Applications of Automatic Speech Recognition and Text-To-Speech Technologies for Hearing Assessment: a Scoping Review

Mohsen Fatehifar, Josef Schlittenlacher, Ibrahim Almufarrij, David Wong, Tim Cootes and Kevin J. Munro

I. ABSTRACT

Objective: Exploring applications of automatic speech recognition and text-to-speech technologies in hearing assessment and evaluations of hearing aids.

Design: Research protocol was registered at the INPLASY database and was performed following the PRISMA scoping review guidelines. A search in ten databases was conducted in January 2023 and updated in June 2024.

Study sample: Studies that used automatic speech recognition or text-to-speech to assess measures of hearing ability (e.g. speech reception threshold), or to configure hearing aids were retrieved. Of the 2942 records found, 28 met the inclusion criteria.

Results: The results indicated that text-to-speech could effectively replace recorded stimuli in speech intelligibility tests, requiring less effort for experimenters, without negatively impacting outcomes (n=5). Automatic speech recognition captured verbal responses accurately, allowing for reliable speech reception threshold measurements without human supervision (n=7). Moreover, automatic speech recognition was employed to simulate participants' hearing, with high correlations between simulated and empirical data (n=14). Finally, automatic speech recognition was used to optimise hearing aid configurations, leading to higher speech intelligibility for wearers compared to the original configuration (n=3).

Conclusions: There is the potential for automatic speech recognition and text-to-speech systems to enhance accessibility of, and efficiency in, hearing assessments, offering unsupervised testing options, and facilitating hearing aid personalisation.

This work was supported by the MRC-DTP under Grant number MR/W007428/1. Kevin J. Munro is supported by the NIHR Manchester Biomedical Research Centre.

Mohsen Fatehifar (Corresponding author) is with Manchester Centre for Audiology and Deafness (ManCAD), University of Manchester (e-mail: mohsen.fatehifar@manchester.ac.uk)

Kevin J. Munro is with Manchester Centre for Audiology and Deafness (ManCAD), University of Manchester and Manchester Academic Health Science Centre, Manchester University Hospitals NHS Foundation Trust, UK (e-mail: kevin.j.munro@manchester.ac.uk)

Josef Schlittenlacher is with the Department of Speech, Hearing and Phonetic Sciences, University College London (e-mail: j.schlittenlacher@ucl.ac.uk)

Ibrahim Almufarrij is with the Manchester Centre for Audiology and Deafness (ManCAD), University of Manchester and Department of Rehabilitation Sciences, College of Applied Medical Sciences, King Saud University, (e-mail: ialmufarrij@ksu.edu.sa)

David Wong is with the Leeds Institute of Health Sciences, University of Leeds (e-mail: d.c.wong@leeds.ac.uk)

Tim Cootes is with the Centre for Imaging Sciences, University of Manchester(e-mail: timothy.f.cootes@manchester.ac.uk)

II. KEYWORDS

Automatic Speech Recognition, Hearing Assessment, Hearing Test, Hearing Aid, Hearing in Noise Test, Speech in Noise, Text To Speech.

III. INTRODUCTION

According to the World Health Organisation [1], 1.5 billion individuals have hearing loss, with 430 million requiring intervention, the most common of which is the provision of hearing aids [2]. It is estimated that by 2050, this number will increase to 2.5 billion with 700 million people requiring intervention. Hearing loss can significantly impact an individual's ability to communicate with ease, leading to stress, anxiety, isolation, depression, and a decline in quality of life. In addition, hearing loss is associated with a variety of long-term health conditions, including dementia [3].

Hearing assessments are typically performed in hospitals and clinics with specialised equipment and professionally qualified staff. However, these are not always available e.g., in developing countries [4]. Even in developed nations, access to these facilities can prove challenging especially in rural areas, and for elderly or infirm individuals [5]. Additionally, in a place where good quality services are readily available, a crisis such as the COVID-19 pandemic can change the situation and make it challenging for people to seek help [6], [7].

In healthcare systems where hearing assessments are available, there can be long waiting times associated with the prescription and fitting of hearing aids. This problem is compounded by the fact that an individual might need to visit the audiologist multiple times to obtain a properly prescribed and fitted hearing aid. Multiple fitting sessions can cause learning effects and fatigue, which obscure the results and make it hard to achieve the best configuration for the patient [8], [9]. Research shows that most people don't come forward to be assessed for hearing aids. Additionally, about half of hearing aid users do not wear them often [10] and poorly fit hearing aid is one of the factors contributing to this problem [11].

One way to address these issues is to develop methods that make hearing and hearing aid assessments easier. Recent

TABLE I
DEFINITION OF TERMS USED IN THIS DOCUMENT

Term	Definitions
Automatic Speech Recognition (ASR)	ASR is a technology that converts spoken language into text or into a representation that helps other machine learning models to make sense of it (e.g. a feature embedding vector).
Text-To-Speech (TTS)	TTS generates a speech signal from the text. In this system, the input is the words, and the output is the audio representing the input words.
Adaptive speech-in-noise Test (SIN)	This is a test of hearing disability that presents speech stimuli in the presence of background noise and aims to measure the highest level of noise (or lowest level of speech) before the speech becomes unintelligible to the participant. In a clinical setting, participants listen to the stimuli and are asked to repeat the words that they were able to understand. A human supervisor then evaluates their response to determine how much of the sentence the participant understood [20].
Signal to Noise Ratio (SNR)	SNR is a measure of the intensity of a signal relative to the intensity of background noise and is measured in decibels (dB).
Speech Reception Threshold (SRT)	SRT is the SNR at which an individual can understand and repeat back spoken words or sentences at criterion performance e.g., 50% correct. A lower SRT indicates a better performance.
Bias	The systematic difference between the measurements obtained from a model and the reference values. (e.g., a hearing test with a bias of +0.5 means that the measured SRT is on average 0.5 higher than the SRT of a clinical test).

progress has been made on this topic [12] and there is evidence that these methods can produce accurate outcomes without human supervision [13].

Advances in machine learning (ML) have enabled intelligent systems to replace humans in various industries. The healthcare industry has also been significantly impacted by the widespread usage of ML [14]. Audiology is no exception, and there have been attempts to leverage ML techniques in various stages of hearing assessment, including:

- Models that can predict optimal stimuli to make the hearing test faster and more accurate [15].
- Classifiers that can detect the type and degree of hearing loss from clinical hearing tests (pure tone audiograms) [16].
- Analysing EEG response signals to acoustic stimuli to detect if the person heard the stimuli [17].
- Self-administered speech intelligibility tests using automatic speech recognition (ASR) or text-to-speech techniques (TTS) [18], [19].

The use of ASR and TTS (see Table I for definitions) can make hearing tests more accessible. For example, ASR can record a person's response to auditory stimuli easily and naturally, which is particularly important for some people who have difficulties using a graphical user interface. Furthermore, using a reliable ASR system might reduce mishearing, miscategorising and other possible human errors. Additionally, TTS enables a flexible generation of natural stimuli (speech) in a controlled manner, which may be both more engaging and ecologically valid than pure tones or other artificial sounds. In the long term, ASR and TTS have the potential to create speech intelligibility tests that mimic a natural conversation, which can be conducted remotely without any specialist equipment.

A. Gap in knowledge

Previous reviews on remote and self-supervised hearing tests have focused on the general use of ML and automated hearing evaluation. Wasmann et al. [21] reviewed automated assessments of hearing but no studies that used ASR or TTS were investigated in their study. Osman [17] conducted a review on the use of ML for the detection of hearing loss, but the scope was limited to detecting hearing loss based on the classification of the auditory brainstem response (an electrophysiological measure of hearing). Almufarrij et al. [13] reviewed remote and self-supervised hearing test tools without focusing on the use of ML.

These studies were not specific to ASR and TTS and did not include all the papers that used these two technologies. There is a need for a scoping review of studies that specifically used ASR or TTS models for hearing tests. Therefore, the aim of this scoping review was to summarise and organise the existing work in this area to provide an overview of the latest advancements in the use of ASR and TTS for the assessment of both hearing and hearing aid fitting. Doing so will identify gaps in previous literature, which will facilitate future research in this domain.

IV. METHOD

The protocol (inplasy.com/inplasy-2023-1-0029) was submitted to the International Platform of Registered Systematic Review and Meta-Analysis Protocols [22] and the review was carried out in accordance with PRISMA scoping review guidelines [23].

A. Eligibility Criteria

This review considered studies that employed ASR or TTS in any aspect of hearing assessment and hearing aid fitting, regardless of whether the methods were conducted remotely or in a controlled setting. The review included theses, conference papers, peer-reviewed papers, book chapters, and preprints. See Table II for the complete inclusion and exclusion criteria.

B. Information Sources

Relevant studies were identified through a systematic literature search that was conducted in January of 2023 and later updated in June of 2024 in the following electronic databases and preprint servers: PubMed, ScienceDirect, Embase, Emcare, Academic Search Premier, IEEE, Acoustical Society of America, Springer, Web of Science, medRxiv, and arXiv. Additionally, studies that were published as conference proceedings and were not indexed on these databases and were known to authors were also added. The identified studies' citations and references were searched for other relevant studies. No restriction on the publication date was imposed.

C. Search Strategy

The search strategy was developed in collaboration with a medical information specialist. The search terms contained related keywords and Medical Subject Headings and were customised for each database. The full search strategy is available in the supplementary material.

TABLE II
INCLUSION AND EXCLUSION CRITERIA

Inclusion	TTS used to convert text to acoustic test stimuli.
	ASR used to capture participants' verbal responses.
	ASR used to optimize hearing aid electroacoustic configurations or other hearing devices.
	ASR used to analyse the auditory stimulus as in a speech intelligibility test procedure i.e., an ASR system adds information about the participant's perception and likely response.
Exclusion	Studies that predicted speech intelligibility in individuals with normal hearing.
	Studies that used ASR for signal processing to alter the output of a hearing aid without patient data from ASR i.e., studies that use ASR without using new or existing individual data (such as hearing thresholds) for personalisation.
	Studies that used ASR to predict speech intelligibility as a function of background noise without giving instructions on how this can be used to set the parameters of a hearing aid.
	Studies that simulated hearing loss to be applied to normal hearing participants. i.e., studies that distort the signal and present it to people with normal hearing.
	Studies using ASR models trained and tested on the same sets of stimuli.
	Publications not written in English.

D. Data Management

Identified studies were exported to the Zotero reference management software to check for any duplicate that might have been missed by the information scientist and to find any retracted studies. The remaining records were exported to an Excel spreadsheet for eligibility checking.

E. Selecting Relevant Records

Initially, The search strategy retrieved 2942 studies and after preliminary screening 1826 of them were selected. Then, two authors (MF and JS) independently read the titles (selecting 151) and abstracts (selecting 49) of the remaining papers. If there was disagreement or uncertainty about inclusion based on the title and abstract, those studies were assigned to two authors (MF and one other author) for a full-text reading and checking against the inclusion and exclusion criteria. Any disagreement between the two authors was resolved by discussion, and if the disagreement was not resolved, a third author was consulted for a final decision. There was a total of 39 disagreements: 21 (53%) when reading the title, 13 (33%) when reading the abstract and 5 (12%) when reading the whole document. Additionally, 11 conference proceedings were also added. The full details of the selection process are shown in Fig. 1.

