scientific reports

Check for updates

OPEN

Automatic development of speech-in-noise hearing tests using machine learning

Sigrid Polspoel[1,2], David R. Moore[3,4], De Wet Swanepoel[5,6], Sophia E. Kramer[1,2] & Cas Smits[2,7]✉

Understanding speech in noisy environments is a primary challenge for individuals with hearing loss, affecting daily communication and quality of life. Traditional speech-in-noise tests are essential for screening and diagnosing hearing loss but are resource-intensive to develop, making them less accessible in low and middle-income countries. This study introduces an artificial intelligence-based approach to automate the development of these tests. By leveraging text-to-speech and automatic speech recognition (ASR) technologies, the cost, time, and resources required for high-quality speech-in-noise testing could be reduced. The procedure, named "Aladdin" (Automatic LAnguage-independent Development of the digits-in-noise test), creates digits-in-noise (DIN) hearing tests through synthetic speech material and uses ASR-based level corrections to perceptually equalize the digits. Traditional DIN tests were compared with newly developed Dutch and English Aladdin tests in listeners with normal hearing and hearing loss. Aladdin tests showed 84% specificity and 100% sensitivity, similar to the reference DIN tests (87% and 100%). Aladdin provides a universal guideline for developing DIN tests across languages, addressing the challenge of comparing test results across variants. Aladdin's approach represents a significant advancement in test development and offers an efficient enhancement to global screening and treatment for hearing loss.

Keywords Artificial intelligence (AI), Synthetic speech, Text-to-speech (TTS), Automatic speech recognition (ASR), Aladdin, Digits-in-noise test

Abbreviations

AI	Artificial intelligence
Aladdin	Automatic language-independent development of the digits-in-noise test
API	Application programming interface
ASR	Automatic speech recognition
dB HL	Decibels hearing level
DIN	Digits-in-noise
FADE	Framework for Auditory Discrimination Experiments
GBFB	Gabor filter bank
HL	Hearing loss
LC	Level corrections
LTASS	Long-term average speech spectrum
MFCCs	Mel-frequency cepstral coefficients
MOS	Mean opinion scale
NVA	Nederlandse vereniging voor audiologie
NH	Normal hearing

[1]Otolaryngology-Head and Neck Surgery, Section Ear and Hearing, Amsterdam UMC location Vrije Universiteit Amsterdam, De Boelelaan, Amsterdam, The Netherlands. [2]Amsterdam Public Health research institute, Quality of Care, Amsterdam, The Netherlands. [3]Division of Patient Services Research, Cincinnati Children's Hospital Medical Center, and Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA. [4]Manchester Centre for Audiology and Deafness, University of Manchester, Manchester, UK. [5]Department of Speech-Language Pathology and Audiology, University of Pretoria, Pretoria, South Africa. [6]Department of Otolaryngology-Head and Neck Surgery, University of Colorado School of Medicine, Aurora, CO, USA. [7]Otolaryngology-Head and Neck Surgery, Ear and Hearing, Amsterdam UMC location University of Amsterdam, Meibergdreef, Amsterdam, The Netherlands. ✉email: c.smits@amsterdamumc.nl

SEM	Standard error of measurement
SNR	Signal-to-noise ratio
SRT	Speech recognition threshold
TTS	Text-to-speech
WHO	World health organization

The ability to understand speech in noisy environments is the most common challenge for individuals with hearing loss, impacting daily communication and quality of life. Traditional speech-in-noise tests, essential for assessing the ability to recognize speech in noise, diagnosing hearing loss and evaluating hearing aid and cochlear implant performance, require extensive resources for development. This limits their availability, especially in low and middle-income countries. The present study introduces an innovative approach using artificial intelligence (AI) to automate the development of such tests. By employing text-to-speech and automatic speech recognition technologies, this approach drastically reduces the cost, time and resources required to develop high-quality speech-in-noise testing accessible worldwide.

Since the establishment of the field of audiology, the importance of using speech-in-noise tests has been recognized[1]. Speech-in-noise tests measure an individual's ability to recognize speech by presenting speech stimuli (usually lists of digits, words or sentences) in the presence of noise. The outcome of the test is either a percentage-correct score or the signal-to-noise ratio (SNR) where the average speech recognition score is equal to a certain percentage-correct[2]. Typically, the listener is presented with a series of speech stimuli and is asked to repeat what they heard. One example of such a test, and the main focus of the present study, is the widely used digits-in-noise (DIN) test[3,4]. It uses digit triplets presented in speech-shaped noise to measure the speech recognition threshold (SRT). The SRT is the signal-to-noise ratio, expressed in dB SNR, at which the listener recognizes 50% of the digit triplets correctly. Apart from a diagnostic test in clinics[3], the DIN test can be completed at home via internet or a smartphone app as a quick self-test for hearing screening[3,5–7], or a test for cochlear implant users[8–10]. The test results strongly correlate with standard audiometric findings[11] and can reach a large global audience without the need of an examiner[11–13] or the requirement to calibrate the devices at home[11,12]. The World Health Organization has adopted this test for its official hearing screening app, called hearWHO[14], and included the DIN test as a recommended option for hearing screening in adults and school-aged children[15].

An essential property of a speech-in-noise test is the precision of the test and, for use as a screening test, its specificity and sensitivity. The precision of the SRT estimated in a speech-in-noise test is expressed by the standard error of measurement (SEM). A small SEM is necessary to produce valid and reliable speech-in-noise test results. The SEM is inversely proportional to the slope of the speech recognition function of the test[16], a sigmoidal function representing the percentage correctly repeated speech items against the SNR at which they are presented. The speech recognition function is described by the SRT (i.e., SNR at 50% correct; the midpoint of the function), and the slope of the speech recognition function at the midpoint. Because each test involves a series of different stimuli, the average speech recognition function is formed by the speech recognition functions of the individual stimuli[17,18]. Differences in intelligibility between the stimuli make the slope of the average speech recognition function shallower and therefore increase the SEM. Thus, ensuring perceptual equivalence among individual speech stimuli is important as it steepens the speech recognition function and minimizes SEM. For hearing screening tests, additional properties determine the quality of a test: sensitivity refers to the ability of the test to identify those with hearing loss and specificity refers to the ability of a test to identify those without hearing loss.

Making the speech stimuli perceptually equivalent is a crucial aspect in the development of an accurate and precise speech-in-noise test. This is generally a costly and labor-intensive process (Fig. 1). It involves the professional recording of speech stimuli uttered by trained speakers and selecting the best utterances. Then a listening experiment is conducted with listeners with normal hearing to use the results for selecting or creating perceptually equivalent stimuli[19,20]. The need for trained speakers, listeners with normal hearing, professional recording equipment, audiometric equipment, and a sound attenuating booth significantly limits the development of new speech-in-noise tests, especially in low and middle-income countries. The aim of the current study was therefore to develop a procedure for the automatic development of speech-in-noise hearing tests using AI (Fig. 1).

In the current study, the procedure of creating a DIN test was drastically shortened and simplified by AI techniques. Specifically, text-to-speech (TTS) and automatic speech recognition systems (ASR) were deployed to create DIN tests. TTS was used to create the synthetic speech material (digits) and ASR to determine level corrections needed to achieve perceptual equality among the synthetic digits (Fig. 1). We refer to this as 'Aladdin'; "Automatic LAnguage-independent Development of the Digits-In-Noise test". Aladdin's goal is to provide a simple, efficient, automatic, low-cost and universal way to create new DIN tests in different languages.

Aladdin requires a high-quality TTS system which, thanks to recent advances in deep neural network techniques, can now approach the naturalness of human speech[21]. Synthetic speech has emerged as a successful alternative to natural speech for creating speech material in various languages, as demonstrated for speech recognition in quiet[22] and speech recognition in noise[23–27]. In addition to a TTS system, Aladdin requires ASR to determine level corrections per digit needed to equalize intelligibility among the digits. Two types of ASRs were considered: off-the-shelve, pre-trained systems with large language models; and a specific system that was trained on the speech stimuli of choice.

There were four parts to the present study. **Part I** – Creating synthetic speech material in five languages and assessing its subjective speech quality. The speech quality of synthetic material and original speech material was measured using the 5-point mean opinion score (MOS) scale. Assessment was done by native listeners in their own country via a calibrated laptop. The goal was to determine if the subjective quality of synthetic speech

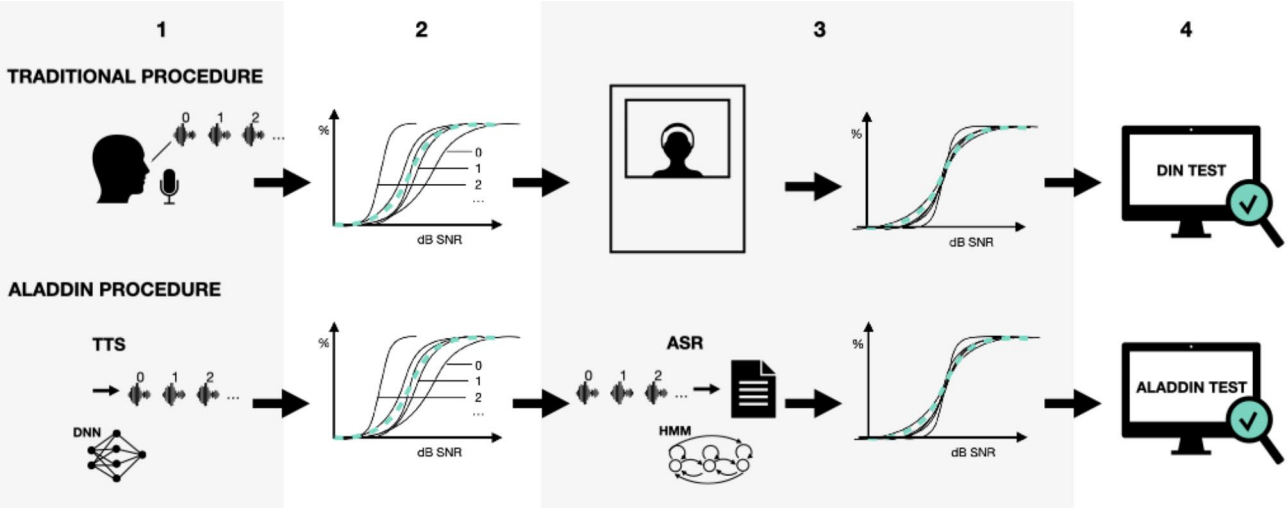

Fig. 1. Schematic representation of the traditional development process of the DIN test's speech material (top panel) and the automated process *Aladdin* (bottom panel). Traditional procedure: (1) Digits are recorded by a professional speaker and the best recordings are selected; (2) Speech recognition functions of the recorded digits differ in slope and SRT, i.e., the intelligibility of the digits is not equal. The blue dotted line represents the average speech recognition function of the digits; (3) A listening experiment with listeners with normal hearing is set up to determine the level corrections per digit needed to achieve equal intelligibility. Next, level corrections are applied, horizontally shifting the speech recognition functions to the mean SRT. Now all digits have the same expected SRT, and the average speech recognition function (blue) has become steeper than in step 2, resulting in a smaller SEM; (4) The test is evaluated with groups of listeners with normal hearing and hearing loss. The automated process, Aladdin, replaces steps 1 and 3 of the traditional development process. Aladdin procedure: (1) A deep neural network (DNN)-based text-to-speech (TTS) system creates digit recordings; (3) Level corrections are determined using the output of a Hidden Markov model-based automatic speech recognition system (ASR). The corrections are subsequently applied to the individual digit recordings.

material was comparable to reference speech material used in audiometric testing. If so, we can conclude that synthetic speech is a viable option for creating speech material for speech-in-noise tests. **Part II** – Comparing speech recognition functions of natural and synthetic digits. A listening experiment was conducted to determine speech recognition functions for digit triplets using male and female Dutch synthetic voices, and a male Dutch natural voice (reference Dutch speech material). SRTs, slopes and variability of the natural and synthetic digit recognition functions were compared. Results yielded level corrections per digit (0–9), useable as a human reference for the ASR. **Part III** - Feeding natural and synthetic digits in noise to different ASRs to determine level corrections for each digit automatically. These level corrections were compared to those based on human listeners from Part II. A strong correlation between the level corrections, as well as a low mean squared error between the human and TTS based level corrections, would suggest that the human listening experiment can be replaced by an ASR. The ASR that resembled the human performance the most was used for the remainder of the study. Next, we applied the ASR-based level corrections to the synthetic digits, making the digits perceptually equally intelligible. At this point, the speech material of the Aladdin-test was created. **Part IV** – Evaluation of the Aladdin test in listeners with normal hearing and listeners with hearing loss. Both groups performed the Aladdin test in Dutch and English, as well as a reference DIN test in both languages. The outcomes and screening characteristics were compared to evaluate whether the Aladdin tests are valid alternatives to the reference DIN tests that were developed the traditional way.

