in release from masking in older adults, was minimally smaller for AI than human speech in Experiment 2 (only trending towards statistical significance). Experiments 1 and 2 may thus suggest that AI-based synthesized speech can provide an alternative way to investigate speech-in-noise perception when the recording of speech materials by a human is financially costly or not feasible for other reasons. Moreover, Google's Wavenet text-to-speech synthesizer used here (van den Oord et al., 2016) enables a variety of acoustic manipulations and use of different languages that could facilitate research in areas where speech materials are hard to come by.

Previous work suggested lower speech intelligibility in the presence of background noise for synthesized speech compared to human speech (Aoki et al., 2022; Cooke et al., 2013; Simantiraki et al., 2018). Interpreting overall differences between speech originating from different speakers or voices can be difficult, as laid out in the introduction, because a number of acoustic differences could cause speech intelligibility differences independently of AI speech. Here, the sets of sentences were matched across human and AI speech in terms of median F0 and duration in order to reduce some acoustic differences. In Experiment 1, speech-intelligibility was better for AI speech than human speech (Fig. 2A; particularly for moderate to low masking levels). This is also consistent with the slightly higher speech intelligibility scores for AI than human speech in Experiment 3 (Fig. 8C). In contrast, intelligibility of AI speech in Experiment 2 was worse compared to human speech (Fig. 5A; particularly for moderate and high masking levels). The difference in overall speech intelligibility for different voices highlights that generalizations about lower speech intelligibility for AI than human speech are challenging. Focusing on interaction designs, as was done here, may facilitate interpretations.

Both Experiments 1 and 2 revealed a steeper slope of the logistic functions that was fit to intelligibility data for AI speech compared to human speech (Figs. 2A and 5A). This means that the variability in speech intelligibility across sentences at different masking levels was lower for AI speech than human speech. The slope differences further make interpretations about overall intelligibility differences between AI speech and human speech more challenging, because interpretations depend on the signal-to-noise ratio used in a specific experiment (often 1–3 levels have been used; e.g., Aoki et al., 2022; Cooke et al., 2013; Cohn & Zellou, 2020; Govender et al., 2019a, 2019b). The reduced variability of intelligibility across sentences with different masking levels perhaps allows using fewer trials to obtain reliable speech-in-noise perception estimates with AI-based synthesized speech compared to human-spoken speech. Overall, the results of Experiment 1 and 2 show that known speech-in-noise perception phenomena can be observed using modern AI-based synthesized speech, and suggest that AI speech could potential be used when speech recordings made by a human are not available or unfeasible to obtain.

6.3 Age-related differences in the ability to distinguish between human and AI speech

The current study shows that older adults, compared to younger adults, are less likely to recognize modern AI-based synthesized speech when they encounter it (Figs. 3 and 6), find AI-based synthesized speech more natural and human-like (Figs. 3 and 6), and are less able to discriminate between human and AI speech (Fig. 9).

There are a few potential factors that could have contributed to the observation that older adults are more likely than younger adults to perceive AI speech as being spoken by a human. One potential explanation is that younger adults have more experience with voice-AI systems compared to younger adults, and as a result are better at recognizing AI speech. However, younger and older adults rated their experience with voice-AI systems, such as Google's Assistant, Amazon's Alexa, or Apple's Siri, about the same, although individuals varied greatly in their rated experience. The absence of a difference in self-reported experience with voice-AI systems has also been reported recently (Zellou et al., 2021) and is more generally consistent with the increasing use of smart phones among older adults (O'Dea, 2021; Statistics-Canada, 2021), the small difference in use of smart homes and speaker between younger and older adults (Laricchia, 2022), and many encounters of voice-AI in everyday life (e.g., grocery self-checkouts, train station announcements, call services). Based on the current data, experience with voice-AI systems is thus unlikely to explain age-group differences in AI-speech recognition.

A related explanation might be that older adults have more experience than younger adults with older less human-like synthesized speech (Klatt, 1980) so that they find modern AI-based synthesized speech more human-like (see discussion in Zellou et al., 2021). However, older adults rated Google's Standard voices—that is, those that are less human-like than Google's Wavenet voices—as more natural and human-like compared to younger adults. It is also unlikely that such potential age differences in experience with older synthesized speech can explain the age differences in perceptual sensitivity to AI speech in Experiment 3, because response bias was removed in these analyses.