F. Data Extraction Process

A data extraction table was designed to extract information from each study in a systematic manner. The primary author (MF) performed the data extraction, while four of the remaining authors individually examined and confirmed the findings on 16% of the studies.

V. RESULTS

Overall, 28 studies were selected that met the inclusion criteria. These studies were divided into four categories based on their objective and how they used ASR and TTS.

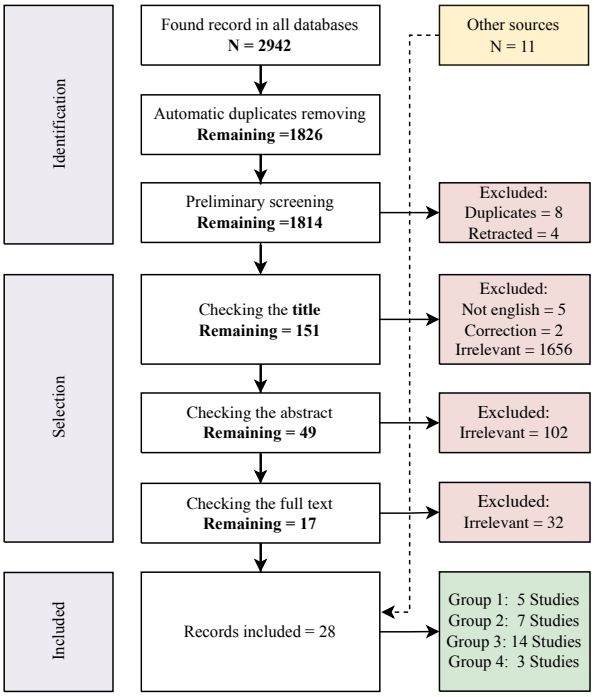

Fig. 1. Article selection process. Based on how ASR and TTS were used, the studies were categorised into four groups (Group 1: TTS for generating the acoustic stimuli, Group 2: ASR for capturing the verbal response, Group 3: ASR for estimating speech test performance), Group 4: ASR for configuration of hearing aid parameters. Two studies were assigned to both Group 1 and Group 2.

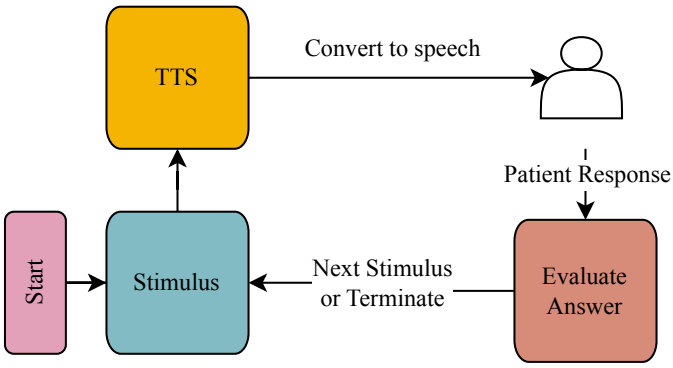

Fig. 2. Diagram showing how TTS is used in speech intelligibility tests. A synthetic stimulus is presented and the participant is asked to repeat the stimulus they hear. This procedure is repeated until the predefined stop condition is satisfied.

A. TTS for generating the acoustic stimuli to be used in speech intelligibility tests

The five (17%) studies in this category replaced the pre-recorded stimuli with sounds synthesised with TTS, which were then used to evaluate the participant's hearing [24], [25], [26], [18], [27]. Their main goal was to reduce the time and effort needed to generate a new dataset of test stimuli. The overall flow of speech intelligibility tests with a TTS system is presented in Fig. 2 and the extracted data is provided in Table I of the supplementary material.

The first research [24] on TTS for stimuli generation was published in 1990. The researchers used the DECtalk program

[28] to synthesise vowels which were then randomly combined to generate word stimuli. They tested their method on 30 participants with normal hearing and 15 participants with hearing loss and reported that 100% of the synthesised stimuli were recognised by the participants. However, they only mentioned achieving a reasonable level of speech recognition accuracy when the synthesised stimuli were distorted. They did not provide specific numerical results about the level of distortion and the accuracy of speech recognition. The authors concluded that due to high accuracy and the ability to freely alter the parameters of synthesised stimuli, their model has the potential to improve the speech intelligibility test procedures.

Advances in machine learning significantly improved the quality of TTS systems, enabling them to generate sentences with human-like voices in real-time. This advance in TTS technology has resulted in more researchers using it for speech intelligibility tests. Nuesse et al. [25] used a commercial TTS system developed by the Acapela group [29] that used a non-uniform unit selection [30] for synthesising the German matrix sentence (OLSA) dataset [31]. The OLSA dataset is a set of 5-word sentences with a predefined grammatical structure, and for each word in a sentence, there are 10 possible options. The stimuli are generated by combining different words. To test their method, the authors evaluated the SRT of 48 participants with normal hearing in a soundproof booth using the 150 sentences from the OLSA dataset [31]. They reported that their method achieved an SRT of +0.5 dB relative to the same test with recorded stimuli, which can be considered negligible. Furthermore, the psychometric functions (showing the relationship between the SNR and correct response percentage) were similar for the two methods. However, the researchers did not examine the effect of synthetic stimuli on participants with hearing problems. They concluded that using synthetic stimuli reduces the cost and time of generating the test without compromising the accuracy.

Ibelings et al. [18] used a new TTS system from the Acapela group [29] to synthesise another German dataset (GöSa [32]), generating 200 sentences with male and female speakers. These sentences were used to evaluate the 25 individuals with normal hearing at home via the Internet. The results indicated a lower SRT of 1.2 dB when using synthetic stimuli compared to natural stimuli, but this was no greater than the differences between different natural speakers. Consequently, the authors reported that the use of synthetic stimuli does not impact the test performance negatively, and it reduces the time and effort required to generate the stimuli.

Ooster et al. [26] designed a remote and automated speech intelligibility test that could be administrated in participants' homes using both TTS and ASR. This study used the same synthesised stimuli as Nuesse et al. [25] and a pretrained ASR system from Amazon. To evaluate the method, OLSA [31] was used in various simulated sound fields (e.g., living room, classroom, and concert hall) on 46 participants with hearing losses from 25 to 60 decibel hearing level (dB HL). The SRT calculated with their model had a bias from 0.7 dB for moderately hearing impaired to 2.2 dB for young people with normal hearing compared to the clinical SRT. The intrasubject standard deviations for participants with normal hearing and hearing loss were 0.63 and 1.01 dB, respectively. Based on these results the authors claimed that their proposed method was valid for a self-supervised hearing test at home.

Polspoel et al. [27] used Google Cloud API to synthesise English and Dutch triplet digits (0-9). To evaluate their system, they recruited 28 participants with normal hearing and 20 participants with hearing loss (47 ± 19 dB HL). Their proposed method had a high Pearson correlation with the reference test for both English (0.95) and Dutch (0.91) digits. Additionally, they also report test-retest reliability close to their reference test for both English (1.7 dB) and Dutch (0.6 dB) digits. With these results, they showed that the TTS system is capable of creating multi-lingual digits-in-noise tests with much less effort compared to traditional methods of generating stimuli.

Except for the first study by Kosai et al. [24], all studies used off-the-shelf proprietary TTS engines capable of generating human-like voices. The results showed a higher SRT than for the traditional method for participants with hearing loss than for normal hearing. However, Nuesse et al. [25] and Ibelings et al. [18] only used participants with normal hearing to evaluate their model; therefore, it was not clear how well their system worked with participants with hearing loss.

Additionally, examined studies [18], [25], [27] synthesised a relatively small and finite number of sentences and manually examined each generated sentence. However, they did not explore the capabilities of TTS for generating stimuli in real-time and whether the TTS could reliably generate high-quality stimuli during the test session or not. If studies were conducted on this topic, audiologists could generate new stimuli for each testing session and reduce the learning effects that are inherent to the speech intelligibility tests with limited vocabulary [33], [34].

B. ASR for capturing the verbal response of the participant

The seven (25%) studies in this category replaced the human supervisor with an ASR system, which was then used to automatically assess participants' responses [35], [36], [37], [26], [38], [19], [39]. The main goal of these studies was to create a speech intelligibility test that could be done without human supervision or even remotely in the participants' homes. The overall flow of the test using an ASR system is presented in Fig. 3 and the extraction table is provided in Table II of the supplementary materials.

In 2015, Meyer et al. [35] built an ASR system with a Hidden Markov Model (HMM) [40] and Mel-frequency cepstrum coefficients(MFCC) [41] trained on a dataset of 23.2 hours of speech. To assess their method, they used the OLSA dataset [31] with their ASR system to calculate the SRT and, compared their result with SRTs obtained in a clinical setting. The exact numbers were not reported and it was only stated that if the participant did not use words that were new to the ASR system (out-of-vocabulary (OOV)), the system could achieve a test-retest standard deviation of 0.5 dB. However, the limitation of using no OOV means that the system is not usable in complex and realistic test settings.

Ooster et al.'s next work [36] built upon their own previous study [35]. They improved the training dataset of the ASR

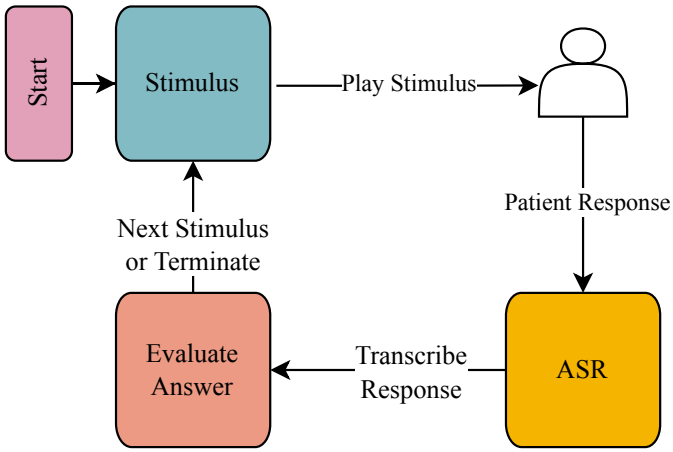

Fig. 3. Diagram showing how speech intelligibility tests with ASR works. In this diagram, the ASR transcribe the participant's response and the evaluation of the response is done with the transcribed text.

by introducing 18 hours of OOV words and trained an ASR system using the new dataset. To evaluate their method, they used 20 listeners with normal hearing and 7 listeners with hearing loss in a soundproof booth with the OLSA dataset [31]. They reported an SRT bias of +0.5 dB for participants with normal hearing and 0.8 dB for participants with hearing loss and the test-retest standard deviation was 0.5 dB and 0.9 dB, respectively. Based on these results, they concluded that the ASR system provides a reliable measurement of SRT for participants with hearing loss and participants with normal hearing.