Part I: subjective quality of synthetic (text-to-speech) speech material for hearing tests

The aim of part I was to select a TTS system suitable for generating high-quality speech material in multiple languages. Generating speech material is the first step in developing a speech-in-noise test (Fig. 1). Although the primary goal of the project was to develop digits-in-noise tests, we chose to evaluate the TTS system not only with digit triplets as speech material but also by examining the application of TTS for other standard types of speech material, namely words and sentences. A TTS system was used to generate synthetic variants of standard audiological speech material in five languages (Dutch, English, French, Spanish, and one tonal language, Mandarin). Both synthetic and natural speech material were evaluated locally by native speakers using a standardized protocol to assess speech quality. They were provided with a calibrated laptop with self-explanatory software to complete the speech rating task. The speech stimuli were presented in both quiet and against low level of background noise. The hypotheses were that speech quality would be rated higher for digit triplets than for words, and lowest for sentences. Additionally, we hypothesized that speech quality would be rated lower in quiet conditions, as low-level imperfections in the synthetic speech would be masked in the

noise condition. Note that this experiment was solely utilized for the evaluation of this phase of the Aladdin test development procedure. It will not be required to conduct this phase in future developments of Aladdin tests in different languages.

Method

Participants

Fifty native listeners participated: 10 Dutch (mean age 38 ± 12 years), 10 American (M = 37 ± 8 years), 10 French (M = 27 ± 1 years), 10 Spanish (M = 36 ± 5 years) and 10 Chinese (Mandarin; M = 31 ± 7 years). The Dutch participants were recruited by the first author of this paper. Cochlear, a partner in this study, assigned an American, French, Spanish, and Chinese contact person to recruit and instruct 10 participants (5 male, 5 female) each in the USA, France, Spain, and China, respectively. Contacts were instructed to select individuals under 50 years old with no self-reported hearing problems and basic proficiency in English to understand the software instructions. Most participants were Cochlear employees, some were acquaintances of the contact people. A calibrated laptop was sent to Cochlear offices in the different countries. After obtaining test results for 10 participants, the equipment was returned and then sent to the next contact person in a different country. This process occurred from October 2021 to July 2023. All participants provided written informed consent and all experiments were performed according to the Netherlands Code of Conduct for Research Integrity. The VUmc medical research ethics committee approved the experimental protocol (protocol number: 2021.0060).

Material

After assessing multiple TTS systems for generating speech material, Google Cloud TTS emerged as the preferred choice. Its selection was based on its superior speech quality, well-documented technology, extensive range of voices and languages, and an application programming interface (API) that streamlines bulk speech production. Currently, Google Cloud TTS supports over 40 languages and offers more than 220 voices by leveraging powerful neural networks known as WaveNet through its API[28]. Dutch voices were chosen based on naturalness by two speech therapists from Amsterdam UMC, while voices for other languages were selected by the first author based on perceived speech quality. Google Cloud TTS produced 16-bit WAV files for digits, words, and sentences at a 44.1 kHz sample rate. Our consideration of multiple languages aligns with Aladdin's goal to automate hearing test creation for any language.

The speech material, created in five languages (Dutch, English, French, Spanish, and Mandarin), included exact replications of digits, words, and sentences used for standard speech audiometry. Examples of standard speech materials in the Netherlands are the original DIN test digits[29], NVA (Nederlandse Vereniging voor Audiologie) words[30], and VU-98 (Vrije Universiteit 1998) sentences[31]. The corresponding countries where the speech material is used are the Netherlands, USA, France, Spain, and China. Table 1 provides an overview of the speech material types used in each country and the synthetic Google voices used for their production. All synthetic speech material was created in both a male and female synthetic voice. Additionally, the experiment included both natural and degraded voices, wherein the natural voice represented the standard speech material and served as a benchmark for the synthetic voices. The degraded voice was a 12-channel noise-carried vocoded voice and added to cover the entire range of the scale. A standard vocoder script for PRAAT was used[32,33].

A listening experiment, following International Telecommunication Union - Telecommunications sector (ITU-T) P-series recommendations for speech quality assessment[19], was conducted to evaluate subjective quality across various speech materials and languages. Self-developed software facilitated listeners, with different mother tongues, to select their native language, rate audio sound files in quiet and noise, and answer the question: "Overall impression - how would you rate the quality of the speech you just heard?". The text "ignore the noise" was added for sound file presentation in background noise. Ratings were provided on the 5-point MOS scale, ranging from "Excellent" (5 points) to "Very poor" (1 point)[34]. Official translations of the scale were utilized, and translations of the "Overall impression" question were authenticated by native speakers. All translations are available in the Supplementary Note 1. The software instructions were presented in English.

The software was installed on a laptop (Dell Latitude 5580 Core 15-6300U, 16 GB) provided with headphones (Sennheiser HD 280 Pro). Participants were asked to complete a 15-minute speech rating task on it. The laptop was calibrated so that all speech sound files were presented binaurally at a level of 65 dB SPL.

	Dutch	English	French	Spanish	Mandarin
Digits	Digits of the DIN test[29]	hearWHO digits[71]	hearWHO digits[14]	hearWHO digits[14]	hearWHO digits[14]
Words	NVA word lists[30]	NU-6[74]	Fournier disyllabic words[75]	Test de Navarram developed by Clinica Universitaria de Navarra (personal communication)	Mandarin PB monosyllable speech test[76]
Sentences	VU-98 sentences[31]	American English Hearing in Noise Test[77]	MBAA2[78]	SharvardCorpus[79]	HOPE Mandarin sentences[80]
Synthetic Voices	M: nl-NL Wavenet B F: nl-NL Wavenet D	M: en-US Wavenet B F: en-US Wavenet F	M: fr-FR Wavenet C F: fr-FR Wavenet D	M: es-ES Wavenet B F: es-ES Wavenet C	M: cmn-CN Wavenet B F: cmn-CN Wavenet A

Table 1. Overview of the speech material types used in each country and the male (M) and female (F) synthetic Google voices used for their production.

Testing procedure

The design of the speech assessment task involved within-subject repeated measures. Participants, guided by a contact person, completed the test independently on a laptop in a quiet environment. Using the software, participants selected language, age, gender, and indicated normal hearing. The test comprised 6 sections, with participants rating sound files on the 5-point MOS scale[34]. Sections were presented in a fixed sequence and included: (1) digits in quiet, (2) digits in noise (3) words in quiet (4) words in noise (5) sentences in quiet and (6) sentences in noise. Each section featured four voices (natural, synthetic male, synthetic female, degraded), each presented randomly four times. Consequently, there were 16 presentations to evaluate in each section, comprising four practice runs with the four voices at the onset of each segment to familiarize the participant with the task (these were not scored). Each voice in each section was scored three times per participant, and these scores were averaged. Section (1) and (2) comprised three digit triplets per presentation, (3) and (4) had six words, and (5) and (6) featured three sentences per presentation. Each presentation lasted around 5 to 6 s. Participants could take breaks between sections. Average scores per section per voice were calculated after collecting all speech rating data, adhering to standard MOS regulations[34]. The experiment had a total duration of approximately 15 min.

Statistical analysis

The Kruskal-Wallis test compared scores of synthetic and natural voices across languages for digits, words, and sentences. It serves as a non-parametric alternative to the F-test due to violations of the normality assumption of the dependent variable.

Results

MOS scores for the different languages for the synthetic speech and natural speech, averaged across participants and repetitive presentations, are shown in Fig. 2.

The majority (16 out of 18) of the synthetic and natural speech sound files received an average MOS score above 4, meaning that all voices, except for the degraded one, were considered to be of good to excellent speech quality. Across languages, there were no significant differences between the male and female synthetic voices and the natural voice in either the digits ($\chi^2(2) = 0.05$, $p = 0.98$), words ($\chi^2(2) = 0.68$, $p = 0.71$), or sentences ($\chi^2(2) = 0.54$ $p = 0.77$) material in quiet. Similarly, there were no significant differences observed in noisy conditions (digits: $\chi^2(2) = 1.23$, $p = 0.54$; words: $\chi^2(2) = 1.27$, $p = 0.53$; sentences: $\chi^2(2) = 0.22$, $p = 0.89$). The degraded voice was, as expected, consistently rated the worst. The results suggest that synthetic speech material is not only suitable for speech-in-noise tests, but also for speech-in-quiet tests. Because ratings of synthetic and natural voices did not differ in any of the three speech materials, the results also suggest that synthetic speech can be effectively used for both digits-in-noise tests and tests involving other words or sentences.

Discussion

Results from the listening experiment with native speakers in five languages showed good to excellent MOS scores that did not differ significantly between synthetic and natural speech in any language across various speech materials (digits, words, sentences) and conditions (quiet and noise). We hypothesized that the synthetic voice would be rated worse on the sentence material compared to the digits material, because longer speech fragments with more prosody and co-articulation could reveal subtle unnatural sounds in the synthetic voice. However, our results do not support this hypothesis. Even the speech quality of daily life sentences presented in quiet was rated as good to excellent, suggesting that modern TTS systems can replace human speakers for creating most of the current types of speech material of hearing tests. This is consistent with studies that have employed synthetic speech for speech-in-noise tests with digit triplets[26], word triplets[23], matrix sentences[27], and everyday sentences[24,25]. Polspoel, et al.[22] demonstrated that synthetic word lists in quiet are even more perceptually equivalent than the commonly used Dutch natural lists for speech audiometry, without requiring adjustments post-creation of the synthetic words. Our results show, in any case, strong evidence that the subjective quality of synthetic speech is not a limiting factor when using synthetic digits in the development of DIN tests, consistent with the objective of this study.

Part II: speech recognition functions of natural and synthetic digits

Further potential differences between the natural and synthetic voices were explored by comparing speech recognition functions of individual digits (Fig. 1). To do so, speech recognition functions of the synthetic male, the synthetic female and the natural male Dutch digits from Part I were determined by presenting the digits at fixed SNRs to a group of listeners with normal hearing. These speech recognition functions were also used to determine the human-derived level corrections required for each set of digits to perceptually equalize them.

Method

Participants

Twenty-four native Dutch-speaking adults (M = 21 ± 2 years) with normal hearing participated. All participants had pure-tone thresholds ≤ 20 dB HL across octave frequencies from 0.25 to 8 kHz in the test ear. Only one ear was tested in the subsequent speech recognition experiment. The mean pure-tone average (PTA) at 0.5, 1, 2, and 4 kHz of the tested ears was 3.1 ± 2.9 dB HL. Left and right ears were alternated between participants, unless only one ear qualified. Recruitment occurred on the university campus, with participants receiving compensation and providing written informed consent. All experiments were performed according to the Netherlands Code of Conduct for Research Integrity. The VUmc medical research ethics committee approved the experimental protocol (protocol number: 2020.0758).