Another potential explanation relates to reduced auditory sensitivity in older compared to younger adults. Estimated audiometric PTAs from digit-in-noise perception thresholds were about 10 dB HL higher in older compared to younger adults. The PTA reflects the average threshold across thresholds for tones at 0.5–4 kHz. Age-related differences in hearing thresholds are typically more pronounced at higher frequencies (e.g., at 8 kHz; Herrmann et al., 2022; Moore, 2007; Plack, 2014; Presacco et al., 2016). Any information in the higher frequencies of the speech signals that contribute to the discrimination between human and AI speech would likely be less available to older than younger adults, potentially explaining their lower performance. Moreover, the age-group difference in the discrimination of human vs AI speech was greater for female than male voices (Fig. 9), that is, for the voice gender for which relevant speech information is present in higher frequencies (Fig. 7). However, discrimination between human and AI speech was overall worse for male compared to female voices, and no relationship between the digit-in-noise perception threshold and the perceptual sensitivity to discriminate between human and AI speech was found. Overall, the results suggest that decline in auditory sensitivity is not a primary driver of the age-related decline in the recognition of AI speech.

Finally, much research shows that older adults are less able than younger adults to recognize and discriminate between different emotional messages conveyed in speech prosody (Allen & Brosgole, 1993; Ben-David et al., 2019; Cohen & Brosgole, 1988; Dupuis & Pichora-Fuller, 2010; Kiss & Ennis, 2001; Martzoukou et al., 2022; Mitchell & Kingston, 2011, 2014; Mitchell et al., 2011; Paulmann et al., 2008), although intelligibility of emotional speech appears to be little affected by aging (Dupuis & Pichora-Fuller, 2014). Older adults have further been suggested to attend more to semantic—that is, the lexical content—than to prosodic aspects of speech, whereas younger adults utilize prosodic information more to make judgements about emotions conveyed in speech (Ben-David et al., 2019). The causes of the reduced sensitivity to emotional speech prosody in older adults are still debated, with some work suggesting a potential basic sensory-processing contribution (Mitchell & Kingston, 2014), whereas others rather suggest non-sensory contributions (Dupuis & Pichora-Fuller, 2015; Orbelo et al., 2005). Recognition of AI speech may rely on the processing of prosodic information, potentially explaining a decline in reduced human vs AI speech discrimination in older adults.

7 Conclusions

The current study explored whether state-of-the-art AI-based synthesized speech may provide a useful alternative to human-recorded speech for the investigation of speech-in-noise perception in younger and older adults. The current study showed for both human and AI speech that speech intelligibility was better for modulated than unmodulated maskers (masking release), and that this masking-release benefit was reduced in older adults. Release-from-masking effects were comparable between human and AI speech, suggesting that modern AI speech could be useful for hearing and speech research. The data further suggest that older adults find AI speech more natural and human-like than younger adults, suggesting

perhaps that research on speech perception in older adults may benefit especially from modern AI-based synthesized speech.

Acknowledgements The research was supported by the Canada Research Chair program (232733) and the Natural Sciences and Engineering Research Council of Canada (Grant ID: RGPIN-2021-02602).

Author contributions BH: designed and conducted research, recorded the data, analyzed the data, and wrote the manuscript.

Data availability The datasets generated during and/or analyzed during the current study are available in the Open Science Framework repository, https://osf.io/4jufv/.

Declarations

Conflict of interest The author has no conflicts or competing interests.

References

Agley, J., Xiao, Y., Nolan, R., & Golzarri-Arroyo, L. (2022). Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior Research Methods, 54*, 885–897.

Allen, R., & Brosgole, L. (1993). Facial and auditory affect recognition in senile geriatrics, the normal elderly and young adults. *International Journal of Neuroscience, 68*, 33–42.

Ammari, T., Kaye, J., Tsai, J. Y., & Bentley, F. (2019) Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction 26*(3), Article No. 17.