Ooster et al. [26] proposed another automatic test that used both ASR and TTS. This study was described in the previous section. They used a commercial ASR system from Amazon. They tested the proposed method on 46 participants with different levels of hearing loss and, as described above, concluded that the discrepancies from a conventional test were small.

The most recent study by Ooster et al. [19] aimed to enhance the accuracy of the ASR system by using a new time-delay neural network [42] with MFCCs [41] and an HMM [40]. This reduced the percentage of unrecognised words from 4.76% to 0.6%. The ASR system was trained on 23 hours of data from Meyer et al. [35] and another dataset with 18 hours of speech [43], [44]. To evaluate the system, 20 listeners with normal hearing, 39 listeners with hearing loss and 14 listeners with cochlear implants were tested in a soundproof booth using the OLSA dataset [31]. The results indicated that compared to the traditional method there was a bias of 1.4 dB for 95% of participants with normal hearing and unaided hearing impaired (i.e., without using their hearing aid) and a bias of 2.1 dB for participants with cochlear implants.

Another study that tried to improve the ASR architecture was undertaken by Nisar et al. [37]. They proposed an adaptive way of giving weights to MFCC features based on the input sound spectrum, leading to enhanced accuracy in the ASR system. They trained the ASR system on a dataset of 3600 utterances and used a dataset of 72 English spondee words (words with two equally stressed syllables. e.g., baseball) for the test. The testing involved 60 participants with various levels of hearing loss and was conducted in a soundproof booth. They did not report the exact SRT bias of their system and only mentioned that it was less than 4.4 dB, which was high compared to other studies. However, the system was able to detect the category of hearing loss (e.g., mild, moderate, and severe) with 96.6% accuracy.

During the COVID-19 pandemic, Bruns et al. [38] developed a fully remote speech intelligibility test. To implement the ASR system, they adopted the model proposed by Peddinti et al. [42], which used a deep neural network with MFCC feature extractor [41] and trained the model on 1000 hours of an in-house German speech dataset. They recruited 16 participants with normal hearing and used the OLSA dataset [31] to test their hearing from their homes in a quiet room. The achieved SRT was 1 dB higher than the clinical SRT for all the participants and they reported a Pearson correlation of 0.93 with the human lead test. This is the only study that used the Internet. The authors concluded that remote testing of hearing with the use of ASR is a valid alternative to the traditional method.

ASR has also been added to the digits-in-noise test. Araiza-Illan et al. [39] proposed a self-supervised digits-in-noise test using an ASR system trained on 1000 hours of Dutch speech [45]. They initially recruited 30 participants with normal hearing to test their ASR in a quiet room and reported a word error rate of 5%. They then selected 6 participants with zero ASR error rates and used bootstrapping to model the effect of the ASR error on the final SRT measurement and reported that if the number of ASR decoding errors was less than 4, their system did not produce more variation than a clinical test (<0.7 dB).

One study [26] used commercial ASRs, while others (85%) trained their model using HMM [40] and MFCCs [41] based model. Nisar et al. [37] proposed a new method to calculate MFCCs, however, they did not provide any metric on its performance to show how much it improved the baseline.

Five studies (71%) used the OLSA [31] as the test stimuli and reported SRT bias of approximately 1 dB. However, Meyer et al. [35] did not compare their method with the clinical SRT. Nisar et al. [37] used an English dataset, and had a system with a high bias (<4.4 dB) compared to the other method.

Regarding the test environment, four studies (57%) conducted the tests in a soundproof room to minimise the effect of surrounding noise on the SRT, while one (14%) of them investigated the effect of different environments and noises on the test. Two studies (28%) conducted the test in a quiet room with one of them being conducted remotely over the Internet. And Only one study did not report the test environment (14%).

C. ASR for estimating speech test performance

The 14 (50%) studies in this category predicted speech intelligibility [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59]. They used ASR to simulate a person with hearing loss, and the simulation must reach the same result as the participant it replaced. Their goal was to analyse the effects of different stimuli and test environments and gain insight

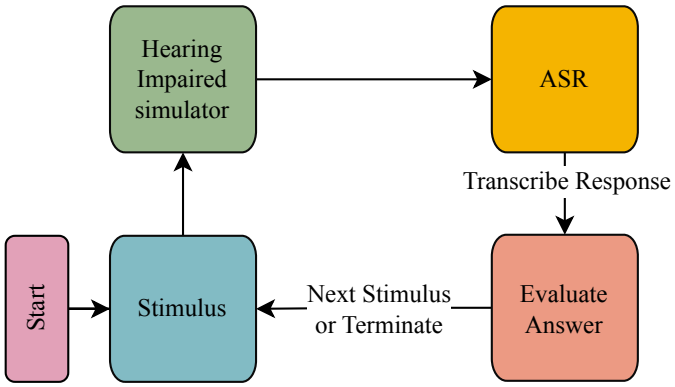

Fig. 4. Diagram showing how ASR is used to simulate speech intelligibility tests. In this system, an ASR and a hearing loss simulator replace a person with hearing loss, and it should achieve a result close to the person it is modelling.

into various situations that can affect speech intelligibility. The overall flow of simulating speech intelligibility tests with ASR is presented in Fig.4 and the extraction table is provided in Tables III, IV and V of the supplementary materials.

Fontan et al. [46] trained the ASR model using SPHINX-3 [60] on 31 hours of French radio broadcast recordings [61]. They evaluated the ASR system on three types of inputs; pseudoword [62] (87.4% accuracy), words [63] (98.3% accuracy) and sentences [64] (90.8% accuracy). The aim of this study was to predict the word identification scores of older adults with hearing loss by using stimuli with various levels of linguistic complexity. They tested their model with 24 participants with hearing loss in a soundproof room using the hearing loss simulator proposed in [65] and reported a correlation of 0.81 for pseudowords, 0.77 for words and 0.71 for sentences between the proposed model and empirical data. Based on these results, the researchers claimed that there is a strong correlation between human and machine results in all three types of stimuli but the pseudowords showed the strongest correlation.

Another study that trained their own ASR was done by Roßbach et al. [47]. They trained a deep learning-based ASR system based on the architecture proposed by [66] on 10 hours of speech from 20 speakers in the shape of the OLSA dataset [31]. The trained model was then used to simulate the SRT measurement procedure. The stimuli for testing were from the OLSA dataset with speech and noise-like maskers generated from [67]. The hearing loss simulation was done by replacing the spectral components below the individual hearing threshold with Gaussian noise at the same level as the individual's hearing threshold. To test their method, they recruited 8 participants with normal hearing and 20 participants with hearing loss to do the speech-in-noise test and reported that their method had an SRT bias of 1.6 dB for participants with normal hearing and 1.4 for participants with hearing loss.

Brochier et al. [48] focused on participants with cochlear implants. They developed an ASR model with a fully computational front-end to simulate cochlear implant perception and to predict phoneme recognition of cochlear implant users. They compared the predictions of their model to data from 35 participants with cochlear implants from [68], [69] and reported a significant correlation for the prediction of consonants (R=0.65) but not for vowels (R=0.38). Predicted SRTs were within 1 dB of those of the cochlear implant users and confusion matrices showed large agreement.

A further set of methods was developed in response to the Clarity Prediction Challenges (CPC) [70], [71]. These aimed to facilitate the development of systems that could estimate the speech intelligibility score of a person with hearing loss from speech stimuli. In this challenge, stimuli in the form of 7 to 10 word-long sentences in noisy environments were simulated with head-related transfer functions (representing hearing loss) and processed by ten hearing aid algorithms. The stimuli were then presented to the listeners, who were asked to repeat what they had heard. Challenge participants were asked to predict how many of the words were recognised with each specific hearing loss condition. CPC1 produced a dataset of 7233 responses from 27 listeners (hearing loss 15 to <80 dB), whereas CPC2 produced a dataset of 10062 responses from 18 listeners (<35 dB and >80 dB) while using more diverse and complex noises and head movements.

The submitted system could be intrusive or non-intrusive. The intrusive system had access to both the enhanced audio and the reference audio with its transcription, while the non-intrusive system had only access to the enhanced signal. Both intrusive and non-intrusive systems could use the input speech alongside metadata (see [70] for the full list) that showed listener and room characteristics. The extraction table includes the best model of each submitted paper (CPC1: N=6, CPC2: N=5) and is provided in Tables IV and V of the supplementary materials.

In CPC1, intrusive systems, on average, performed better as they had access to the clean reference data [49], [50], [53]. The winner of the CPC1 [72] was an intrusive system, but they did not use ASR. The best ASR based system [49] used an ASR model to create a representation for both the reference speech and the one enhanced by the hearing aid. They compared the two created representations with each other to calculate speech intelligibility and achieve a correlation of 0.76 on the open dataset.

In CPC2, Huckvale et al. extended their previous model [72] by using Wav2Vec [73] and fine-tuning it on the Cambridge read news dataset [74], achieving a correlation score of 0.78. Tu et al. [58] also extended on previous entries to CPC1 [49], [51]. Both models used pre-trained transformer-based ASR. The intrusive model compared the features generated by the ASR for the clean reference and target speech (correlation score = 0.77), while the non-intrusive system estimated the uncertainty of the ASR system (correlation score = 0.72).

The winner [55] of this challenge was a non-intrusive system that used pre-trained WavLM [75] and Whisper [76] models to extract features from speech signal. Extracted features are then mapped to the speech intelligibility score using transformer models. With this system, they managed to achieve a correlation score of 0.78. One common approach in CPC2 was the use of foundation models like Whisper [76] to extract features from the input speech and use another machine learning algorithm to map the extracted features to

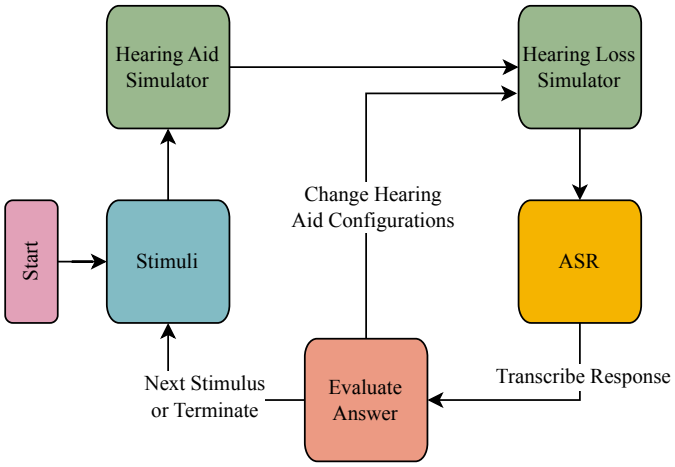

Fig. 5. Diagram showing how ASR is used to fit a hearing aid. In this diagram, an ASR is used to evaluate a particular hearing aid configuration and change the parameters to achieve the best results.

intelligibility score ([57], [59]).