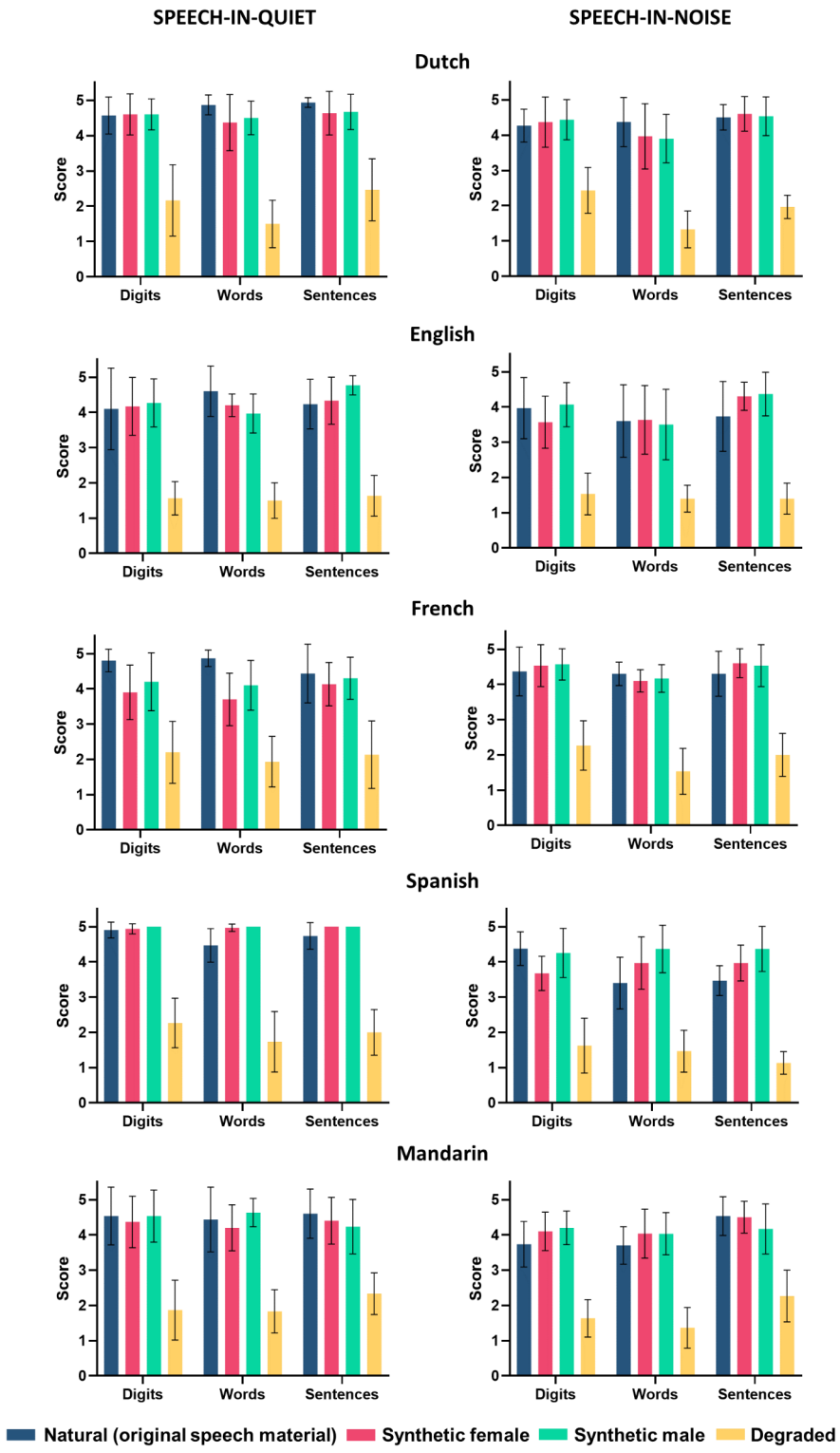

Fig. 2. Average MOS scores for three types of speech material (digits, words and sentences) in four voices. The different voices include: natural, synthetic female, synthetic male and degraded. Speech quality ratings for speech stimuli presented in quiet are shown on the left hand side, speech quality ratings for speech stimuli presented in low-level of noise are shown on the right hand side. Scores are the means (± SD) of 10 participants per language.

Materials

The synthetic (female and male voice) and natural Dutch digits (0 to 9) from Part I were used for further evaluation. The original natural digits and speech-shaped, stationary masking noise were taken from the standard Dutch DIN test[29]. These natural digits were pronounced by a male professional speaker. While recommending the use of female voices for future automatically generated DIN tests ('Aladdin-tests') for consistency, this study also included a male synthetic voice for better comparison with the natural male voice.

The average spectrum of the female synthetic digits (long-term average speech spectrum, LTASS) was adjusted to match the 'idealized speech spectrum' from the speech intelligibility index (SII) standard: a constant sound pressure spectrum level from 100–500 Hz and decreasing from 500–9500 Hz at a rate of 9 dB per octave. The LTASS adjustment involved correcting the synthetic speech spectrum by the difference between the LTASS of the synthetic and the idealized LTASS using PRAAT[33]. This ensures a uniform speech (and noise) spectrum for all future Aladdin test materials, preventing different effects of audibility on DIN test results between Aladdin tests. Note that the masking noise is therefore identical for all Aladdin tests, with its spectrum mirroring the idealized LTASS. The spectra of the natural digits and the corresponding noise were left unadjusted. The LTASS of the male synthetic digits and noise was aligned with natural DIN material using PRAAT in a similar manner. Subsequently, the root mean square (RMS) levels of all synthetic and natural digits were equalized. Following the approach outlined by Smits, et al.[29], 120 unique digit triplets were constructed from the 10 digits in each voice, with each triplet containing three unique digits. Silent intervals of 500 ms preceded the first digit and followed the final digit, while 200 ms intervals separated the digits. The testing equipment for the listening task included Sennheiser HDA200 headphones connected to a Dell Optiplex780 PC via a Sound Blaster Creative soundcard (THX).

Testing procedure

The experimental study employed a within-subject repeated measures design. Testing took place in a sound-treated booth and stimuli were monaurally presented to the listener via headphones. DIN tests were presented at fixed SNRs of -14, -12, -10, -8, -6, and 0 dB, with an overall presentation level set at 65 dB SPL. Participants were divided equally, with one-third starting with the female synthetic DIN test, one-third with the male synthetic DIN test, and one-third with the male natural DIN test. Each test comprised 120 triplets organized into 6 lists of 20 triplets, ensuring each digit appeared 6 times per list. Participants encountered one list per presentation level for each DIN test, maintaining consistent presentation numbers per digit and signal-to-noise ratio (SNR) for speech recognition function estimation. The order of the 6 lists remained fixed for each participant across the three DIN tests to minimize the potential effect of learning effects. To mitigate fatigue, the SNR order alternated between easier (-4, -6, and 0 dB SNR) and more challenging (-14, -12, and − 10 dB SPL) conditions. Each DIN test began with a practice list to minimize potential learning effects.

The participants were verbally instructed to repeat three digits after each trial and encouraged to guess if uncertain. When they could not understand one or more digits, they could say "blank" (e.g., "4, blank, 7"). In data analysis, blanks were replaced with a random digit, resulting in a fixed 10% guess rate per digit. Stimuli were not repeated, and no feedback was given.

Results

Figure 3 [a-c] shows the speech recognition (performance-intensity) functions for the individual digits in noise for respectively the natural male, synthetic female and synthetic male voice. Each dot represents the digit-specific mean percentage correct of the 24 participants with normal hearing for each presentation level. The lines represent maximum likelihood fits of logistic functions to the raw data per digit. The lower asymptote (guess rate) was set at 10%, which means the speech recognition function is described by: $10 + 90/\left[1 + e^{(-4S(SNR-SRT))}\right]$

, where S represents the slope of the speech recognition function. The mean (± standard deviation) SRTs and mean slopes of the digits were − 9.4 ± 3.2 dB SNR and 14.9 ± 4.1%-points/dB for the natural voice, -8.6 ± 3.3 dB

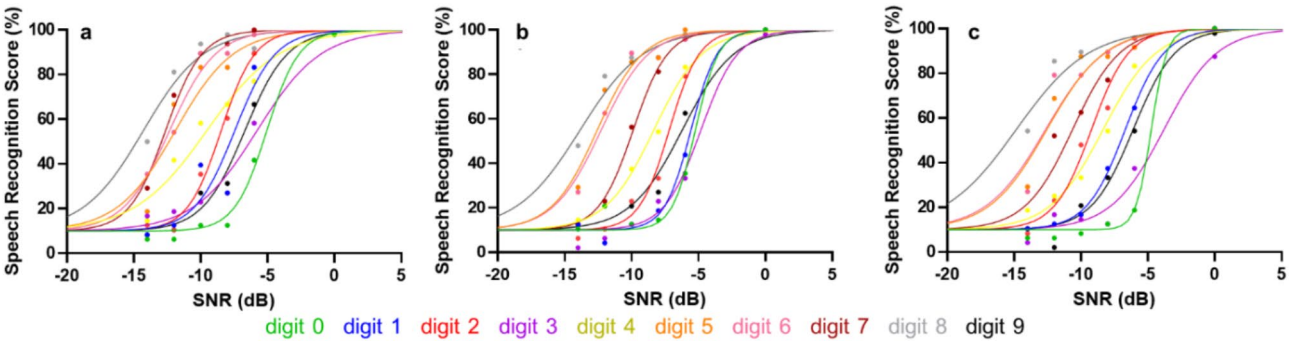

Fig. 3. Speech recognition functions (percent correct as a function of SNR) of the individual digits in noise for listeners with normal hearing. The digits were uttered in 3 voices: a natural male (a), a synthetic female (b) and a synthetic male (c) voice. Each dot shows the average score of 24 participants. The lines show the maximum likelihood fits to the raw data. The lower asymptote (guess rate) was set at 0.1.

SNR and 18.1 ± 6.0%-points/dB for the female synthetic voice, and − 8.9 ± 3.7 dB SNR and 16.6 ± 9.9%-points/dB for the male synthetic voice (Fig. 3). These mean SRTs were not significantly different (ANOVA, $F(2,27) = 0.144$, $p = 0.86$), and neither were the mean slopes (ANOVA, $F(2,27) = 0.51$, $p = 0.61$).

The level corrections per digit needed to achieve perceptual equivalence were calculated by subtracting the (voice-specific) mean SRT from the digit-specific SRT. Level corrections shifted the individual speech recognition functions horizontally so that all SRTs aligned on the average SRT (see Fig. 1). The level corrections of the natural male and synthetic male voice were strongly correlated (Pearson, $r = 0.96$, $p < 0.001$) (Fig. 4). There was also a strong correlation between level corrections of the natural male and synthetic female voice (Pearson, $r = 0.95$, $p < 0.001$). These level corrections, obtained from human listeners, serve as a benchmark for the ASR in Part III.

Discussion

Results showed no significant differences between natural and synthetic voices in mean DIN SRTs and slopes of the speech recognition functions, indicating similarity in digit intelligibility across these different Dutch voices. The variability in digit-specific speech recognition is also very similar between the natural and synthetic voices. These result are in line with the results from studies with natural and synthetic German Matrix sentences[27] and German everyday sentences[25]. In these studies they determined SRTs and slopes of speech recognition functions and showed that they compare very well between the natural and synthetic speech. In contrast, Polspoel, et al.[22] identified small but significant differences in the mean SRTs and slopes between natural and synthetic word lists, both of which were generated using the same synthetic male and female Dutch voices utilized in the present study.

Part III: determining level corrections from ASR

Part II demonstrated the traditional procedure to determine level corrections needed to perceptually equalize stimuli (i.e., digits) of a speech-in-noise test. The aim of part III was to determine whether it was possible to use ASR to estimate human speech recognition functions and the associated level corrections. ASR derived speech recognition functions and level corrections for Dutch natural and synthetic digits were compared to those obtained from human listeners in Part II (Fig. 1). If ASR derived level corrections closely resemble human derived level corrections, then ASR could potentially serve as a satisfactory alternative for conducting listening experiments in development of DIN tests.

Method

Cloud-based ASRs

First, three off-the-shelf cloud-based ASRs were assessed: Google Cloud ASR[35], Microsoft Azure Bing Speech API[36], and IBM Watson Speech to Text[37]. These systems are typically trained on extensive datasets, with Google Cloud's ASR, for instance, trained on millions of hours of audio and billions of text sentences, supporting over 110 languages. Utilizing advanced deep learning neural network algorithms, these systems convert speech to text

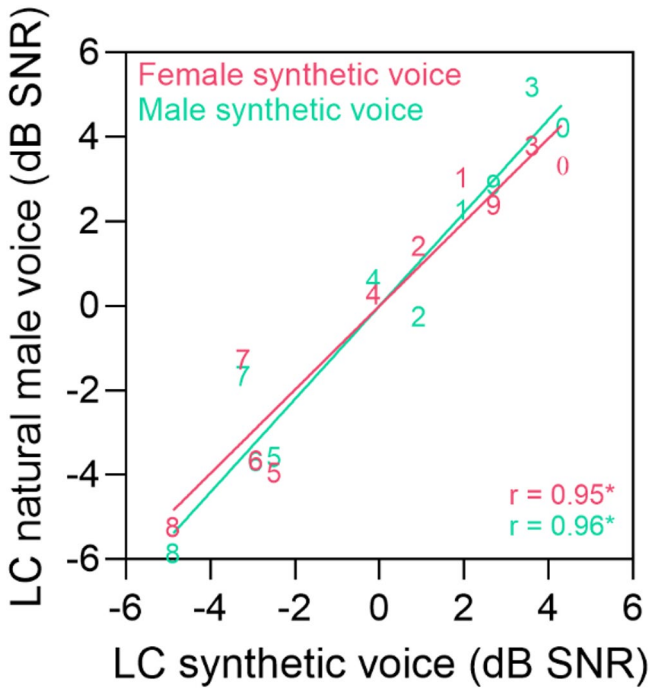

Fig. 4. Level corrections (LC) for the natural male digits as a function of the LC of the synthetic male digits (turquoise) and the the synthetic female digits (pink) based on a listening experiment with 24 human listeners.

and offer real-time processing and customization through APIs. Despite their similarities, these systems vary in target applications, data processing, model training, technology, training data, and supported languages.