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.

Ben-David, B. M., Gal-Rosenblum, S., van Lieshout, P. H. H. M., & Shakuf, V. (2019). Age-related differences in the perception of emotion in spoken language: The relative roles of prosody and semantics. *Journal of Speech, Language, and Hearing Research, 62*, 1188–1202.

Bentley, F., LuVogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018) Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2*(3), Article No. 91.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science, 58*, 739–753.

Bilger, R. C. (1984). *Manual for the clinical use of the revised SPIN Test*. The University of Illinois.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341–345.

Brown, L., Mahomed-Asmail, F., De Sousa, K. C., & Swanepoel, D. W. (2019). Performance and reliability of a smartphone digits-in-noise test in the sound field. *American Journal of Audiology, 28*, 736–741.

Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods, 50*, 2586–2596.

Buteau, E., & Lee, J. (2021). Hey alexa, why do we use voice assistants? The driving factors of voice assistant technology use. *Communication Research Reports, 38*, 336–345.

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science, 11*, 464–473.

Cohen, E. S., & Brosgole, L. (1988). Visual and auditory affect recognition in senile and normal elderly persons. *International Journal of Neuroscience, 43*, 89–101.

Cohn, M., Raveh, E., Predeck, K., Gessinger, I., Möbius, B., & Zellou, G. (2020). Differences in gradient emotion perception: Human vs. alexa voices. In: *Proceedings of Interspeech*. Shanghai, China, (pp. 1818–1822).

Cohn, M., & Zellou, G. (2020). Perception of concatenative vs. neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes. In: *Proceedings of Interspeech*. Shanghai, China, (pp. 1733–1737).

Cohn, M., Liang, K.-H., Sarian, M., Zellou, G., & Yu, Z. (2021). Speech rate adjustments in conversations with an Amazon alexa socialbot. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2021.671429

Cohn, M., Segedin, B. F., & Zellou, G. (2022). Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics, 90*, 101123.

Cohn, M., & Zellou, G. (2021). Prosodic differences in human- and alexa-directed speech, but similar local intelligibility adjustments. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2021.675704

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

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

Cruickshanks, 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*, 879–886.

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

De Sousa, K. C., Swanepoel, D. W., Moore, D. R., Myburgh, H. C., & Smits, C. (2020). Improving sensitivity of the digits-in-noise test using antiphasic stimuli. *Ear and Hearing, 41*, 442–450.

Drager, K. D. R., Clark-Serpentine, E. A., Johnson, K. E., & Roeser, J. L. (2006). Accuracy of repetition of digitized and synthesized speech for young children in background noise. *American Journal of Speech-Language Pathology, 15*, 155–164.

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. (2010). Use of affective prosody by young and older adults. *Psychology and Aging, 25*, 16–29.

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

Dupuis, K., & Pichora-Fuller, M. K. (2015). Aging affects identification of vocal emotions in semantically neutral sentences. *Journal of Speech, Language, and Hearing Research, 58*, 1061–1076.

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01694-3

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*, 18–25.

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.

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.

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

Goman, A. M., & Lin, F. R. (2016). Prevalence of hearing loss by severity in the United States. *American Journal of Public Health, 106*, 1820–1822.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*, 93–104.

Govender, A., Wagner, A. E., & King, S. (2019a). Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise. In: *Proceedings of Interspeech*, Graz, Austria, (pp. 1551–1555).

Govender, A., Valentini-Botinhao, C., & King, S. (2019b). Measuring the contribution to cognitive load of each predicted vocoder speech parameter in DNN-based speech synthesis. In *Proceedings 10th ISCA Speech Synthesis Workshop* (pp. 121–126). https://doi.org/10.21437/SSW.2019-22.

Greene, B. G., Logan, J. S., & Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, & Computers, 18*, 100–107.

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

Henry, M. J., Herrmann, B., Kunke, D., & Obleser, J. (2017). Aging affects the balance of neural entrainment and top-down neural modulation in the listening brain. *Nature Communications, 8*, 15801.

Herrmann, B., Buckland, C., & Johnsrude, I. S. (2019). Neural signatures of temporal regularity processing in sounds differ between younger and older adults. *Neurobiology of Aging, 83*, 73–85.