This group consists of studies that used ASR to investigate the effects of various stimuli and situations by simulating the speech-in-noise test. Among these studies, only one study (7%) [48] focused on participants with cochlear implants. This lack of research in this category indicates that there needs to be more research on participants with cochlear implants to better investigate the potential of ASR for them.

An important point to consider is that the model proposed by Roßbach et al. [47] used different stimuli for training and testing but the same noise is used, hence, it is unclear how the model will perform in the presence of unseen noise.

Models submitted to the clarity challenge all had the same dataset which made the comparison easier. There was a trend of using pre-trained ASR models to extract features from input speech and this is much more prominent in the second challenge where a non-intrusive model with a pre-trained model outperformed the intrusive systems.

D. ASR for configuration of hearing aid parameters

The three studies (10%) in this category consisted of research that uses ASR to find an optimum configuration for hearing aids [77], [78], [79]. They evaluated the intelligibility of hearing aid outputs by measuring how well the ASR can understand the altered signal and aimed to achieve the maximum score by optimising different hearing aid parameters (e.g., insertion gains). The overall architecture of the studies that used ASR for fitting hearing aids is presented in Fig. 5 and the extraction table is provided in Table VI of the supplementary materials.

Fontan et al. [77] used the SPHINX-3 [60] ASR system and trained it on 31 hours of French radio broadcast recordings [61]. Their objective was to determine the insertion gains of a hearing aid in a way that would maximise the performance of the ASR system. To achieve this, they took a hearing aid that was fitted based on CAM2 [80], a validated generic prescription method, and tested 625 predetermined gain functions to identify the optimal insertion gains. For evaluation, the researchers used 60 disyllabic nouns [63] and 40 sentences from the French hearing in noise test [64] as stimuli and tested this method on 24 participants with hearing loss in a soundproof booth. They reported that the ASR-based configuration resulted in a higher mean intelligibility score than the CAM2 configuration (98.2% compared to 96.5%). However, given the small magnitude of this increase and both values close to the maximum, it is not clear that this has a clinically significant impact. They also asked the participant to score their comfort level for both configurations and reported a higher comfort score when the hearing aid was set up based on the ASR (8.4 compared to 7 out of 10). This increase in comfort level might be due to the fact that the method sets less amplification for higher frequencies, which leads to higher pleasantness [81].

Gonçalves Braz et al. [78] expanded Fontan's work by using genetic algorithms [82] and expanding the search space (set of all possible values) for fitting parameters. The authors aimed to optimise two parameters: insertion gains and compression threshold. The insertion gains were optimised across five frequencies with a step size of 0.1 dB and in the range of ± 10 dB to the prescribed insertion gains. The search for compression threshold was done between 20 dB SPL to 50 dB SPL with a step size of 1 dB. Regarding the ASR system, the authors used the Julius 4.4.2 system [83] and trained it on 100 hours of French radio broadcasting recordings [61], [84]. To evaluate the model, they used 60 disyllabic nouns [63] and fitted the simulated hearing aid based on the audiograms of 12 people with hearing loss. With the proposed configuration, the ASR system achieved an intelligibility score of 98%, surpassing the 88% achieved when listening to the output of hearing aids configuring based on CAM2 [80]. To check the consistency of the system, they repeated the procedure 12 times for each audiogram and achieved a correlation of >0.95 between each audiogram's results.

The final study [79] was a continuation of their previous research [78] and used their proposed model to find the best attack and release time constants, which determine how quickly a hearing aid adjusts its amplification as a function of input level. The search space for attack time spanned from 100 to 500 ms and for the release time, it extended from 300 to 2000 ms with a step size of 10 ms. To evaluate the effectiveness of the optimisation of the time constants, they used the same 12 audiograms as the earlier study [78]. While they reported an increase in the ASR intelligibility score compared to CAM2 [80] fitted hearing aid (92% compared to 88%), the results showed no improvement from the configuration obtained from [78]. As for consistency, they ran the experiment twice for each participant and reported there was no statistical difference between the results of the two experiments.

Models that optimise hearing aids have two main components, the first is the hearing aid simulator, which the studies want to find the best configuration for, and the hearing loss simulator. The hearing loss simulator's job is to degrade the signal to replicate a person with hearing loss. All the papers in this group used the same hearing loss simulator [65] which can simulate the loss of audibility and recruitment in a person.

Studies in this category were all conducted by the same

group, which incrementally complement each other to cover all major settings of hearing aids (insertion gains, compression threshold, and time constants). They all used French radio broadcasting to train the ASR model. While one (33%) of the studies investigated the comfort level of human participants, the other two (67%) only reported the score of the ASR system when the hearing aid's output was fed to it.

Furthermore, one (33%) of the studies used a predefined set of parameters, however, two (67%) of them use evolutionary algorithms [85] to find the optimal values faster and by doing so, they were able to expand their search space.

VI. Discussion

This scoping review identified studies that used ASR and TTS technologies for assessments of both hearing and hearing aid fitting, and grouped them into four categories, based on how they used these technologies. There has been less research on creating synthetic speech or for the automatic configuration of hearing aids compared to simulated SRT measurement and ASR operated tests.

The dominant language with the exception of the Clarity challenge was German, with a few studies in French. there is a lack of diversity in using other languages and speech intelligibility tests datasets. Additionally, there is a lack of diversity in the researchers themselves. For example, in the "ASR for configuration of hearing aid parameters" group, all three studies were conducted by the same group. Similarly in the "ASR for capturing the verbal response of the participant" category, five out of the eight studies were conducted by the same research group. Consequently, the studies in each group are very similar to each other and there is a need for more researchers to evaluate these topics independently and to bring forth new ideas and innovations to this domain.

A. TTS for generating the acoustic stimuli to be used in speech intelligibility tests

TTS can be used for generating stimuli for speech intelligibility tests. The studies in this category investigated the effect of using machine generated stimuli instead of using prerecorded speech. The current studies are mostly on German datasets with a limited vocabulary (OLSA) and one study synthesised English and Dutch digits [27]. However, there is no proof to date that the approach generalises to a variety of factors like the voice gender, the used dataset, and the language of the stimuli. Thus, there is a need for more research to explore methods for creating TTS models that can create synthetic speech in other languages and other speech intelligibility tests datasets and evaluate them in speech intelligibility tests and for participants with hearing loss.

The SRT bias for participants with hearing loss tends to be different from participants with normal hearing when using TTS or ASR. Thus, the effect of TTS should be investigated on both types of listeners.

Compared to other speech intelligibility tests (digit-in-noise and word-in-noise), the sentence-in-noise test has a more diverse vocabulary and uses stimuli with a more complex structure and is closer to natural speech. However, the choices of words are still limited and reusing this limited vocabulary in multiple test sessions leads to learning effects [33]. We believe that a better way of employing the TTS system is to generate new and meaningful stimuli with different words for every test, thereby preventing any learning effect. In this method, since the stimuli are generated at the test time, pre-recorded stimuli are no longer useful since we do not know the stimuli beforehand. However, no research has been conducted to date that uses TTS in a more flexible manner than generating a predefined set of sentences.

Using TTS to create new stimuli has its own challenges. The first problem is to have an algorithm to select stimuli with proper words and sentences that are suitable for speech intelligibility tests. Having a limited number of words makes it easier to create a high-quality TTS system. However, when the stimuli are generated by an algorithm for an unknown sentence structure, and the vocabulary is unlimited, it becomes challenging to prove that a TTS system produces all stimuli correctly.

B. ASR for capturing the verbal response of the participant

The studies in this category investigated the effectiveness of using ASR for evaluating participants' responses during a hearing test. Test-retest reliability measures the method's consistency by comparing the measured SRT of the same person across multiple experiments. This was around 0.5 dB in normal hearing [86] and 0.9 dB in hearing impaired [87] for a clinical test. Three of the reviewed studies reported this metric and their results were in the acceptable range. However, other studies did not report this metric which makes it hard to evaluate their consistency.

One of the main advantages of using ASR instead of a human supervisor is that people can test their hearing without supervision. However, only two studies conducted their testing in a normal environment. To achieve the goal of an unsupervised speech intelligibility test, more studies need to focus on conducting the test in" everyday" locations like the home of the participants and investigate ways to improve the performance of their system in such environments.

Creating an ASR that can discriminate a limited number of words (e.g., the OLSA dataset) in a quiet and controlled environment is relatively easy. However, creating an ASR that achieves high accuracy in an uncontrolled environment is challenging. High accuracy is necessary because otherwise it cannot be distinguished if a wrong response was due to the participant giving a wrong response or the ASR system not recognising a correct response of the participant. The system needs to consider various acoustic environments, background noise, and uncalibrated devices. Adding this to our suggestion of using TTS for creating a new stimulus for each test session means that the ASR will have a harder job, as it needs to accurately recognise a much more diverse set of vocabulary. While this is a challenging task, we believe that more research and effort into this topic can lead to a fully automated and reliable test that can be done without visiting a clinic.

Some important questions were beyond the scope of the studies in the current review. however, they are worth inves-

tigating in future studies. These include doing a comparison of fitted hearing aids based on ASR or human SRT measurements. Doing so can better show the applicability of an ASR-based test, as opposed to only comparing the SRTs. Secondly, Using the results of hearing measurements done by multiple trained audiologists instead of one, yields a more accurate ground truth and a better comparison of the SRT measured by ASR and an audiologist.

C. ASR for estimating speech test performance

The studies in this category used ASR to investigate the effects of various stimuli and listening environments by simulating a measure of speech intelligibility. The first study on this subject was done by Fontan et al. [46] using French stimuli of varying linguistic complexity to predict word identification scores. Brochier et al. [48] was the only study that focused on people with cochlear implants and There is a clear need for more research on this approach for cochlear implants.

The clarity challenge introduced two datasets for this task. Using a standard dataset not only makes comparison of different submitted models possible but is also beneficial for comparison of future models as other researchers can run their system on the clarity challenge dataset and compare their results with other systems that used the same dataset.

This challenge had two non-intrusive and intrusive modes, however, while in the first challenge intrusive models outperformed non-intrusive systems, Large foundation ASR models like Whisper [76] model had a big impact on the second challenge and enabled non-intrusive models to outperform the intrusive ones.

One point to consider about the studies submitted to both clarity challenges is that only a few of them were published in peer-reviewed journals ([88], [89], [90]) and the rest were published as pre-print or conference proceeding with some of them providing limited information regarding their used method.

Unlike other groups, studies in this section did not report the test-retest reliability of their proposed method. This is because, with the same input, the ASR will always perform the same and generate the same output, thus, the test-retest reliability of these models is perfect.

D. ASR for configuration of hearing aid parameters

The final category investigated if ASR can be used to quickly and automatically compare different hearing aid settings and find the most suitable configurations for the person that yields the highest speech intelligibility.

French disyllabic nouns were used in all three studies. Fontan et al. [77] also used a French speech-in-noise test [64]. The limited diversity of datasets stems from all three studies having been done by the same group. In their first study, they compared participants' comfort levels while using hearing aids fitted based on ASR and the CAM2 method, but unfortunately, they did not report the comfort level in their other two studies.