The natural male and spectrum-adjusted synthetic female Dutch digits were each used to create 120 digit triplets. These digit triplets were mixed with corresponding noise from Part II across 24 SNRs (-4.5 to 7 dB SNR in 0.5 dB steps). The reason for using such small SNR increments was motivated by the discovery of a distinct tipping point where the digits were either recognized by the ASR or not. The range of SNRs was based on performance of the ASR system and chosen to cover the steep part of the speech recognition function. The digit triplets in noise were evaluated across all three ASRs via an API in Python, yielding scores ranging from 0 to 100% for most digits. The output was standardized to always consist of 3 characters, replacing non-digit responses with blanks and converting words to digits where applicable. Single-digit responses were correctly positioned, with the remaining characters as blanks, later substituted with random digits (0 to 9) to introduce a 10% guessing chance per digit presentation. Additional triplets with lower SNRs (-5 to +7.5 dB SNR in 0.5 dB steps) were created for the natural voice, specifically for Google Cloud and Azure ASR, due to overly high scores and lack of accurate speech recognition function estimations. For the analysis of digit-specific speech recognition functions, only SNRs yielding scores between 20 and 90% were considered because the SRT is the important parameter for calculation of level corrections and recognition probabilities did not always reach 100%. This is because the ASR systems are not familiar with the limited set of possible outcomes (i.e., digits 0 to 9).

FADE ASR

FADE stands for simulation Framework for Auditory Discrimination Experiments and is an ASR system based on Hidden Markov models that simulates the human speech recognition process[38–40]. FADE has demonstrated in previous research to accurately predict SRTs of various speech materials[41,42]. Unlike the off-the-shelf (standard) ASRs mentioned earlier, this particular system employs identical speech and noise stimuli for both training and testing purposes. As a result, it yields much lower SRTs that closely align with those of humans[39]. It uses front-end auditory spectro-temporal features for the initial processing of the audio signals before further analysis. It has been demonstrated that the Gabor filter bank (GBFB) features provide more accurate FADE predictions than standard Mel-Frequency Cepstral Coefficients (MFCC) features and more accurately resembles human auditory processing. FADE operates with a distinct training and testing phase. During the training phase, the system, utilizing hidden Markov models (HMMs), is exposed to labeled audio data to learn patterns and features associated with different stimuli. Techniques like feature extraction are used to capture relevant characteristics. Parameters are adjusted based on feedback from the training data to improve performance. In the testing phase, the trained system is evaluated on unseen data to assess its ability to discriminate between stimuli. In the datasets for training and testing, the 10 digits were mixed with the speech-shaped noise using random temporal offsets, such that a repetition of the same waveform was practically impossible[39]. The provided noise signal in FADE needed to be at least one minute long.

The FADE source code and instructions are available on GitHub[43] and were installed on a Linux OS (Ubuntu). Note that FADE works independently of languages and thus the same approach could be used for different language variants of the DIN test. The Dutch natural male and spectrum-adjusted synthetic female digits from Part II with the associated LTASS noise were used as input for FADE.

Initially, FADE was run with default parameters and trained on various SNRs (in 3 dB steps) using digit sets mixed with random fragments of speech-shaped noise. For each SNR, a separate ASR model was tested across SNRs, and the SRT was estimated. The optimal training SNR, yielding the lowest SRT, was determined to be -9 dB and was used for subsequent testing. To improve the precision of SRT estimation, the default step size was refined from 3 dB to 1 dB, with both training and test sample sizes set to 1000. The procedure was also repeated using the non-default MFCC front-end feature. However, as level corrections (LC) from GBFB correlated more strongly with human results, GBFB was selected as the standard, aligning with Schädler, et al.[44] who preferred GBFB over MFCC due to its enhanced robustness against additive noise. The output generated by the model at the optimal training SNR of -9 dB was subsequently used for comparison with human results. Binary test data (0 or 1) at each SNR indicated whether a digit was correctly recognized by the ASR. These data were used to estimate the average and digit-specific speech recognition functions by performing maximum likelihood fits. The guess rate, i.e., the probability of correct recognition of a digit at very low SNRs, differed between digits and runs but averaged 0.1 across the ten digits. These differences are probably inherent in training and testing with the FADE ASR system where no a priori knowledge is provided about the occurrence of the various stimuli. The SRT of each digit was obtained by taking the midpoint between the guess rate and 1 (i.e., the SRT of the guess rate corrected speech recognition function).

Then, level corrections were determined from these digit-specific speech recognition functions and applied to the digit wav-files. A new run through FADE was performed to generate new speech recognition functions. Then, the average and digit-specific speech recognition functions were determined again. This process was repeated until the slope of the average speech recognition function did not improve by more than 1%-point/dB. The summed level corrections from all runs yielded the final level corrections of the FADE system, which were then compared to the human-based corrections from Part II. The Root mean square (RMS) error of the level corrections was calculated by taking the square root of the average of the squared differences between the human-based and FADE ASR-based level corrections.

Results

Cloud-based ASRs

Figure 5 [b-d] shows average recognition scores and fitted speech recognition functions for three pre-trained cloud-based ASR systems. For comparison, speech recognition functions of the same synthetic female digits derived from human listeners (see Part II) are shown in Fig. 5 [a]. The ASR derived speech recognition functions

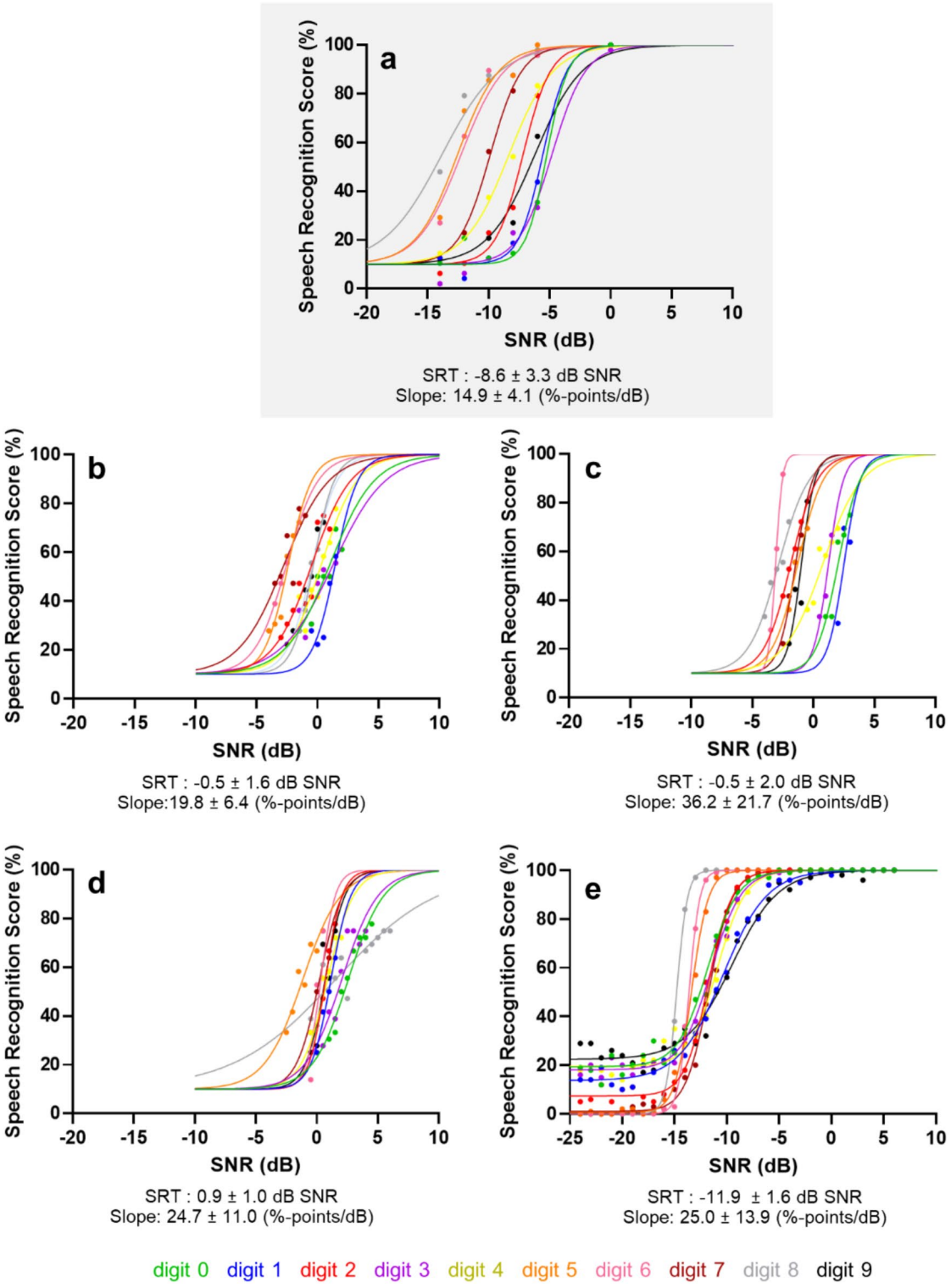

Fig. 5. Speech recognition functions of the synthetic female Dutch digits in noise based on three cloud-based ASRs. Reference human-derived speech recognition functions are presented in (**a**). Google Cloud ASR is presented in (**b**), Microsoft Azure Bing Speech API in (**c**), IBM Watson Speech to Text in (**d**), and FADE ASR (after the final run) in (**e**). The dots are the percentage correct scores based on 12 presentations per digit per SNR. The lines show the maximum likelihood fits to the raw data. The lower asymptote (guess rate) was set at 0.1 in a, b, c and d for all digits, and estimated from the data per digit in e.

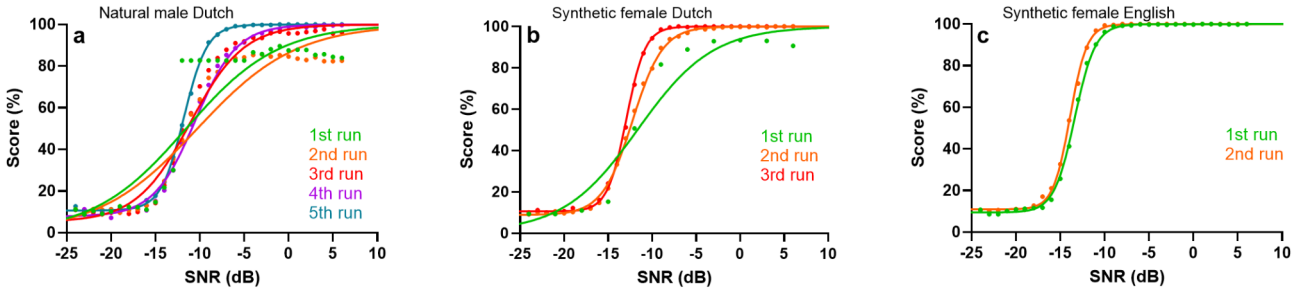

Fig. 6. Average speech recognition functions of digits based on FADE ASR. Results for natural male Dutch digits are presented in (**a**), synthetic female Dutch digits in (**b**) and synthetic female English digits in (**c**). The different functions are the result of the different runs through the system. After each run, level corrections were determined to achieve equal intelligibility and the level-corrected digits were again fed to the system until the slope stopped increasing more than 1%-point/dB.

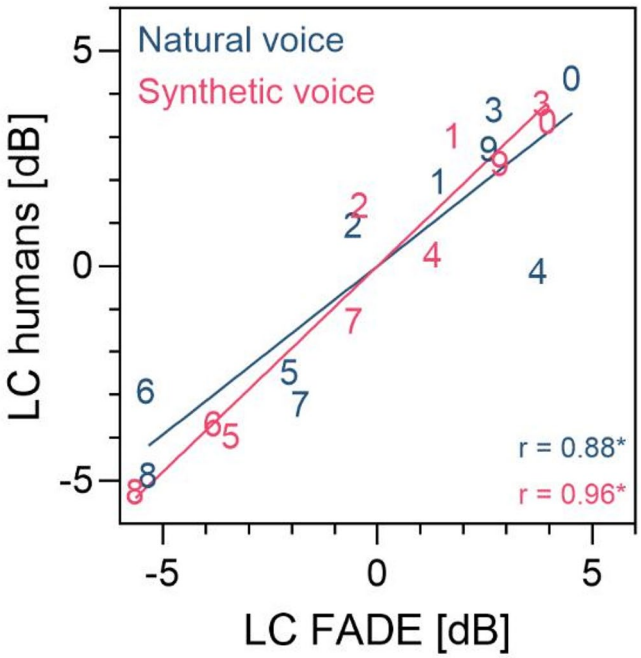

Fig. 7. Level corrections (LC) for natural (male voice) and synthetic (female voice) Dutch digits based on the listening experiment with human listeners from Part II as a function of the LC based on FADE ASR.

have significantly steeper slopes (one-way ANOVA, $F(3,36) = 4.002$, $p = 0.015$) and significantly higher SRTs (one-way ANOVA, $F(3,36) = 39.934$, $p < 0.001$) compared to the human derived speech recognition functions. The variance of the digit SRTs was significantly greater for humans compared to the ASRs (Levene's test, $F(3,36) = 7.465$, $p < 0.001$). This indicated that the digits were more equally intelligible for ASRs than for humans.