Herrmann, B., Maess, B., & Johnsrude, I. S. (2018). Aging affects adaptation to sound-level statistics in human auditory cortex. *The Journal of Neuroscience, 38*, 1989–1999.

Herrmann, B., Maess, B., & Johnsrude, I. S. (2022). A neural signature of regularity in sound is reduced in older adults. *Neurobiology of Aging, 109*, 1–10.

Holder, J. T., Levin, L. M., & Gifford, R. H. (2018). Speech recognition in noise for adults with normal hearing: Age-normative performance for AzBio, BKB-SIN, and QuickSIN. *Otology & Neurotology, 39*, e972–e978.

IEEE. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*(3), 225–246.

Irsik, V. C., Almanaseer, A., Johnsrude, I. S., & Herrmann, B. (2021). Cortical responses to the amplitude envelopes of sounds change with age. *The Journal of Neuroscience, 41*, 5045–5055.

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

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods, 8*, 614–629.

Kim, S. (2021). Exploring how older adults use a smart speaker-based voice assistant in their first interactions: Qualitative study. *JMIR Mhealth and Uhealth, 9*, e20427.

Kiss, I., & Ennis, T. (2001). Age-related decline in perception of prosodic affect. *Applied Neuropsychology, 8*, 251–254.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 67*, 971–995.

Koole, A., Nagtegaal, A. P., Homans, N. C., Hofman, A., Baatenburg de Jong, R. J., & Goedegebure, A. (2016). Using the digits-in-noise test to estimate age-related hearing loss. *Ear and Hearing, 37*, 508–513.

Laricchia, F. (2022) Smart home product ownership rates in the U.S. 2020. Retrieved July 8, 2022, from https://www.statista.com/statistics/799584/united-states-smart-home-device-survey-by-age/

Lewis, J. R. (2018). Investigating MOS-X ratings of synthetic and human voices. *Voice Interaction Design, 2*, 1–22.

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

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*, 433–442.

Liu, C., & Jin, S.-H. (2019). Psychometric functions of vowel detection and identification in long-term speech-shaped noise. *Journal of Speech, Language, and Hearing Research, 62*, 1473.

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.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

MacPherson, A., & Akeroyd, M. A. (2014). Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey. *Trends in Hearing, 18*, 2331216514537722.

Martzoukou, M., Nasios, G., Kosmidis, M. H., & Papadopoulou, D. (2022). Aging and the perception of affective and linguistic prosody. *Journal of Psycholinguistic Research*. https://doi.org/10.1007/s10936-022-09875-7

Masalski, M., Adamczyk, M., & Morawski, K. (2021). Optimization of the speech test material in a group of hearing impaired subjects: A feasibility study for multilingual digit triplet test development. *Audiology Research, 11*, 342.

McCloy, D. R., Panfili, L., John, C., Winn, M., Wright, R. A. (2018). Gender, the individual, and intelligibility. In: *176th Meeting of the acoustical society of America*. Victoria, BC, Canada.

McDermott Josh, H., & Simoncelli Eero, P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron, 71*, 926–940.

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

Milne-Ives, M., de Cock, C., Lim, E., Shehadeh, M. H., de Pennington, N., Mole, G., Normando, E., & Meinert, E. (2020). The effectiveness of artificial intelligence conversational agents in health care: Systematic review. *Journal of Medical Internet Research, 22*, e20346.

Mitchell, R. L. C., & Kingston, R. A. (2011). Is age-related decline in vocal emotion identification an artefact of labelling cognitions? *International Journal of Psychological Studies, 3*, 156–163.

Mitchell, R. L. C., & Kingston, R. A. (2014). Age-related decline in emotional prosody discrimination. *Experimental Psychology, 61*, 215–223.

Mitchell, R. L. C., Kingston, R. A., & Barbosa Bouças, S. L. (2011). The specificity of age-related decline in interpretation of emotion cues from prosody. *Psychology and Aging, 26*, 406–414.

Moore, B. C. J. (2007). *Cochlear hearing loss: Physiological psychological and technical issues*. Wiley.