Furthermore, the researchers compared ASR scores for speech generated by hearing aids set up using the prescription formula of CAM2 and set up using their own ASR algorithm. Based on the results they concluded that their proposed system reaches a better configured hearing aid. However, this is not surprising since, during their own method to set up the hearing aid, they start with the CAM2 setting and choose the parameters to maximise the ASR score. Although a clear potential was demonstrated, it is necessary to evaluate the settings with human participants rather than an evaluation metric that is highly similar to the metric that was used for the optimisation.

An unexplored area in this topic is the involvement of the patient by simulating the hearing aid with different configurations and letting the person see how different configurations would change the possible output of the hearing aid. By doing this, the patient can see the benefit of a hearing aid before ever fitting one, and they can test different configurations from their home. However, automatically presenting good candidate fittings of hearing aids is challenging. One approach was done by Nielsen et al. [91] using active learning, and this may be improved by including ASR based suggestions. AI, TTS and ASR systems can be helpful to mitigate these problems. TTS can be used to accurately simulate different types of sentences and stimuli. ASR can be used to test the person's hearing after altering the hearing aid configuration and optimisation algorithms can assist in finding the best configurations without requiring a complex setting of parameters as an audiologist would do.

VII. Conclusion

ASR and TTS have been used for several purposes in speech intelligibility testing and to set up hearing devices. Both ASR and TTS have the potential to be used in hearing assessment and hearing aid fitting, improving accuracy, and decreasing the reliance on human experts. Research priorities include creating remote and unsupervised speech intelligibility tests, creating more natural stimuli using TTS, and creating hearing aid simulators usable by the hearing aid users themselves.

VIII. Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank Dr David Moore, professor of auditory neuroscience, Cincinnati Children's Hospital, for his help and guidance during the preparation of this manuscript.

References

[1] "World report on hearing," https://www.who.int/publications-detail-redirect/9789240020481, 2021.
[2] D. G. Blazer and D. L. Tucci, "Hearing loss and psychiatric disorders: A review," *Psychol. Med*, vol. 49, no. 6, pp. 891–897, Apr. 2019.
[3] T. D. Griffiths, M. Lad, S. Kumar, E. Holmes, B. McMurray, E. A. Maguire, A. J. Billig, and W. Sedley, "How Can Hearing Loss Cause Dementia?" *Neuron*, vol. 108, no. 3, pp. 401–412, Nov. 2020.
[4] J. Fagan and M. Jacobs, "Survey of ENT services in Africa: Need for a comprehensive intervention," *Glob. Health Action.*, vol. 2, no. 1, p. 1932, Nov. 2009.

[5] A. M. Planey, "Audiologist availability and supply in the United States: A multi-scale spatial and political economic analysis," *Soc. Sci. Med.*, vol. 222, pp. 216–224, Feb. 2019.

[6] G. Grasselli, A. Zangrillo, A. Zanella, M. Antonelli, L. Cabrini, A. Castelli, D. Cereda, A. Coluccello, G. Foti, R. Fumagalli, G. Iotti, N. Latronico, L. Lorini, S. Merler, G. Natalini, A. Piatti, M. V. Ranieri, A. M. Scandroglio, E. Storti, M. Cecconi, A. Pesenti, and COVID-19 Lombardy ICU Network, "Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy," *JAMA*, vol. 323, no. 16, pp. 1574–1581, Apr. 2020.

[7] Centers for Disease Control and Prevention, "Interim US Guidance for Risk Assessment and Public Health Management of Healthcare Personnel with Potential Exposure in a Healthcare Setting to Patients with Coronavirus Disease (COVID-19)," 2020.

[8] K. C. Hustad and M. A. Cahill, "Effects of Presentation Mode and Repeated Familiarization on Intelligibility of Dysarthric Speech," *Am. J. Speech Lang. Pathol.*, vol. 12, no. 2, pp. 198–208, May 2003.

[9] C. Sorin and C. Thouin-Daniel, "Effects of auditory fatigue on speech intelligibility and lexical decision in noise," *J. Acoust. Soc. Am.*, vol. 74, no. 2, pp. 456–466, Aug. 1983.

[10] H. Dillon, J. Day, S. Bant, and K. J. Munro, "Adoption, use and non-use of hearing aids: A robust estimate based on Welsh national survey statistics," *Int. J. Audiol.*, vol. 59, no. 8, pp. 567–573, Jul. 2020.

[11] A. McCormack and H. Fortnum, "Why do people fitted with hearing aids not wear them?" *Int. J. Audiol.*, vol. 52, no. 5, pp. 360–368, May 2013.

[12] J. P. Whitton, K. E. Hancock, J. M. Shannon, and D. B. Polley, "Validation of a Self-Administered Audiometry Application: An Equivalence Study: Equivalence of Mobile and Clinic-Based Tests," *Laryngoscope.*, vol. 126, no. 10, pp. 2382–2388, Oct. 2016.

[13] I. Almufarrij, H. Dillon, P. Dawes, D. R. Moore, W. Yeung, A.-P. Charalambous, C. Thodi, and K. J. Munro, "Web- and app-based tools for remote hearing assessment: A scoping review," *Int. J. Audiol.*, vol. 62, no. 8, pp. 699–712, Aug. 2023.

[14] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and Robust Machine Learning for Healthcare: A Survey," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 156–180, 2021.

[15] M. Cox and B. de Vries, "Bayesian Pure-Tone Audiometry Through Active Learning Under Informed Priors," *Front. Digit. Health*, vol. 3, p. 723348, Aug. 2021.

[16] K. Taylor and W. Sheikh, "Automated Hearing Impairment Diagnosis Using Machine Learning," in *2022 Intermt. Eng. Technol. Comput. IETC*. Orem, UT, USA: IEEE, May 2022, pp. 1–6.

[17] R. A. Osman and H. A. Osman, "On the Use of Machine Learning for Classifying Auditory Brainstem Responses: A Scoping Review," *IEEE Access*, vol. 9, pp. 110 592–110 600, 2021.

[18] S. Ibelings, T. Brand, and I. Holube, "Speech Recognition and Listening Effort of Meaningful Sentences Using Synthetic Speech," *Trends in Hearing*, vol. 26, p. 233121652211306, Jan. 2022.

[19] J. Ooster, L. Tuschen, and B. T. Meyer, "Self-conducted speech audiometry using automatic speech recognition: Simulation results for listeners with hearing loss," *Computer Speech & Language*, vol. 78, p. 101447, Mar. 2023.

[20] T. D. J. Van and J. L. Yanz, "Speech Recognition Threshold in Noise," *J. Speech Lang. Hear. Res.*, vol. 30, no. 3, pp. 377–386, Sep. 1987.

[21] J.-W. Wasmann, L. Pragt, R. Eikelboom, and D. W. Swanepoel, "Digital Approaches to Automated and Machine Learning Assessments of Hearing: Scoping Review," *J. Med. Internet Res.*, vol. 24, no. 2, p. e32581, Feb. 2022.

[22] M. Fatehifar, J. Schlittenlacher, D. Wong, and K. Munro, "Applications Of Automatic Speech Recognition And Text-To-Speech Models To Detect Hearing Loss: A Scoping Review Protocol," INPLASY, Tech. Rep., Jan. 2023.

[23] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, S. Hempel, E. A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M. G. Wilson, C. Garritty, S. Lewin, C. M. Godfrey, M. T. Macdonald, E. V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, Ö. Tunçalp, and S. E. Straus, "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation," *Ann. Intern. Med.*, vol. 169, no. 7, pp. 467–473, Oct. 2018.

[24] M. Kosai, J. Kohda, J. Udaka, and Y. Koike, "Speech audiometric trial using synthetic vowels produced with DECtalk," *Med. Inform. (Lond)*, vol. 15, no. 4, pp. 309–318, Jan. 1990.

[25] T. Nuesse, B. Wiercinski, T. Brand, and I. Holube, "Measuring Speech Recognition With a Matrix Test Using Synthetic Speech," *Trends in Hearing*, vol. 23, p. 233121651986298, Jan. 2019.

[26] J. Ooster, M. Krueger, J.-H. Bach, K. C. Wagener, B. Kollmeier, and B. T. Meyer, "Speech Audiometry at Home: Automated Listening Tests via Smart Speakers With Normal-Hearing and Hearing-Impaired Listeners," *Trends in Hearing*, vol. 24, p. 233121652097001, Jan. 2020.

[27] S. Polspoel, D. R. Moore, D. W. Swanepoel, S. E. Kramer, and C. Smits, "Global access to speech hearing tests," Jun. 2024.

[28] S. Lock and C. K. Leong, "Program library for DECtalk text-to-speech system," *Beh. Res. Meth. Instr. Comp.*, vol. 21, no. 3, pp. 394–400, May 1989.

[29] "Acapela," https://www.acapela-group.com/.

[30] J. R. Bellegarda, *TTS Unit Selection*. Cham: Springer International Publishing, 2007, pp. 71–76.

[31] K. Wagener, V. Kühnel, and B. Kollmeier, "Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test," *Z. Audiol.*, vol. 38, pp. 4–15, 1999.

[32] B. Kollmeier and M. Wesselkamp, "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2412–2421, Oct. 1997.

[33] T. Willberg, V. Sivonen, S. Hurme, A. A. Aarnisalo, H. Löppönen, and A. Dietz, "The long-term learning effect related to the repeated use of the Finnish matrix sentence test and the Finnish digit triplet test," *Int. J. Audiol.*, vol. 59, no. 10, pp. 753–762, Oct. 2020.

[34] A. Schlueter, U. Lemke, B. Kollmeier, and I. Holube, "Normal and Time-Compressed Speech: How Does Learning Affect Speech Recognition Thresholds in Noise?" *Trends Hear.*, vol. 20, p. 233121651666988, Jan. 2016.

[35] B. T. Meyer, B. Kollmeier, and J. Ooster, "Autonomous measurement of speech intelligibility utilizing automatic speech recognition," in *Interspeech 2015*. ISCA, Sep. 2015, pp. 2982–2986.

[36] J. Ooster, R. Huber, B. Kollmeier, and B. T. Meyer, "Evaluation of an automated speech-controlled listening test with spontaneous and read responses," *Speech Communication*, vol. 98, pp. 85–94, Apr. 2018.

[37] S. Nisar, M. Tariq, A. Adeel, M. Gogate, and A. Hussain, "Cognitively Inspired Feature Extraction and Speech Recognition for Automated Hearing Loss Testing," *Cogn Comput*, vol. 11, no. 4, pp. 489–502, Feb. 2019.

[38] T. Bruns, J. Ooster, M. Stennes, and J. Rennies, "Automated Speech Audiometry for Integrated Voice Over Internet Protocol Communication Services," *Am J Audiol*, vol. 31, no. 3S, pp. 980–992, Sep. 2022.

[39] G. Araiza-Illan, L. Meyer, K. P. Truong, and D. Başkent, "Automated Speech Audiometry: Can It Work Using Open-Source Pre-Trained Kaldi-NL Automatic Speech Recognition?" *Trends in Hearing*, vol. 28, p. 23312165241229057, Jan. 2024.