FADE ASR

There was a total of five runs for the Dutch natural digits and three runs for the Dutch synthetic digits through the FADE ASR before no further improvement in average speech recognition function was found. Figure 6 [a, b] shows the average speech recognition functions after the different runs for respectively the Dutch natural and synthetic digits. Figure 5 [e] shows the digit-specific speech recognition functions after the final run for the Dutch synthetic digits. It is evident that, as anticipated, the application of level corrections resulted in steeper speech recognition functions. Note that for the initial runs recognition probabilities sometimes did not reach 100%. See the discussion for further information.

The cumulative total of level corrections across various runs for each digit, meaning the level difference between the digits from the final set compared to the original digits, were compared with the reference human-derived level corrections determined in Part II. Figure 7 shows strong correlations between human and ASR-derived level corrections, for both natural (Pearson correlation, $r = 0.88$, $p < 0.001$) and synthetic ($r = 0.96$,

$p < 0.001$) digits. Root mean square (RMS) error of the level corrections was 1.6 dB for the natural voice and 0.9 dB for the synthetic voice.

Discussion

The notably higher SRTs and steeper slopes of cloud based off-the-shelf ASR-generated speech recognition functions, along with moderate correlations (Pearson, $0.54 \leq r \leq 0.79$) between the ASR-derived and human-derived level corrections (see Supplementary Fig. 1), suggested that these pre-trained ASRs were not suitable for Aladdin's intended purpose. This is likely because their design prioritized minimizing word error rate rather than mimicking human auditory behavior. The poorer performance of ASR systems compared to human performance on speech recognition tasks is usually referred to as the man-machine performance gap[45,46]. These ASR systems consist of an auditory model and a language model and were not specifically trained for the current task. Sometimes it is possible to train these systems for specific tasks or fine-tune certain parameters. It is likely that this will improve general performance (lower word error rates) but it is doubtful whether differences in recognition rate between digits will become more human-like. More recent ASR systems, such as Whisper, an end-to-end system, have already demonstrated better performance in speech recognition tasks compared to the systems we tested[47]. It is possible that these newer systems could provide more accurate estimations of the level corrections.

Results showed strong correlations between FADE-derived and human-derived level corrections We expect that one reason for the strong correlation between FADE-derived and human-derived level corrections is that digits represent a well-known and highly defined set of stimuli. FADE is trained on this set of digits and is thus optimized for discriminating between ten digits based on differences in spectro-temporal features. In human speech recognition, multiple factors often play a role; in addition to bottom-up factors, these may include top-down factors such as cognition, working memory, and the use of context. These factors are less important in closed-set speech materials, although depending on the size and content of the set, variables such as the number of response alternatives or the number of alternatives a listener considers may exceed the actual number of stimuli. However, for the DIN test stimulus set, this is unlikely to play a significant role, as digits 0 through 9 are learned from a very young age. It is therefore plausible that applying our method to other closed-set speech materials may yield less favorable results. Further research is needed to confirm this.

An unexpected observation when using FADE ASR to determine speech recognition functions was that, initially, during the first run, the average speech recognition function did not always reach 100% (see Fig. 6). This was found to be due to the heterogeneity of the digit stimuli. When recognition performance varied significantly between the digits, the trained FADE ASR model showed unexpected and unrealistic behavior. For example, the speech recognition function of one or more digits was not a continuously increasing (S-shaped) function but instead decreased with increasing SNR after reaching a maximum. After applying level corrections, the digit-specific speech recognition functions became more homogeneous, and eventually FADE ASR produced speech recognition functions that closely resembled those of human listeners.

Part IV: comparing Aladdin to original DIN test

The results of Part I and Part II demonstrated that the subjective quality of synthetic speech is high and that the speech recognition functions were comparable to those of natural speech. Thus, the synthetic speech material appeared suitable for speech-in-noise tests. The results of Part III showed that FADE clearly outperformed the three pre-trained ASRs and therefore we considered FADE ASR as the preferred technique to estimate the level corrections. In part IV we used FADE ASR for creating Aladdin tests and validated these tests and the entire Aladdin procedure. A Dutch and an English DIN test were automatically created, and the test-retest reliability and screening characteristics were compared with two standard (reference) tests. It was decided to create English and Dutch Aladdin tests because Dutch listeners generally have sufficient English proficiency to complete a DIN test[48], unlike with other languages such as French, Spanish, or Mandarin.

Method

Participants

Twenty-eight listeners with normal hearing (NH, $M = 25.5 \pm 4.5$ years) and 20 listeners with hearing loss (HL, $M = 59.2 \pm 14.2$ years) participated in the listening experiment. Pure-tone audiometry was conducted using a standard clinical audiometer (Decos audiology) and Sennheiser HDA200 headphones. Listeners with normal hearing had pure-tone thresholds of ≤ 20 dB HL across all octave frequencies from 250 Hz to 8 kHz in both ears and were recruited from the VU University campus. Listeners with hearing loss were recruited through flyers at the Amsterdam UMC ENT department and a post on a Facebook group for individuals with hearing loss. All participants were native Dutch speakers with at least basic knowledge of English, received compensation, and provided written informed consent. All experiments were performed according to the Netherlands Code of Conduct for Research Integrity. The VUmc medical research ethics committee approved the experimental protocol (protocol number: 2023.0202).

Material

Participants in the study performed four different DIN tests: a Dutch reference DIN test, an English reference DIN test, a Dutch Aladdin test, and an English Aladdin test. The reference Dutch DIN test was Smits, et al.[29] 's original DIN test, featuring natural male digits that had been level-corrected through a human listening experiment. The reference English DIN test was developed by Motlagh Zadeh, et al.[49], employing level-corrected natural female digits spoken by a Midwestern American English speaker. Both the Dutch and English Aladdin tests were developed using an identical procedure: first Google Cloud TTS was used to generate sets of synthetic

digits (see Part I for details and the voices used). Second, the synthetic digits were spectrum adjusted to match the average spectrum of the Dutch and English digits to the standard, idealized, speech spectrum (see Part II). Note that the average speech spectrum and the spectrum of the masking noise from the Dutch and English Aladdin tests are identical. Third, level corrections were derived from running FADE ASR as described in Part III. For the English digits two runs of FADE were needed to perceptually equalize the digits, see Fig. 6 [c]. Finally, for each language a set of 120 unique digit triplets was created by concatenating series of three digits.

Testing procedure

The study used a within-subject repeated measures design. Participants completed the DIN tests independently, seated in a sound-treated booth with a computer, without a test administrator present. Stimuli were presented diotically through Sennheiser HDA200 headphones connected to a Dell Optiplex780 PC via a Sound Blaster Creative soundcard (THX). The order of the four types of DIN tests was counterbalanced using a Latin square design across all participants. Each test involved presenting 24 digit triplets in a one-down, one-up adaptive manner, similar to the original DIN presentation method[3]. The SRT was determined as the average SNR of the last 20 out of 24 presentations. Participants performed each type of test twice to assess test-retest accuracy. Before actual testing started a practice DIN test was used to familiarize participants with the task. For listeners with normal hearing, the presentation level was fixed at 65 dB SPL. For those with hearing loss, the presentation level was adjusted individually by presenting digits in quiet until a comfortable level was determined. Thirteen of the twenty listeners with hearing loss required levels above 65 dB SPL, with an average level of 87.5 dB SPL.

Code availability

Statistical analysis were conducted with IBM SPSS Statistics, version 28.0.1.1 (15). Graphs were created with GraphPad Prism, version 10.2.0 (392). FADE software is available on GitHub (https://github.com/m-r-s/fade).

Results

Figure 8 shows the SRTs of the Aladdin test as a function of the SRTs of the reference DIN test in Dutch (left panel) and English (right panel). The circles and squares represent the test-retest SRT averages of the participants. There is a strong correlation between the Aladdin SRTs and the reference DIN SRTs for both the Dutch (Pearson correlation, $r = 0.91$, $p < 0.0001$) and English ($r = 0.95$, $p < 0.0001$) versions across all listeners.

Figure 9 illustrates the test-retest correlation of SRT scores for the four DIN tests. The SEM and SRTs for the listeners with normal hearing, the listeners with hearing loss and for the combined groups are presented in Table 2. The SEM represents the measurement error of the SRT estimate[16]. The SEMs remained generally small, indicating high precision of all DIN tests.

The screening characteristics of the Dutch and English reference DIN and Aladdin test were evaluated by assessing their ability to discriminate between participants with normal hearing and hearing loss (Fig. 10). There was a significant correlation between the SRT and PTA of the better ear for all four tests in the hearing loss group (all $p < 0.01$). The cutoff SRT was selected as the 95th percentile of the group of listeners with normal hearing. All four tests achieved perfect classification accuracy for all listeners with a hearing loss in the better ear (> 20 dB

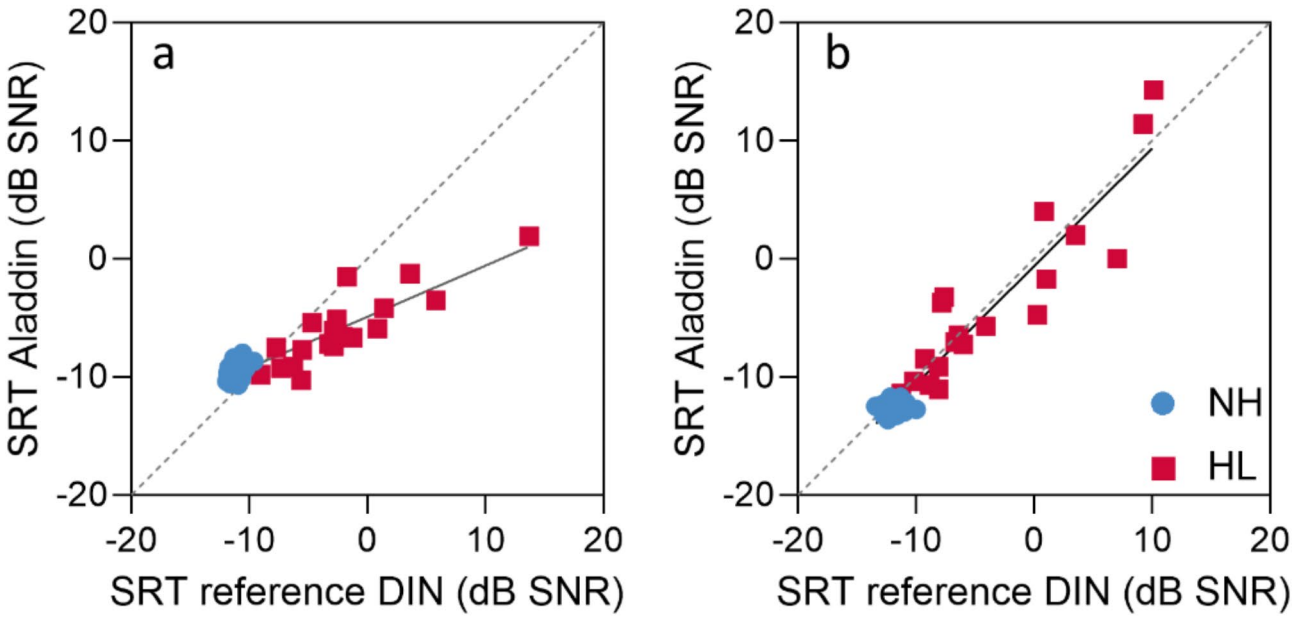

Fig. 8. Speech recognition thresholds (SRTs) of the Aladdin test as a function of the reference DIN test. Results for the Dutch tests are presented in (**a**) and results for the English tests in (**b**). The blue circles represent the listeners with normal hearing (NH), the red squares the listeners with hearing loss (HL). The regression concerns both groups combined.

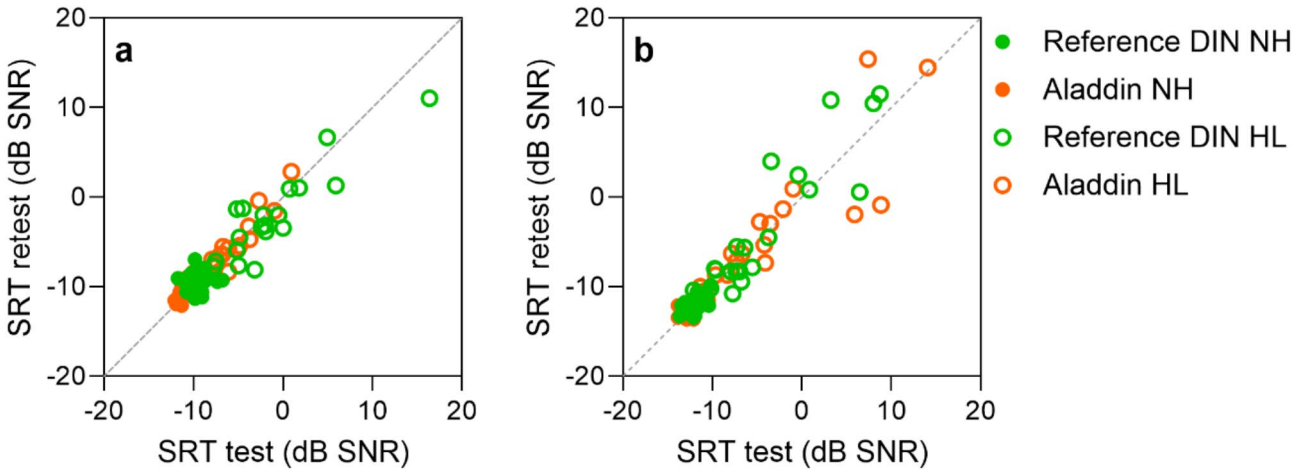

Fig. 9. DIN retest against DIN test SRTs. The reference DIN test datapoints are shown in green, the Aladdin test datapoints in orange. Results from Dutch DIN tests are presented in (**a**) and from English DIN tests in (**b**). The filled circles represent the scores of the normal hearing (NH) listeners, the unfilled circles of the listeners with hearing loss (HL).