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.

O'Brien, K., Liggett, A., Ramirez-Zohfeld, V., Sunkara, P., & Lindquist, L. A. (2020). Voice-controlled intelligent personal assistants to support aging in place. *Journal of the American Geriatrics Society, 68*, 176–179.

O'Dea, S. (2021) Smartphone ownership in the U.S. 2015–2021. https://www.statista.com/statistics/489255/percentage-of-us-smartphone-owners-by-age-group/.

Orbelo, D. M., Grim, M. A., Talbott, R. E., & Ross, E. D. (2005). Impaired comprehension of affective prosody in elderly subjects is not predicted by age-related hearing loss or age-related cognitive decline. *Journal of Geriatric Psychiatry and Neurology, 18*, 25–32.

Panfili, L. M., Haywood, J., McCloy, D. R., Souza, P. E., & Wright, R. A. (2017). The UW/NU Corpus, Version 2.0 https://depts.washington.edu/phonlab/projects/uwnu.php.

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). How aging affects the recognition of emotional speech. *Brain and Language, 104*, 262–269.

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.

Plack, C. J. (2014). *The sense of hearing*. Psychology Press.

Polkosky, M. D., & Lewis, J. R. (2003). Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology, 6*, 161–182.

Potgieter, J. M., Swanepoel, W., & Smits, C. (2018). Evaluating a smartphone digits-in-noise test as part of the audiometric test battery. *South African Journal of Communication Disorders, 65*, e1–e6.

Presacco, A., Simon, J. Z., & Anderson, S. (2016). Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. *Journal of Neurophysiology, 116*, 2346–2355.

Raitio, T., Suni, A., Vainio, M., & Alku, P. (2014). Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise. *Computer Speech & Language, 28*, 648–664.

Richter F (2020) Smart speaker adoption continues to rise. Retrieved June 30, 2022, from https://www.statista.com/chart/16597/smart-speaker-ownership-in-the-united-states/

Ross, B., Dobri, S., & Schumann, A. (2021). Psychometric function for speech-in-noise tests accounts for word-recognition deficits in older listeners. *The Journal of the Acoustical Society of America, 149*, 2337–2352.

Salza, P. L., Foti, E., Nebbia, L., & Oreglia, M. (1996). MOS and pair comparison combined methods for quality evaluation of text-to-speech systems. *Acta Acustica United with Acustica, 82*, 650–656.

Simantiraki, O., Cooke, M., & King, S. (2018). Impact of different speech types on listening effort. In: *Proceedings of Interspeech*. Hyderabad, India.

Simpson, C. A., & Hart, S. G. (1977). Required attention for synthesized speech perception for two levels of linguistic redundancy. *The Journal of the Acoustical Society of America, 61*, S7–S7.

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., & Houtgast, T. (2005). Results from the Dutch speech-in-noise screening test by telephone. *Ear and Hearing, 26*, 89–95.

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.

Smits, C., Kramer, S. E., & Houtgast, T. (2006). Speech reception thresholds in noise and self-reported hearing disability in a general adult population. *Ear and Hearing, 27*, 538–549.

Statistics-Canada (2021) Table 22-10-0115-01 Smartphone use and smartphone habits by gender and age group. Retrieved July 8, 2022, from https://www150.statcan.gc.ca/t151/tbl151/en/tv.action?pid=2210011501

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

Taylor, P., & Isard, A. (1997). SSML: A speech synthesis markup language. *Speech Communication, 21*, 123–133.

Thomas, K. A., & Clifford, S. (2017). Validity and mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77*, 184–197.

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: *Proceedings 9th ISCA Workshop on Speech Synthesis Workshop* (SSW 9), (p. 125).

Wingfield, A., Lindfield Kimberly, C., & Goodglass, H. (2000). Effects of age and hearing sensitivity on the use of prosodic information in spoken word recognition. *Journal of Speech, Language, and Hearing Research, 43*, 915–925.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*, 2064–2072.

Zellou, G., Cohn, M., & Ferenc Segedin, B. (2021). Age- and gender-related differences in speech alignment toward humans and voice-AI. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2020.600361