[40] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb./1989.

[41] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[42] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low Latency Acoustic Modeling Using Temporal Convolution and LSTMs," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 373–377, Mar. 2018.

[43] "King-ASR-L-092," http://en.htrs.demoboc.cn/datacenter/searchdetails/288.html.

[44] "King-ASR-L-182," http://en.htrs.demoboc.cn/datacenter/searchdetails/361.html.

[45] N. Oostdijk *et al.*, "The spoken dutch corpus. Overview and first evaluation." in *LREC*. Athens, Greece, 2000, pp. 887–894.

[46] L. Fontan, T. Cretin-Maitenaz, and C. Füllgrabe, "Predicting Speech Perception in Older Listeners with Sensorineural Hearing Loss Using Automatic Speech Recognition," *Trends in Hearing*, vol. 24, p. 233121652091476, Jan. 2020.

[47] J. Roßbach, B. Kollmeier, and B. T. Meyer, "A model of speech recognition for hearing-impaired listeners based on deep learning," *J. Acoust. Soc. Am.*, vol. 151, no. 3, pp. 1417–1427, Mar. 2022.

[48] T. Brochier, J. Schlittenlacher, I. Roberts, T. Goehring, C. Jiang, D. Vickers, and M. Bance, "From Microphone to Phoneme: An End-to-End Computational Neural Model for Predicting Speech Perception With Cochlear Implants," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 11, pp. 3300–3312, Nov. 2022.

[49] Z. Tu, N. Ma, and J. Barker, "Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-impaired Listeners," Jul. 2022.

[50] C. O. Mawalim, B. A. Titalim, and M. Unoki, "CPC1 E031 system description," in *Proc. 2nd Clarity Workshop Mach. Learn. Chall. Hear. Aids Clarity-2022 Online*, 2022.

[51] Z. Tu, N. Ma, and J. Barker, "Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-intrusive Speech Intelligibility Prediction," Jul. 2022.

[52] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids," Aug. 2022.

[53] N. Kamo, K. Arai, A. Ogawa, S. Araki, T. Nakatani, K. Kinoshita, M. Delcroix, T. Ochiai, and T. Irino, "Conformer-based fusion of text, audio, and listener characteristics for predicting speech intelligibility of hearing aid users," in *Proc. 2nd Clarity Workshop Mach. Learn. Chall. Hear. Aids*, 2022.

[54] J. Roßbach, R. Huber, S. Röttges, C. F. Hauth, T. Biberger, T. Brand, B. T. Meyer, and J. Rennies, "Speech intelligibility prediction for hearing-impaired listeners with the LEAP model." in *INTERSPEECH*, 2022, pp. 3498–3502.

[55] S. Cuervo and R. Marxer, "Temporal-hierarchical features from noise-robust speech foundation models for non-intrusive intelligibility prediction," in *Proc. ISCA Clarity-2023*, 2023.

[56] M. Huckvale and G. Hilkhuysen, "Combining acoustic phonetic linguistic and audiometric data in an intrusive intelligibility metric for hearing-impaired listeners," in *Proc. ISCA Clarity-2023*, 2023.

[57] R. Mogridge, G. Close, R. Sutherland, S. Goetze, and A. Ragni, "Pre-trained intermediate ASR features and Human memory simulation for non-intrusive speech intelligibility prediction in the Clarity Prediction Challenge 2," in *Proc. ISCA Clarity-2023*, 2023.

[58] Z. Tu, N. Ma, and J. Barker, "Intelligibility prediction with a pretrained noise-robust automatic speech recognition model," Oct. 2023.

[59] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep Learning-Based Non-Intrusive Multi-Objective Speech Assessment Model With Cross-Domain Features," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 54–70, 2023.

[60] K. Seymore, S. Chen, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, and E. Thayer, "The 1997 CMU Sphinx-3 English Broadcast News Transcription System," in *Proc. 1998 DARPA Speech Recognit. Workshop*, 1998.

[61] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Interspeech 2009*. ISCA, Sep. 2009, pp. 2583–2586.

[62] L. Dodelé and D. Dodelé, "L'audiométrie vocale en présence de bruit et filetest AVfB," *Cah Audit.*, vol. 13, no. 6, pp. 15–22, 2000.

[63] J.-E. Fournier, *Audiométrie Vocale: Les Épreuves d'intelligibilité et Leurs Applications Au Diagnostic, à l'expertise et à La Correction Prothétique Des Surdités*. Maloine, 1951.

[64] V. Vaillancourt, C. Laroche, C. Mayer, C. Basque, M. Nali, A. Eriks-Brophy, S. D. Soli, and C. Giguère, "Adaptation of the hint (hearing in noise test) for adult canadian francophone populations: Adaptación del hint (prueba de audición en ruido) para poblaciones de adultos canadienses francófonos," *Int. J. Audiol.*, vol. 44, no. 6, pp. 358–361, 2005.

[65] Y. Nejime and B. C. J. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *J. Acoust. Soc. Am.*, vol. 102, no. 1, pp. 603–615, Jul. 1997.

[66] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-Discriminative Training of Deep Neural Networks," in *Interspeech*, 2013, pp. 2345–2349.

[67] W. Schubotz, T. Brand, B. Kollmeier, and S. D. Ewert, "Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features," *J. Acoust. Soc. Am.*, vol. 140, no. 1, p. 524, Jul. 2016.

[68] C. M. McKay and H. J. McDermott, "Perceptual performance of subjects with cochlear implants using the Spectral Maxima Sound Processor (SMSP) and the Mini Speech Processor (MSP)," *Ear Hear.*, vol. 14, no. 5, pp. 350–367, Oct. 1993.

[69] B. Munson, G. S. Donaldson, S. L. Allen, E. A. Collison, and D. A. Nelson, "Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability," *J. Acoust. Soc. Am.*, vol. 113, no. 2, pp. 925–935, Feb. 2003.

[70] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. V. Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 3508–3512.

[71] J. Barker, M. A. Akeroyd, W. Bailey, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor, "The 2nd Clarity Prediction Challenge: A Machine Learning Challenge for Hearing Aid Intelligibility Prediction," in *ICASSP 2024 - 2024 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*, Apr. 2024, pp. 11 551–11 555.

[72] M. Huckvale and G. Hilkhuysen, "ELO-SPHERES intelligibility prediction model for the Clarity Prediction Challenge 2022," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 3934–3938.

[73] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Adv. Neural Inf. Process. Syst.*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[74] Robinson, Tony, Fransen, Jeroen, Pye, David, Foote, Jonathan, Renals, Steve, Woodland, Phil, and Young, Steve, "WSJCAM0 Cambridge Read News," p. 3670016 KB, 1995.

[75] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[76] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," in *INTERSPEECH 2023*, Aug. 2023, pp. 2798–2802.

[77] L. Fontan, M. Le Coz, C. Azzopardi, M. A. Stone, and C. Füllgrabe, "Improving hearing-aid gains based on automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 148, no. 3, pp. EL227–EL233, Sep. 2020.

[78] L. Gonçalves Braz, L. Fontan, J. Pinquier, M. A. Stone, and C. Füllgrabe, "OPRA-RS: A Hearing-Aid Fitting Method Based on Automatic Speech Recognition and Random Search," *Front. Neurosci.*, vol. 16, p. 779048, Feb. 2022.

[79] L. Fontan, L. Gonçalves Braz, J. Pinquier, M. A. Stone, and C. Füllgrabe, "Using Automatic Speech Recognition to Optimize Hearing-Aid Time Constants," *Front. Neurosci.*, vol. 16, 2022.

[80] B. C. J. Moore, B. R. Glasberg, and M. A. Stone, "Development of a new method for deriving initial fittings for hearing aids with multi-channel compression: CAMEQ2-HF," *Int. J. Audiol.*, vol. 49, no. 3, pp. 216–227, Mar. 2010.

[81] B. C. J. Moore, C. Füllgrabe, and M. A. Stone, "Determination of Preferred Parameters for Multichannel Compression Using Individually Fitted Simulated Hearing Aids and Paired Comparisons," *Ear Hear.*, vol. 32, no. 5, pp. 556–568, Sep. 2011.

[82] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: Past, present, and future," *Multimed. Tools Appl.*, vol. 80, no. 5, pp. 8091–8126, Feb. 2021.

[83] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proc. APSIPA ASC 2009 Asia-Pac. Signal Inf. Process. Assoc. 2009 Annu. Summit Conf.* Asia-Pacific Signal and Information Processing Association, 2009, pp. 131–137.

[84] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Proc. Fifth Int. Conf. Lang. Resour. Eval. LREC06*. Genoa, Italy: European Language Resources Association (ELRA), May 2006.

[85] P. A. Vikhar, "Evolutionary algorithms: A critical review and its future prospects," in *2016 Int. Conf. Glob. Trends Signal Process. Inf. Comput. Commun. ICGTSPICC*. Jalgaon, India: IEEE, Dec. 2016, pp. 261–265.

[86] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2801–2810, Jun. 2002.

[87] K. C. Wagener and T. Brand, "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters La inteligibilidad de frases en silencio para sujetos con audición normal y con hipoacusia: La influencia del procedimiento de medición y de los parámetros de enmascaramiento," *Int. J. Audiol.*, vol. 44, no. 3, pp. 144–156, Mar. 2005.

[88] X. Zhou, C. O. Mawalim, B. Angela Titalim, and M. Unoki, "Incorporating the Digit Triplet Test in A Lightweight Speech Intelligibility Prediction for Hearing Aids," in *2023 Asia Pac. Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC*. Taipei, Taiwan: IEEE, Oct. 2023, pp. 1593–1600.

[89] C. O. Mawalim, B. A. Titalim, S. Okada, and M. Unoki, "Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss," *Applied Acoustics*, vol. 214, p. 109663, Nov. 2023.

[90] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, "Non-Intrusive Speech Intelligibility Prediction for Hearing-

Impaired Users Using Intermediate ASR Features and Human Memory Models," in *ICASSP 2024 - 2024 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*. Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 306–310.

[91] J. Nielsen, J. Nielsen, and J. Larsen, "Perception-based Personalization of Hearing Aids using Gaussian Processes and Active Learning," *IEEE/ACM Trans. Audio Speech Lang. Process.*, pp. 1–1, 2014.