		Reference DIN			Aladdin	
		SRT ± SD (dB SNR)	SEM (dB)		SRT ± SD (dB SNR)	SEM (dB)
Dutch	NH	-9.6 ± *0.7*	1.0		-11.1 ± *0.5*	0.4
	HL	-2.0 ± *5.2*	1.7		-6.0 ± *3.1*	0.8
	NH + HL	-6.4 ± *5.1*	1.4		-9.0 ± *6.3*	0.6
English	NH	-11.9 ± *0.7*	0.5		-12.6 ± *0.4*	0.5
	HL	-3.6 ± *6.7*	2.3		-4.0 ± *7.3*	2.5
	NH + HL	-8.4 ± *6.0*	1.5		-9.0 ± *3.2*	1.7

Table 2. SRTs and standard errors of measurement (SEMs) for the reference DIN and Aladdin tests in Dutch and english, shown for the normal hearing (NH) group, the hearing loss (HL) group, and the combined group (NH + HL).

HL). Note that three listeners in the group with hearing loss had one ear in the normal-hearing range, so they should not be detected as having hearing loss by the test. The Dutch and English reference DIN tests correctly classified 27 out of 31 listeners with normal hearing in the better ear. Both Aladdin tests correctly classified 26 out of 31 listeners with normal hearing in the better ear. This yields specificity and sensitivity of 84% and 100% for both Aladdin tests and 87% and 100% for both reference DIN tests.

General discussion

The Aladdin study successfully achieved its objective to generate reliable Dutch and English digits-in-noise (DIN) tests automatically. Utilizing synthetic speech, both Dutch and English versions of the Aladdin test were developed, with digits perceptually equalized through an Automatic Speech Recognition system (ASR). The resulting tests demonstrated similar test characteristics compared to the reference DIN test in their respective languages.

The Aladdin study consisted of four parts. Part I assessed the subjective quality of synthetic speech material in five languages. Results from a listening experiment with native speakers showed good to excellent MOS scores that did not differ significantly between synthetic and natural speech in any language across various speech materials (digits, words, sentences) and conditions (quiet and noise). These results, and the additional benefits of ease of use and affordability, advocate for the use of synthetic speech as the standard approach for creating speech tests.

Presently, Google Cloud TTS supports more than 220 voices across over 40 languages and variants, including various national accents for commonly spoken languages like English (American, British, South-African, Indian, Australian accents). It also includes less commonly spoken languages such as Icelandic, Serbian, Filipino and Hebrew. Although several languages spoken by ethnic groups in low-to middle-income countries are not yet covered by Google TTS, it is anticipated that more languages will be added in the future, as well as regional accents. Moreover, Google TTS already provides a Custom Voice feature, enabling users to train a personalized voice model using their own audio recordings, thereby creating a unique voice. AI advancements even allow for the creation of high-quality synthetic speech across multiple languages utilizing the same voice[50]. These aspects present a promising avenue for future research aimed at enhancing the Aladdin methodology, thereby

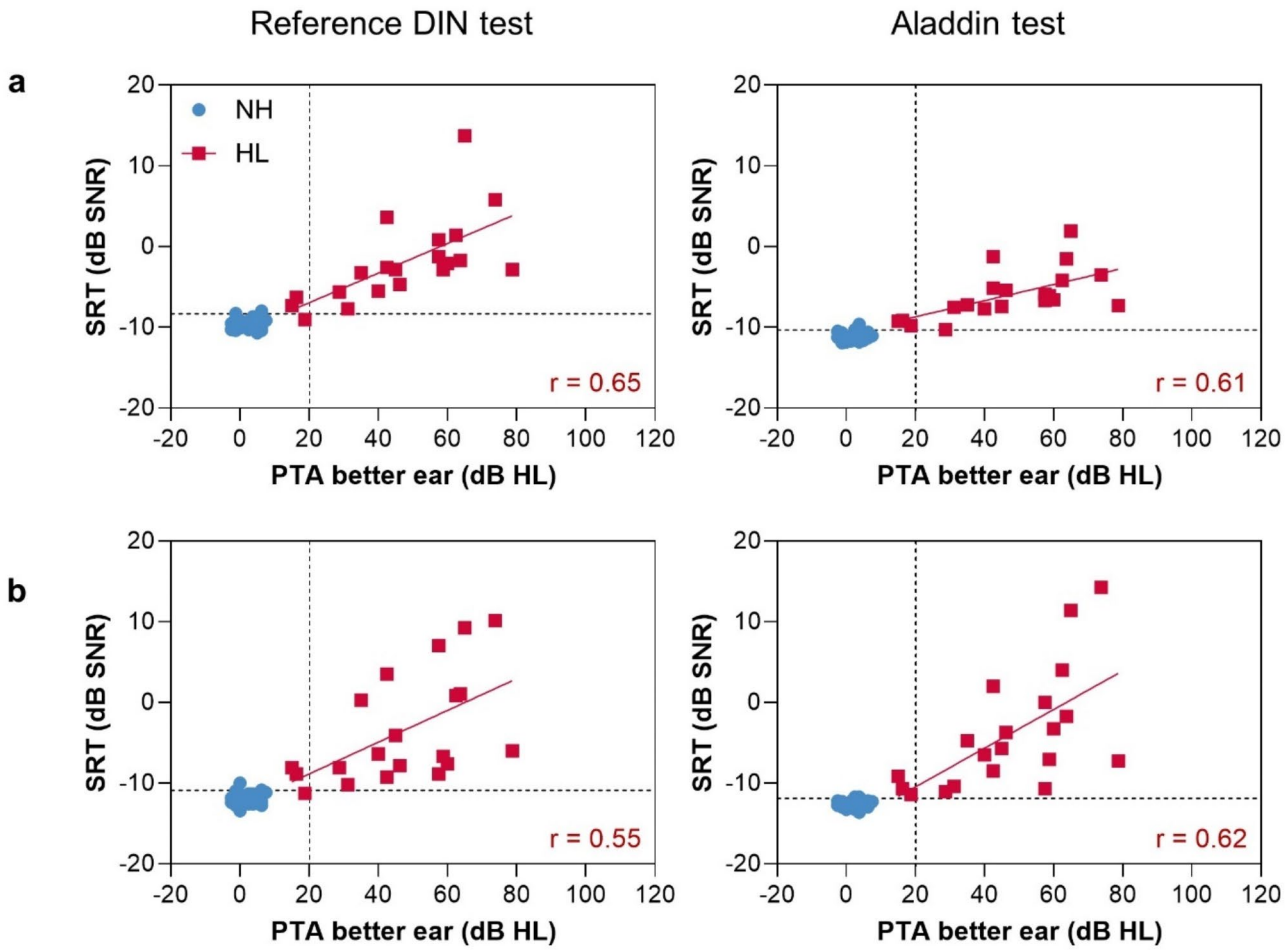

Fig. 10. Scatter plots of the DIN test SRTs as a function of the better ear PTA. The reference DIN tests are presented in the left panel, the Aladdin tests in right panel Dutch tests are presented in (**a**) and English tests in (**b**). Vertical dashed lines at 20 dB HL distinguish between listeners with normal hearing and listeners with hearing loss in the better ear. The horizontal dashed lines denote the SRT values corresponding to the pass/fail criteria for each respective test, determined by the 95th percentile (inclusive) of SRTs within the normal hearing group. The dashed lines divide the plots into quadrants: the lower left and upper right quadrant reflect the correctly classified cases (true negatives and true positives respectively), the lower right quadrant represents false negatives, and the upper left quadrant false positives. Note that three listeners in the hearing loss group had PTA in the better ear within normal limits.

facilitating greater comparability of DIN tests across languages and accents. A potential limitation of using synthetic speech is the bandwidth of the material. Analysis of the spectra of speech files created by the Google TTS engine shows that the upper limit is approximately 12 kHz. This is insufficient to capture all extended high frequencies (EHF, frequencies > 8 kHz). Additionally, it is unclear whether the EHF information in synthetic speech is comparable to that in natural speech. In the Aladdin procedure, we limit the bandwidth of the digit speech material to 9.5 kHz, so the limited bandwidth of the synthetic speech is not an important issue. In a study by Nuesse, et al.[27], the TTS system from the Acapela Group was identified as the best after evaluating a large number of TTS systems. The sample rate of the sound files from this TTS system is 22.1 kHz, meaning that no speech information above approximately 11 kHz will be present in the synthetic speech. Given the recent focus on EHF information in speech material[51,52], the bandwidth of the generated synthetic speech should be considered when selecting TTS systems or using synthetic speech for specific experiments.

In Part II comparisons were made between natural and synthetic voices by assessing speech recognition functions of individual digits presented at fixed SNRs to listeners with normal hearing. We did not find significant differences between natural and synthetic voices in mean DIN SRTs and slopes of the speech recognition functions. Strong correlations among the voices may suggest that the level corrections are robust to type of voice, however, this has to be confirmed for other languages. Further research is needed to assess whether speech recognition of synthetic digits in other languages (e.g., those from Part I) is comparable to that of natural digits. However, Dutch is a relatively less spoken language (approximately 23 million native speakers worldwide), and other more widely spoken languages usually have better coverage by TTS systems (more available voices in Google TTS), potentially resulting in synthetic speech that even more closely resembles natural speech.

The speech recognition functions were used to determine level corrections and to perceptually equalize the digits. The intelligibility of Dutch digits is highly variable (see Fig. 3), which means that the precision of a DIN test, where the digits are not perceptually equalized, is low. By applying the level corrections, the slope of the average speech recognition function increases because the digit-specific speech recognition functions align. The standard error of measurement (SEM) of an adaptive speech-in-noise test is inversely proportional to the slope of the average speech recognition function and the square root of the number of presentations[16]. For the natural digits in Fig. 3a, the slope of the average function is approximately 9%-points/dB, while the average of the slopes of the digit-specific speech recognition functions is about 16%-point/dB. This slope will be approximately equal to the slope of the average speech recognition function after applying the level corrections. Note that slopes of the digit triplet speech recognition functions constructed from these digit-specific speech recognition functions are even steeper than slopes of average digit speech recognition functions[17]. To achieve the same precision (SEM) with a DIN test where the digits are not perceptually equalized, approximately three times $[\left(\frac{16}{9}\right)^2]$ as many presentations are needed compared to a DIN test where this optimization has been applied.

It is common practice in the development of a speech-in-noise test to determine level corrections through a listening experiment with listeners with normal hearing and a single type of noise. The test is then optimized for individuals with normal hearing using that specific noise, but not necessarily for individuals with hearing loss, hearing aids, or cochlear implants. Additionally, the level corrections may differ when using other types of background noise. Determining level corrections is an intensive process, and therefore it is not always realistic to establish these level corrections separately for each noise type or user group. Moreover, the results of the tests would no longer be directly comparable. In general, these disadvantages do not outweigh the potential increase in test precision. Previous research has also shown that, for the digits used in the Dutch DIN test, the level corrections for participants with hearing loss and those with normal hearing are strongly correlated[17]. Similarly, the level corrections for digits presented in steady-state noise, 16 Hz interrupted noise, and 32 Hz interrupted noise were also strongly correlated[53].