Applications of Automatic Speech Recognition and Text-To-Speech Technologies for Hearing Assessment: a Scoping Review

TABLE I
THIS TABLE SUMMARISES THE PAPERS THAT USE TTS IN THEIR METHOD

	(Kosai, 1990) [1]	(Nuesse, 2019) [2]	(Ooster, 2020) [3]	(Ibelings, 2022) [4]	(Polspoel, 2024) [5]
Participants' hearing level	30 NH, 15 HI	48 NH	46 HI from <25 to 60 dB HL	25 NH	NH: 28, HI: 20
Test environment	Unknown	Soundproof booth	Simulated living room, classroom, and concert hall.	Home of participants	Soundproof booth
ML Model's features	DECtalk [6]	Acapela [7]	ASR: ASR from Amazon TTS: Acapela [7]	Acapela cloud [7]	Google could API
ML Model's metrics	Unknown	Unknown	Insertion Error: 1.9% Deletion Error: 6.1%	Unknown	unknown
Method bias	Unknown	Same Slope 0.5 dB higher SRT	SRT bias from 0.7 dB for HI to 2.2 dB for NH.	1.2 dB lower SRT compared to natural stimuli.	R= 0.95 (English), R= 0.91 (Dutch)
Stimuli dataset	Unknown	OLSA [8]	OLSA [8]	GöSa [9]	Triplet digits
Training Data	Unknown	Unknown	Unknown	Unknown	Unknown
Test-Retest reliability of proposed method	Unknown	Unknown	Intrasubject standard deviation of 0.63 dB for NH and 1.01 dB for HI (Gold standard clinical test is 0.5 dB and 0.9 dB, respectively test).	Unknown	0.6 (Dutch), 1.7 (English)
Noise Generation algorithm	Low pass filter with 3350, 2240, 1800, 1400, 1120, 900, 710, 560 and 450Hz cut-off frequencies.	Superimposing from the test dataset.	Superimposing from the test dataset.	Superimposing from the test dataset.	Speech shape noise
Masking Noise Level	Unknown	Noise at 65 dB SPL and changing stimuli level to get SNR of -11, -8.5 and -6.	Constant 65 dB SPL	SNR of -4, -6 and -8 dB	Overall presentation at a comfortable level.
Stimuli level	Unknown	Adaptive	Adaptive	Adaptive	Overall presentation at a comfortable level.
Conclusion	An early attempt to use TTS for speech intelligibility tests. With TTS it is possible to freely alter the parameters of the stimuli. 100% articulation on undistorted synthesised vowels.	Using TTS can make the development of tests faster. Synthetic stimuli are able to generate comparable results to human-generated stimuli.	Used TTS to generate stimuli and ASR to capture the patient response.	Used TTS to generate stimuli for finding the SRT and showed it can achieve comparable results to natural speech while reducing the complexity of generating a new dataset.	Provided a tool for doing the DIN test with synthetic speech and ASR system. And reported high correlation with reference test.

NH: Normal Hearing, HI: Hearing Impaired, dB HL: Decibels at Hearing level, dB SPL: Decibels at Sound Pressure Level, DIN: Digits-in-Noise

TABLE II

THIS TABLE SUMMARISES THE PAPERS THAT USE ASR TO CAPTURE PARTICIPANTS' VERBAL RESPONSES.

	(Meyer, 2015) [10]	(Ooster, 2018) [11]	(Nisar, 2019) [12]	(Ooster, 2020) [3]	(Bruns, 2022) [13]	(Ooster, 2023) [14]	(Araiza-Illan, 2024) [15]
Participants' hearing level	Unknown	20 NH, 7 HI from 26 to 42 dB HL	60 from 0 to +90 dB HL	46 from <25 to 60 dB HL	16 NH	20 NH, 39 HI, 14 CI	NH: 6
Test environment	Unknown	Soundproof booth	Soundproof booth	Room with simulated acoustics of living room, classroom, concert hall.	Quiet office room and via VOIP.	Soundproof booth	Quiet room
ML Model's features	HMM [16] MFCCs [17]	DNN-HMM [18] MFCCs [17]	HMM [16] Weighted MFCCs	ASR: From Amazon TTS: Acapela [7]	MFCCs [19]	Time delay neural network [20] GMM-HMM [16] MFCCs [19]	Kaldi-NL
ML Model's metrics	WER: No new words: 0.66%, with new words: 22.7%	Insertion Error 2.9% Deletion Error 0.9%	Unknown	Insertion Error 1.9% Deletion Error 6.1%	WER 2.83%	Insertion Error 3% Deletion Error 0.6%	WER 5%
Method' bias	Unknown	NH: 0.5 dB HI: 0.8 dB	SRT Bias of <4.4 dB. Accuracy of 96.67% for detecting HL.	On average 1.40 dB. 0.23 dB difference in intrasubject standard deviation between reference and ASR based test	Correlation of r = 0.93. SRT is 1 dB higher with the proposed method	NH and unaided HI: 1.38 dB bias. Aided HI and CI listeners: 2.05 dB bias.	<0.7 dB with 4 errors)
Stimuli dataset	OLSA [8]	OLSA [8]	72 English spondee	OLSA [8]	OLSA [8]	OLSA [8]	Triplet digits (0-9) [21])
Training Data	27170 utterance (23.2 hours) 10 males, 10 females	Same as [10]. 18 hours of OOV words	3600 utterances	N/A	1000 hours German speech (8000 hours after noise adding and augmentation)	Same as [10]. 18 hours of speech from 40 speakers [22] [23].	1000h of Dutch speech [24]
Test-Retest reliability of proposed method	0.5 dB for the test without OOV.	NH: 0.5 dB HI: 0.9 dB	Unknown	Intrasubject standard deviation of 0.63 dB for NH and 1.01 dB for HI (This is 0.5 dB and 0.9 dB for the clinical test).	Unknown	Unknown	Unknown
Noise Generation algorithm	Unknown	Unknown	Unknown	Superimposing from test dataset	Superimposing the speech material	Superimposing the speech material	Speech shape noise
Masking Noise Level	Unknown	65 dB SPL	Unknown	Constant 65 dB SPL	Start from 0 dB	fixed at 65 dB	fixed at 65 dB
Conclusion	ASR is good for capturing the response when there are no OOV words, but not when there are OOVs.	Reliable SRT measurement can be obtained with the ASR system.	Altered MFCCs to increase the ASR accuracy for detecting SRT.	Used TTS to generate the stimuli and ASR to capture the patient response.	Used ASR to detect SRT of patients over the Internet.	Improved the baseline ASR system by using state-of-the-art models.	Used ASR for self-supervised DIN and evaluated using bootstring simulation.

NH: Normal Hearing, HI: Hearing Impaired, CI: Cochlear Implant, dB HL: Decibels at Hearing Level, dB SPL: Decibels at Sound Pressure Level, MFCC: Mel-Frequency Cepstrum Coefficients, HMM: Hidden Markov Model, GMM: Gaussian mixture model, OOV: Out-Of-Vocabulary, WER: Word Error Rate, DIN: Digits-in-Noise

TABLE III

THIS TABLE SUMMARISES THE PAPERS THAT USE ASR TO SIMULATE THE HEARING TEST PROCEDURE.

	(Fontan, 2020) [25]	(Roßbach ,2022) [26]	(Brochier , 2022) [27]
Participants' hearing levels	24 HI with >20 dB difference in low and high frequency sensitivity.	8 NH, 20 HI with mild to moderate HL	CI users
Test environment	Soundproof room	Unknown	N/A
ML Model's features	SPHINX-3 [28] MFCCs [19] HMM [16]	Model in [18] HMM [16] MFCCs [17]	DNN with GRU [29] HMM [16]
ML Model's metrics	Logatoms: 87.4% Words: 98.3% Sentences: 90.8%	Unknown	Phoneme frame error rate and confusion
Method bias	Correlation between ASR and human: Logatoms: 0.81 Words: 0.77 Sentences: 0.71	SRT bias of baseline and proposed model: NH: 1.6 dB ($p<0.01$) HI: 1.4 dB ($p<0.001$)	MSE between ASR and CI users: 0.48% for Vowels, 0.28% for consonants. High correlation for consonants (R = 0.65) but not for vowels (R = 0.38). Predicted SRTs within 1 dB of published studies with CI users.
Stimuli dataset	68 logatoms [30]. 60 nouns [31]. 40 sentences [32].	OLSA [8]	McKay [33] Munson [34] Friesen [35] Schvartz-Leyzac [36]
Training Data	31 hours of French radio recording [37]	10 hours of OLSA-like speech from 20 speakers with added noise from -10 to 20 dB SNR.	TIMIT [38]
Noise Generation algorithm	None	Speech and noise like masker.	20-talker babble
Masking Noise Level	Speech presented at a soft level of 50 dB SPL	SNR of -30 dB to 10 dB	Quiet and adaptive
Conclusion	Predict the SI of HI, using stimuli with different complexity. Reported strong correlation between the SI measured by humans and ASR.	Used DNN to create an ASR system to detect SRT of NH and HI with the presence of different types of masker noise. It should be noted that they used the same noise in both training and testing	CI users phoneme perception prediction. Good accuracy for consonants, but not for vowels.

NH: Normal Hearing, HI: Hearing Impaired, CI: Cochlear Implant, dB SPL: Decibels at Sound Pressure Level, MFCC: Mel-Frequency Cepstrum Coefficients, HMM: Hidden Markov model, DNN: Deep Neural Network, GRU: Gated recurrent unit, SI: Speech Intelligibility, MSE: Mean Square Error.

TABLE IV
THIS TABLE SUMMARISES PAPERS IN THE FIRST CLARITY CHALLENGE

Paper	ML Model's features	Method' bias (Correlation)	Intrusive	Training Data	Conclusion
(Tu, 2022) [39]	Transformer [40] CNN	0.76	Yes	LibriSpeech [41] CPC1 [42]	Measured SI by comparing the hidden representation of generated sound with the reference.
(Tu, 2022) [43]	Transformer [40]	0.73	No	LibriSpeech [41] CPC1 [42] CEC1 [44]	Used deep ensemble [45] to measure the uncertainty of the ASR model as a representation of SI.
(Mawalim, 2022) [46]	Time delay neural networks [20]	0.67	Yes	LibriSpeech [41] CPC1 [42]	Used ASR to determine the difficulty of understanding the presented speech.
(Zezario, 2022) [47]	MOSA-Net [48] WavLM [49]	0.65	No	CPC1 [42]	Trained a multi-branch speech intelligibility prediction to convert features extracted by WavLM to SI score.
(Kamo, 2022) [50]	CNN Conformer [51]	0.60	Yes	LibriSpeech [41] CPC1 [42]	Ensemble of 10 neural network models. Used Conformer model to fuse different extracted features.
(Roßbach, 2022) [52]	LEAP [53] Time delay neural networks [20] MFCCs [17] HMM [16]	0.54	No	CPC1 [42] 8k hours of in-house German dataset.	Adopted LEAP [53] for people with hearing loss. And created a mapping from the model output to the SI score.