In the present study, level corrections were primarily determined through listening experiments using monaural presentation of the stimuli. This choice was motivated by the fact that this was also the presentation method used during the development of the Dutch DIN test[3] and because the standard version of FADE ASR does not account for binaural aspects. However, the DIN test is also frequently presented diotically (Part IV) or in antiphasic presentation[48,54]. Diotic DIN SRTs are approximately 1 dB lower than monaural DIN SRTs due to binaural summation[48]. It is reasonable to assume that the effect of binaural summation on the intelligibility of digits in noise does not depend on the specific digit, as binaural summation is known to be independent of frequency[55]. Antiphasic presentation has been widely employed for screening purposes[54]. In the antiphasic DIN test, speech is presented out of phase, while noise is presented in phase. These antiphasic DIN tests are significantly more sensitive to unilateral and conductive hearing loss compared to diotic DIN tests[54]. Smits, et al.[48] found approximately 5 dB lower SRTs for the antiphasic DIN compared to the diotic DIN. This difference, known as the Binaural Intelligibility Level Difference (BILD), is comparable across other language variants of the DIN[54,56]. In contrast to binaural summation, the BILD for a specific digit presented in noise is likely dependent on the spectrum of that digit, as the difference between diotic and antiphasic detection of tones in noise is highly frequency-dependent[57] and binaural masking level differences for speech recognition in noise are dependent primarily on interaural phase differences in the frequencies below about 500 Hz[58]. This implies that for a set of digits with equal intelligibility in monaural or diotic presentation, digits with relatively more low-frequency energy are easier to understand in antiphasic presentation compared to digits with less low-frequency energy. A pilot experiment with 12 listeners with normal hearing (unpublished data) indeed showed differences in level corrections between diotic and antiphasic presented digits. In particular the Dutch digit 8 was much better intelligible in antiphasic presentation than in diotic presentation which can be explained by the relatively high level of low-frequency energy compared to the level of high-frequency energy for this digit. In summary, the level corrections derived from monaural or diotic presentations are not optimal for an antiphasic DIN test. However, for the reasons outlined above, it is valid to determine the level corrections for only one DIN test variant and to apply these without modification to the other variants. In most cases, monaural or diotic presentation will be chosen during development, but for specific screening purposes, antiphasic presentation may also be used.

In Part III, ASR systems were evaluated for generating speech recognition functions and determining level corrections for perceptually equalizing digit intelligibility. The ASR derived level corrections were compared to the human-derived level corrections of the natural and synthetic digits from Part II. Results showed that the pre-trained open-source ASRs produced SRTs that were significantly higher (poorer), and with steeper slopes, compared with human-based evaluations in Part II. Consequently, we utilized FADE, an ASR system better suited for mimicking human auditory behavior. FADE-derived level corrections demonstrated strong correlations with human-derived level corrections, particularly for synthetic digits. It has to be confirmed that the FADE ASR system provides accurate level corrections in other languages (e.g., tonal languages) as well. However, results from other studies[59,60] have demonstrated that FADE ASR predicts SRTs in Cantonese and Mandarin as accurately as in English, German, Polish, Russian and Spanish. Therefore, we believe it is a realistic expectation that our approach will also work for a large number of other languages.

Some DIN tests use babble noise[61,62], and FADE has been shown to successfully simulate the outcome of speech recognition for different noise maskers, including multitalker babble[63]. Overall, the purpose of FADE is to provide a structured framework for assessing and evaluating auditory discrimination abilities. It was originally designed to predict speech recognition thresholds of the German Matrix sentence test in noise[40], and was later extended to predict the outcomes of other speech recognition tests such as DIN and the Göttingen everyday sentences speech test (GÖSA)[41]. FADE uses no empirical reference and works in different languages and for various noise conditions[63]. FADE's predictions in auditory discrimination experiments aligned closely with

empirical data for listeners with and without hearing loss[38,64], consistent with our findings. To our knowledge, the specific application of determining level corrections for speech material items to achieve perceptual equivalence has not been one of the applications of FADE until now. Note that not all ASRs potentially suitable for the Aladdin procedure were considered in this study. For example, Kaldi, an ASR used for research and development in the field of speech recognition, could possibly also be employed[65,66]. As previously indicated, further research is required to determine whether FADE can also be applied to determine level corrections for speech materials from other speech in noise tests. It would also be interesting to investigate whether FADE ASR could be used to determine level corrections for antiphasic digit presentation. FADE has been used to incorporate some form of binaural hearing[60,67], but in that implementation, features from both left and right channels were provided as input to the model. This approach is not appropriate for the specific antiphasic listening condition. A front-end binaural model preceding the FADE model might provide a more suitable solution[68].

In Part IV, the Aladdin test, developed from ASR-based level-corrected synthetic digits, was validated against established reference DIN tests in Dutch and English. Dutch-speaking listeners with normal hearing and with hearing loss performed both reference and Aladdin tests in each language. Strong correlations between Aladdin and reference test SRTs were observed for both Dutch and English versions ($r = 0.91$ and $r = 0.95$, respectively). SEMs, indicating SRT estimation errors, were generally low across all tests. Both Aladdin tests achieved perfect classification accuracy for participants with hearing loss. The SEM and variance in SRTs were significantly greater among participants with hearing loss compared to those with normal hearing in the English DIN tests (Table 2). This difference is likely attributable to the older age and lower English proficiency of participants with hearing loss, compared to those with normal hearing. Additionally, SEMs are expected to be lower for listeners with normal hearing than for listeners with hearing loss because the speech recognition function is shallower for listeners with hearing loss[16,69]. Smits, et al.[48] found no significant differences in SRTs between Dutch and English (US) DIN tests among young adult Dutch listeners with normal hearing and Kaandorp, et al.[70] demonstrated that linguistic abilities have minimal impact on Dutch DIN test results for young adult university students, suggesting that a basic proficiency in the language is adequate for performing the test. While no significant linear correlation was found between English SRTs and age for either test, participants aged 60 and above displayed significantly greater SRT variability in our study (see Supplementary Fig. 2). Potgieter, et al.[71] observed that non-native English speakers with limited self-reported English proficiency performed significantly worse on the South African DIN test compared to both native speakers and non-native speakers with higher self-reported English proficiency. A limitation of our study is the sole assessment of English proficiency through a single yes-or-no question ("Do you have at least a basic understanding of the English language?"), masking variations in proficiency levels and inadvertently including participants with insufficient English proficiency. Ideally, native English participants would have been included in the study.

Significant correlations were observed between the PTA of the better ear and the DIN SRT across all four DIN tests in listeners with hearing loss (all $0.55 \leq r \leq 0.65$). This aligns with findings from De Sousa, et al.[54] who reported that, in a group of listeners with normal hearing or symmetric sensorineural hearing loss, the better ear PTA was significantly correlated with SRTs on the South African English diotic DIN test.

The Aladdin approach not only simplifies and shortens the test development procedure of a DIN test, it also provides a universal guideline for the development of comparable DIN tests in different languages. Currently, comparing DIN test results across languages is not straightforward due to the many variants of the test; including variations in speech spectrum, masking noise, number of trials, antiphasic versus diotic presentation, target audience and test procedures[4]. With the Aladdin procedure, many of these factors are kept constant. Importantly, the spectra and bandwidth of all Aladdin tests are identical. Thus, "Aladdin" provides a *language-independent* approach to developing DIN tests, potentially making it applicable even in low- and middle-income countries where limited resources and audiometric equipment hinder test development. However, more research is needed to create and validate Aladdin tests in more languages, as it is still dependent upon the available languages of the TTS system.

In conclusion, the present Aladdin project has demonstrated that valid Dutch and English DIN tests can be generated fully automatically in a uniform manner, without the need for expensive audiometric equipment, professional speakers, and extensive listening experiments. We expect that our approach will also work for other languages, although additional research is still required. The proposed procedure has a broader application in audiology beyond automating DIN test generation. Conceptually, its methodology could extend to the creation of different speech recognition tests utilizing closed-set stimuli. Examples of such tests are the Matrix sentence test[18,72] and Coordinate Response Measure[73] but further research is needed to determine whether this approach is indeed suitable for these types of tests.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 15 August 2024; Accepted: 27 March 2025
Published online: 15 April 2025

References

1. Carhart, R. Tests for selection of hearing aids. *Laryngoscope* **56**, 780–794 (1946).
2. Smits, C., De Sousa, K. C. & Swanepoel, W. An analytical method to convert between speech recognition thresholds and percentage-correct scores for speech-in-noise tests. *J. Acoust. Soc. Am.* **150**, 1321–1331. https://doi.org/10.1121/10.0005877 (2021).
3. Smits, C., Goverts, T. S. & Festen, J. M. The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *J. Acoust. Soc. Am.* **133**, 1693–1706. https://doi.org/10.1121/1.4789933 (2013).

4. Van den Borre, E., Denys, S., van Wieringen, A. & Wouters, J. The digit triplet test: A scoping review. *Int. J. Audiol.* **60**, 946–963 (2021).
5. Zokoll, M. A., Wagener, K. C., Brand, T., Buschermöhle, M. & Kollmeier, B. Internationally comparable screening tests for listening in noise in several European Languages: The German digit triplet test as an optimization prototype. *Int. J. Audiol.* **51**, 697–707. https://doi.org/10.3109/14992027.2012.690078 (2012).
6. Watson, C. S., Kidd, G. R., Miller, J. D., Smits, C. & Humes, L. E. Telephone screening tests for functionally impaired hearing: Current use in seven countries and development of a US version. *J. Am. Acad. Audiol.* **23**, 757–767. https://doi.org/10.3766/jaaa.23.10.2 (2012).
7. Potgieter, J. M., Swanepoel, W. & Smits, C. Evaluating a smartphone digits-in-noise test as part of the audiometric test battery. *S Afr. J. Commun. Disord.* **65**, e1–e6. https://doi.org/10.4102/sajcd.v65i1.574 (2018).
8. Wasmann, J. A., Huinck, W. J. & Lanting, C. P. Remote cochlear implant assessments: validity and stability in Self-Administered Smartphone-Based testing. *Ear Hear.* **45**, 239–249. https://doi.org/10.1097/aud.0000000000001422 (2024).
9. de Graaff, F. et al. Our experience with home self-assessment of speech recognition in the care pathway of 10 newly implanted adult cochlear implant users. *Clin. Otolaryngol.* **44**, 446–451. https://doi.org/10.1111/coa.13307 (2019).
10. Cullington, H. E. & Aidi, T. Is the digit triplet test an effective and acceptable way to assess speech recognition in adults using cochlear implants in a home environment? *Cochlear Implants Int.* **18**, 97–105. https://doi.org/10.1080/14670100.2016.1273435 (2017).
11. Smits, C., Kapteyn, T. S. & Houtgast, T. Development and validation of an automatic speech-in-noise screening test by telephone. *Int. J. Audiol.* **43**, 15–28. https://doi.org/10.1080/14992020400050004 (2004).
12. Culling, J. F., Zhao, F. & Stephens, D. The viability of speech-in-noise audiometric screening using domestic audio equipment. *Int. J. Audiol.* **44**, 691–700. https://doi.org/10.1080/14992020500267017 (2005).
13. Smits, C. & Houtgast, T. Results from the Dutch speech-in-noise screening test by telephone. *Ear Hear.* **26**(1), 89–95 (2005).
14. De Sousa, K. C. et al. Global use and outcomes of the hearwho mHealth hearing test. *Digit. Health.* **8**, 20552076221113204. https://doi.org/10.1177/20552076221113204 (2022).
15. Hearing screening: considerations for implementation. Licence: CC BY-NC-SA 3.0 IGO (World Health Organization, Geneva, 2021).
16. Smits, C., Festen, J. M., Swanepoel, D. W., Moore, D. R. & Dillon, H. The one-up one-down adaptive (staircase) procedure in speech-in-noise testing: standard error of measurement and fluctuations in the tracka). *J. Acoust. Soc. Am.* **152**, 2357–2368. https://doi.org/10.1121/10.0014898 (2022).
17. Smits, C. & Houtgast, T. Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests. *J. Acoust. Soc. Am.* **120**, 1608–1621. https://doi.org/10.1121/1.2221405 (2006).
18. Kollmeier, B. et al. The multilingual matrix test: principles, applications, and comparison across Languages: A review. *Int. J. Audiol.* **54 Suppl 2**, 3–16. https://doi.org/10.3109/14992027.2015.1020971 (2015).
19. International Telecommunication Union. *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm (ITU-T Recommendation P.835)* (ITU-T, 2003).
20. Akeroyd, M. A. et al. International collegium of rehabilitative audiology (ICRA) recommendations for the construction of multilingual speech tests. ICRA working group on multilingual speech tests. *Int. J. Audiol.* **54** (Suppl 2), 17–22. https://doi.org/10.3109/14992027.2015.1030513 (2015).
21. Ning, Y., He, S., Wu, Z., Xing, C. & Zhang, L. J. A review of deep learning based speech synthesis. *Appl. Sci.* **9**, 4050. https://doi.org/10.3390/app9194050 (2019).
22. Polspoel, S., Holtrop, F. S., Bosman, A. J., Kramer, S. E. & Smits, C. Measurement and optimisation of the perceptual equivalence of the Dutch consonant-vowel-consonant (CVC) word lists using synthetic speech and list pairs. *Int. J. Audiol.* **64**(1), 35–42. https://doi.org/10.1080/14992027.2024.2306186 (2025).
23. Génin, A. et al. Development and validation of a French speech-in-noise self-test using synthetic voice in an adult population. *Front. Audiol. Otol.* **2** https://doi.org/10.3389/fauot.2024.1292949 (2024).
24. Herrmann, B. The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults. *Int. J. Speech Technol.* **26**, 1–21. https://doi.org/10.1007/s10772-023-10027-y (2023).
25. Ibelings, S., Brand, T. & Holube, I. Speech recognition and listening effort of meaningful sentences using synthetic speech. *Trends Hear.* **26**, 233121652211306. https://doi.org/10.1177/23312165221130656 (2022).
26. Kropp, M. H., Hocke, T., Agha-Mir-Salim, P. & Müller, A. Evaluation of a synthetic version of the digits-in-noise test and its characteristics in CI recipients. *Int. J. Audiol.* **60**, 507–513. https://doi.org/10.1080/14992027.2020.1839678 (2021).
27. Nuesse, T., Wiercinski, B., Brand, T. & Holube, I. Measuring speech recognition with a matrix test using synthetic speech. *Trends Hear.* **23**, 2331216519862982. https://doi.org/10.1177/2331216519862982 (2019).
28. Google *Google Cloud Text-to-Speech*, (2021).
29. Smits, C., Theo Goverts, S. & Festen, J. M. The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *J. Acoust. Soc. Am.* **133**, 1693–1706. https://doi.org/10.1121/1.4789933 (2013).
30. Bosman, A. J. & Smoorenburg, G. F. Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment. *Audiology* **34**, 260–284 (1995).
31. Versfeld, N. J., Daalder, L., Festen, J. M. & Houtgast, T. Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *J. Acoust. Soc. Am.* **107**, 1671–1684. https://doi.org/10.1121/1.428451 (2000).
32. Winn, M. *Praat vocoder script*, https://github.com/ListenLab/Vocoder/blob/main/praat_vocoder.txt (n.d.).
33. Boersma, P. & Weenink, D. PRAAT: Doing phonetics by computer (Version 5.3.51). (2007).
34. Viswanathan, M. & Viswanathan, M. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Comput. Speech Lang.* **19**, 55–83. https://doi.org/10.1016/j.csl.2003.12.001 (2005).
35. Google *Google Cloud Speech-to-Text*, (2022).
36. Microsoft.Corporation. Microsoft Azure Bing Speech. (2022).
37. IBM.Corporation. IBM Watson Speech to Text. (2022).
38. Schädler, M. R., Warzybok, A. & Kollmeier, B. Objective prediction of hearing aid benefit across listener groups using machine learning: speech recognition performance with binaural Noise-Reduction algorithms. *Trends Hear.* **22**, 2331216518768954. https://doi.org/10.1177/2331216518768954 (2018).
39. Schädler, M. R., Warzybok, A., Ewert, S. D. & Kollmeier, B. A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception. *J. Acoust. Soc. Am.* **139**, 2708. https://doi.org/10.1121/1.4948772 (2016).
40. Schädler, M. R., Warzybok, A., Hochmuth, S. & Kollmeier, B. Matrix sentence intelligibility prediction using an automatic speech recognition system. *Int. J. Audiol.* **54 Suppl 2**, 100–107. https://doi.org/10.3109/14992027.2015.1061708 (2015).
41. Huelsmeier, D., Warzybok, A. & Schaedler, M. R. Extension of the framework for auditory discrimination experiments (FADE) to predict the goettingen (everyday) sentence speech test. *Speech Communication; 13th ITG-Symposium* (2018).
42. Schädler, M. Optimization and Evaluation of an Intelligibility-Improving Signal Processing Approach (IISPA) for the Hurricane Challenge 2.0 with FADE. *Interspeech*. 1331-1335 (2020).
43. Schädler, M. R. FADE - A Simulation framework for auditory discrimination experiments. https://doi.org/10.5281/zenodo.3734164.
44. Schädler, M., Meyer, B. & Kollmeier, B. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *J. Acoust. Soc. Am.* **131**, 4134–4151. https://doi.org/10.1121/1.3699200 (2012).

45. Jürgens, T., Brand, T. & Kollmeier, B. Modelling the human-machine gap in speech reception: microscopic speech intelligibility prediction for normal-hearing subjects with an auditory model. *Interspeech*. 410–413 (2007).
46. Jürgens, T. & Brand, T. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *J. Acoust. Soc. Am.* **126**, 2635–2648 (2009).
47. Slaney, M. & Fitzgerald, M. B. Comparing human and machine speech recognition in noise with QuickSIN. *JASA Express Lett.* **4** https://doi.org/10.1121/10.0028612 (2024).
48. Smits, C., Watson, C., Kidd, G., Moore, D. & Goverts, T. A comparison between the Dutch and American-English digits-in-noise (DIN) tests in normal-hearing listeners. *Int. J. Audiol.* **55**, 1–8. https://doi.org/10.3109/14992027.2015.1137362 (2016).
49. Motlagh Zadeh, L. et al. Extended high-frequency hearing enhances speech perception in noise. *Proc. Natl. Acad. Sci. U S A.* **116**, 23753–23759. https://doi.org/10.1073/pnas.1903315116 (2019).
50. Zhang, J. X., Ling, Z. H. & Dai, L. R. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 540–552 (2019).
51. Delaram, V. et al. Gender and speech material effects on the long-term average speech spectrum, including at extended high frequencies. *J. Acoust. Soc. Am.* **156**, 3056–3066. https://doi.org/10.1121/10.0034231 (2024).
52. Polspoel, S., Kramer, S. E., van Dijk, B. & Smits, C. The importance of extended High-Frequency speech information in the recognition of digits, words, and sentences in quiet and noise. *Ear Hear.* **43**, 913–920. https://doi.org/10.1097/aud.0000000000001142 (2022).
53. Smits, C. & Houtgast, T. Recognition of digits in different types of noise by normal-hearing and hearing-impaired listeners. *Int. J. Audiol.* **46**, 134–144. https://doi.org/10.1080/14992020601102170 (2007).
54. De Sousa, K. C., Swanepoel, W., Moore, D. R., Myburgh, H. C. & Smits, C. Improving sensitivity of the digits-in-noise test using antiphasic stimuli. *Ear Hear.* **41**, 442–450. https://doi.org/10.1097/aud.0000000000000775 (2020).
55. Hirsh, I. J. Binaural summation; a century of investigation. *Psychol. Bull.* **45**, 193–206. https://doi.org/10.1037/h0059461 (1948).
56. Ceccato, J. C. et al. French version of the antiphasic Digits-in-Noise test for smartphone hearing screening. *Front. Public. Health*. **9**, 725080. https://doi.org/10.3389/fpubh.2021.725080 (2021).
57. Durlach, N. I. Equalization and cancellation theory of binaural masking-level differences. *J. Acoust. Soc. Am.* **35**, 1206–1218 (1963).
58. Levitt, H. & Rabiner, L. R. Binaural release from masking for speech and gain in intelligibility. *J. Acoust. Soc. Am.* **42**, 601–608. https://doi.org/10.1121/1.1910629 (1967).
59. Scharf, M., Hochmuth, S., Wong, L., Kollmeier, B. & Warzybok, A. Lombard Effect for Bilingual Speakers in Cantonese and English: importance of spectro-temporal features. *Interspeech* 1377–1381 (2022).
60. Hülsmeier, D., Schädler, M. R. & Kollmeier, B. D. A. R. F. A data-reduced FADE version for simulations of speech recognition thresholds with real hearing aids. *Hear. Res.* **404**, 108217. https://doi.org/10.1016/j.heares.2021.108217 (2021).
61. Wilson, R. H., Burks, C. A. & Weakley, D. G. Word recognition of digit triplets and monosyllabic words in multitalker babble by listeners with sensorineural hearing loss. *J. Am. Acad. Audiol.* **17**, 385–397. https://doi.org/10.3766/jaaa.17.6.2 (2006).
62. Moore, D. R. et al. FreeHear: A new Sound-Field Speech-in-Babble hearing assessment tool. *Trends Hear.* **23**, 2331216519872378. https://doi.org/10.1177/2331216519872378 (2019).
63. Schädler, M. R., Hülsmeier, D., Warzybok, A., Hochmuth, S. & Kollmeier B. Microscopic multilingual Matrix test predictions using an ASR-based speech recognition model. *Interspeech*. 610–614 (2016).
64. Kollmeier, B., Schädler, M. R., Warzybok, A., Meyer, B. T. & Brand, T. Sentence recognition prediction for Hearing-impaired listeners in stationary and fluctuation noise with FADE: empowering the Attenuation and distortion concept by Plomp with a quantitative processing model. *Trends Hear.* **20** https://doi.org/10.1177/2331216516655795 (2016).
65. Povey, D. et al. The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (2011).
66. Araiza-Illan, G., Meyer, L., Truong, K. P. & Başkent, D. Automated speech audiometry: can it work using Open-Source Pre-Trained Kaldi-NL automatic speech recognition?? *Trends Hear.* **28**, 23312165241229057. https://doi.org/10.1177/23312165241229057 (2024).
67. Zedan, A., Jürgens, T., Williges, B., Hülsmeier, D. & Kollmeier, B. Modelling speech reception thresholds and their improvements due to Spatial noise reduction algorithms in bimodal cochlear implant users. *Hear. Res.* **420**, 108507. https://doi.org/10.1016/j.heares.2022.108507 (2022).
68. Hauth, C. F., Berning, S. C., Kollmeier, B. & Brand, T. Modeling binaural unmasking of speech using a blind binaural processing stage. *Trends Hear.* **24**, 2331216520975630. https://doi.org/10.1177/2331216520975630 (2020).
69. Smits, C. & Festen, J. M. The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: Steady-state noise. *J. Acoust. Soc. Am.* **130**, 2987–2998. https://doi.org/10.1121/1.3644909 (2011).
70. Kaandorp, M. W., De Groot, A. M. B., Festen, J. M., Smits, C. & Goverts, S. T. The influence of lexical-access ability and vocabulary knowledge on measures of speech recognition in noise. *Int. J. Audiol.* **55**, 157–167. https://doi.org/10.3109/14992027.2015.1104735 (2016).
71. Potgieter, J. M., Swanepoel, W., Myburgh, H. C. & Smits, C. The South African english smartphone Digits-in-Noise hearing test: effect of age, hearing loss, and speaking competence. *Ear Hear.* **39**, 656–663. https://doi.org/10.1097/aud.0000000000000522 (2018).
72. Houben, R. et al. Development of a Dutch matrix sentence test to assess speech intelligibility in noise. *Int. J. Audiol.* **53**, 760–763. https://doi.org/10.3109/14992027.2014.920111 (2014).
73. Bolia, R. S., Nelson, W. T., Ericson, M. A. & Simpson, B. D. A speech corpus for multitalker communications research. *J. Acoust. Soc. Am.* **107**, 1065–1066. https://doi.org/10.1121/1.428288 (2000).
74. Tillman, T. W. & Carhart, R. *An Expanded Test for Speech Discrimination Utilizing CNC Monosyllabic Words: Northwestern University Auditory Test No. 6* (USAF School of Aerospace Medicine Brooks Air Force Base, 1966).
75. Fournier, J. E. Audiométrie vocale: les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités. Maloine (1951).
76. Fei, J., Aiting, C., Yang, Z., Xin, X. & Dongyi, H. Development of a script of phonemically balanced monosyllable lists of Mandarin–Chinese. *J. Otology*. **5**, 8–19. https://doi.org/10.1016/S1672-2930(10)50003-5 (2010).
77. Nilsson, M., Soli, S. D. & Sullivan, J. A. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.* **95**, 1085–1099. https://doi.org/10.1121/1.408469 (1994).
78. James, C. J. et al. The French MBAA2 sentence recognition in noise test for cochlear implant users. *Int. J. Audiol.* **62**, 304–311. https://doi.org/10.1080/14992027.2022.2045368 (2023).
79. Aubanel, V., Lecumberri, M. L. G. & Cooke, M. The Sharvard corpus: A phonemically-balanced Spanish sentence resource for audiology. *Int. J. Audiol.* **53**, 633–638. https://doi.org/10.3109/14992027.2014.907507 (2014).
80. Xi, X. et al. Development of a corpus of Mandarin sentences in babble with homogeneity optimized via psychometric evaluation. *Int. J. Audiol.* **51**, 399–404. https://doi.org/10.3109/14992027.2011.642011 (2012).

Acknowledgements

We acknowledge the contributions of the following individuals to our research: Hans van Beek for software development; Filiep Vanpoucke for facilitating outreach via Cochlear, and all international contacts for participant recruitment in our multilingual experiment; and Marc Schädler for the generous assistance in setting up

FADE. David Moore was co-funded by the NIHR Manchester Biomedical Research Centre (NIHR203308). We acknowledge the hearWHO language development partners for using the original DIN recordings in Part I. This collaboration project is co-funded by PPP Allowance awarded by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.

Author contributions

C.S. acquired funding, conceived the project, and designed the experiments. S.P. performed the experiments, analyzed the data and wrote the manuscript with input from all authors. C.S., D. R. M., D. S. and S.E.K. substantially revised the manuscript. All authors have approved the submitted version.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-96312-z.

Correspondence and requests for materials should be addressed to C.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025