SI: Speech Intelligibility, MFCC: Mel-Frequency Cepstrum Coefficients, HMM: Hidden Markov model, CNN: Convolutional Neural Network

TABLE V
THIS TABLE SUMMARISES PAPERS IN THE SECOND CLARITY CHALLENGE

Paper	ML Model's features	Method' bias (Correlation)	Intrusive	Training Data	Conclusion
(Cuervo, 2023) [54]	Whisper [55] WavLM [49] Transformers [40]	0.78	No	CPC2 [56]	Used pretrained large foundation models to extract features from speech and used transformers to combine extracted features.
(Huckvale, 2023) [57]	Wav2Vec [58]	0.78	Yes	CPC2 [56] Cambridge Read News [59]	Combined speech features with the metadata provided in the challenge using a neural network model.
(Mogridge, 2023) [60]	Whisper [55] Bi-LSTM [61]	0.77	No	CPC2 [56]	Used Whisper as an audio feature extractor and then an ensemble of two Bi-LSTM models [61] for making the prediction.
(Tu, 2023) [62]	Transformers [40]	0.77	Intrusive and Non-intrusive	Simulated noisy LibriSpeech [41]	Proposed a non-intrusive system that converts entropy to SI and an intrusive system that compares the generated features of reference and enhanced signal.
(Zezario, 2023) [48]	Whisper [55] CNN Bi-LSTM [61] Attention	0.76	No	CPC2 [56]	Used Whisper to extract features from speech and used a neural network to convert features to SI.

SI: Speech Intelligibility, Bi-LSTM: Bidirectional Long Short-term Memory, CNN: Convolutional Neural Network

TABLE VI
THIS TABLE SUMMARISES THE PAPERS THAT USE ASR TO CONFIGURE HEARING AIDS.

	(Fontan, 2020) [63]	(Gonçalves, 2022) [64]	(Fontan, 2022) [65]
Participants' hearing level	24 HI with HL <75 dB HL	12 HI	12 HI
Test environments	Soundproof room	Unknown	Unknown
ML Model's features	SPHINX-3 [28] MFCCs [19] HMM [16]	ASR engine Julius 4.4.2 [66] GMM HMM [16]	ASR engine Julius 4.4.2 [66] GMM HMM [16]
ML Model's metrics	Logatoms: 87.4%, Words: 98.3%, Sentences: 90.8%	Unknown	Unknown
Method' bias	ASR based configuration had a higher intelligibility compared to CAM2 with $p<0.001$ and $p=0.002$ for words and sentences. (98.25% compared to 96.5%). ASR based configuration had a higher speech pleasantness compared to CAM2 with $p<0.001$ and $p=0.002$ for words and sentences. (8.4 compared to 7 out of 10).	98% ASR intangibility score for ASR configured HA compared to 88% for CAM2 (P=0.002).	92% ASR intangibility score for ASR configured HA compared to 88% for CAM2. No improvement over [64].
Stimuli dataset	60 disyllabic nouns [31]. 40 sentences from French hearing in noise test [32].	French disyllabic nouns [31].	French disyllabic nouns [31].
Training Data	31 hours of French radio broadcast recording [37].	ESTER [67] and ESTER2 [37] ("100 hours French radio recording").	ESTER [67] and ESTER2 [37] ("100 hours French radio recording").
Test-Retest reliability of proposed method	Unknown	Pearson correlation of >0.95 for 12 repetitions which shows a low test-retest standard deviation.	P-value=0.1 Based on two repetitions (i.e., no significant difference exists)
Stimuli level	60 dB SPL	65 and 85 dB SPL	65 and 85 dB SPL
Configuration search algorithm	625 predetermined gain functions.	Three genetic algorithms [68]	The best search algorithm of [64].
Searched parameters	Insertion Gain	Insertion gains in five frequencies (step size of 0.1 dB and ± 10 dB to the prescribed insertion gain). Compression threshold (step size of 1 dB, range between 20 to 50 dB SPL).	Attack (100 to 500 ms) and release (300 to 2000 ms) time with 10 ms step size.
Conclusion	Used ASR to evaluate different configurations of HA and find a better optimisation for the prescribed HA.	Increased the search space for the insertion gain by using genetic algorithms.	Extended the work of [64] and optimised attach and release time of their HA prescription.

NH: Normal Hearing, HI: Hearing Impaired, dB HL: Decibels at Hearing Level, dB SPL: Decibels at Sound Pressure Level, MFCC: Mel-Frequency Cepstrum Coefficients, HMM: Hidden Markov Model, DNN: Deep Neural Network, GMM: Gaussian mixture model

REFERENCES

[1] M. Kosai, J. Kohda, J. Udaka, and Y. Koike, "Speech audiometric trial using synthetic vowels produced with DECtalk," *Med. Inform. (Lond)*, vol. 15, no. 4, pp. 309–318, Jan. 1990.

[2] T. Nuesse, B. Wiercinski, T. Brand, and I. Holube, "Measuring Speech Recognition With a Matrix Test Using Synthetic Speech," *Trends in Hearing*, vol. 23, p. 233121651986298, Jan. 2019.

[3] J. Ooster, M. Krueger, J.-H. Bach, K. C. Wagener, B. Kollmeier, and B. T. Meyer, "Speech Audiometry at Home: Automated Listening Tests via Smart Speakers With Normal-Hearing and Hearing-Impaired Listeners," *Trends in Hearing*, vol. 24, p. 233121652097001, Jan. 2020.

[4] S. Ibelings, T. Brand, and I. Holube, "Speech Recognition and Listening Effort of Meaningful Sentences Using Synthetic Speech," *Trends in Hearing*, vol. 26, p. 233121652211306, Jan. 2022.

[5] S. Polspoel, D. R. Moore, D. W. Swanepoel, S. E. Kramer, and C. Smits, "Global access to speech hearing tests," Jun. 2024.

[6] S. Lock and C. K. Leong, "Program library for DECtalk text-to-speech system," *Beh. Res. Meth. Instr. Comp.*, vol. 21, no. 3, pp. 394–400, May 1989.

[7] "Acapela," https://www.acapela-group.com/.

[8] K. Wagener, V. Kühnel, and B. Kollmeier, "Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test," *Z. Audiol.*, vol. 38, pp. 4–15, 1999.

[9] B. Kollmeier and M. Wesselkamp, "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2412–2421, Oct. 1997.

[10] B. T. Meyer, B. Kollmeier, and J. Ooster, "Autonomous measurement of speech intelligibility utilizing automatic speech recognition," in *Interspeech 2015*. ISCA, Sep. 2015, pp. 2982–2986.

[11] J. Ooster, R. Huber, B. Kollmeier, and B. T. Meyer, "Evaluation of an automated speech-controlled listening test with spontaneous and read responses," *Speech Communication*, vol. 98, pp. 85–94, Apr. 2018.

[12] S. Nisar, M. Tariq, A. Adeel, M. Gogate, and A. Hussain, "Cognitively Inspired Feature Extraction and Speech Recognition for Automated Hearing Loss Testing," *Cogn Comput*, vol. 11, no. 4, pp. 489–502, Feb. 2019.

[13] T. Bruns, J. Ooster, M. Stennes, and J. Rennies, "Automated Speech Audiometry for Integrated Voice Over Internet Protocol Communication Services," *Am J Audiol*, vol. 31, no. 3S, pp. 980–992, Sep. 2022.

[14] J. Ooster, L. Tuschen, and B. T. Meyer, "Self-conducted speech audiometry using automatic speech recognition: Simulation results for listeners with hearing loss," *Computer Speech & Language*, vol. 78, p. 101447, Mar. 2023.

[15] G. Araiza-Illan, L. Meyer, K. P. Truong, and D. Başkent, "Automated Speech Audiometry: Can It Work Using Open-Source Pre-Trained Kaldi-NL Automatic Speech Recognition?" *Trends in Hearing*, vol. 28, p. 23312165241229057, Jan. 2024.

[16] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb./1989.

[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[18] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-Discriminative Training of Deep Neural Networks," in *Interspeech*, 2013, pp. 2345–2349.

[19] C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi, "Mel-frequency cepstral coefficient analysis in speech recognition," in *2006 Int. Conf. Comput. Inform.* Kuala Lumpur, Malaysia: IEEE, Jun. 2006, pp. 1–5.

[20] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low Latency Acoustic Modeling Using Temporal Convolution and LSTMs," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 373–377, Mar. 2018.

[21] C. Smits, S. Theo Goverts, and J. M. Festen, "The digits-in-noise test: Assessing auditory speech recognition abilities in noise," *J. Acoust. Soc. Am.*, vol. 133, no. 3, pp. 1693–1706, Mar. 2013.

[22] "King-ASR-L-092," http://en.htrs.demoboc.cn/datacenter/searchdetails/288.html.

[23] "King-ASR-L-182," http://en.htrs.demoboc.cn/datacenter/searchdetails/361.html.

[24] N. Oostdijk *et al.*, "The spoken dutch corpus. Overview and first evaluation." in *LREC*. Athens, Greece, 2000, pp. 887–894.

[25] L. Fontan, T. Cretin-Maitenaz, and C. Füllgrabe, "Predicting Speech Perception in Older Listeners with Sensorineural Hearing Loss Using Automatic Speech Recognition," *Trends in Hearing*, vol. 24, p. 233121652091476, Jan. 2020.

[26] J. Roßbach, B. Kollmeier, and B. T. Meyer, "A model of speech recognition for hearing-impaired listeners based on deep learning," *J. Acoust. Soc. Am.*, vol. 151, no. 3, pp. 1417–1427, Mar. 2022.

[27] T. Brochier, J. Schlittenlacher, I. Roberts, T. Goehring, C. Jiang, D. Vickers, and M. Bance, "From Microphone to Phoneme: An End-to-End Computational Neural Model for Predicting Speech Perception With Cochlear Implants," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 11, pp. 3300–3312, Nov. 2022.

[28] K. Seymore, S. Chen, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, and E. Thayer, "The 1997 CMU Sphinx-3 English Broadcast News Transcription System," in *Proc. 1998 DARPA Speech Recognit. Workshop*, 1998.

[29] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Sep. 2014.

[30] L. Dodelé and D. Dodelé, "L'audiométrie vocale en présence de bruit et filetest AVfB," *Cah Audit.*, vol. 13, no. 6, pp. 15–22, 2000.

[31] J.-E. Fournier, *Audiométrie Vocale: Les Épreuves d'intelligibilité et Leurs Applications Au Diagnostic, à l'expertise et à La Correction Prothétique Des Surdités*. Maloine, 1951.

[32] V. Vaillancourt, C. Laroche, C. Mayer, C. Basque, M. Nali, A. Eriks-Brophy, S. D. Soli, and C. Giguère, "Adaptation of the hint (hearing in noise test) for adult canadian francophone populations: Adaptación del hint (prueba de audición en ruido) para poblaciones de adultos canadienses francófonos," *Int. J. Audiol.*, vol. 44, no. 6, pp. 358–361, 2005.

[33] C. M. McKay and H. J. McDermott, "Perceptual performance of subjects with cochlear implants using the Spectral Maxima Sound Processor (SMSP) and the Mini Speech Processor (MSP)," *Ear Hear.*, vol. 14, no. 5, pp. 350–367, Oct. 1993.

[34] B. Munson, G. S. Donaldson, S. L. Allen, E. A. Collison, and D. A. Nelson, "Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability," *J. Acoust. Soc. Am.*, vol. 113, no. 2, pp. 925–935, Feb. 2003.

[35] L. M. Friesen, R. V. Shannon, D. Baskent, and X. Wang, "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1150–1163, Aug. 2001.

[36] K. C. Schvartz-Leyzac, T. A. Zwolan, and B. E. Pfingst, "Effects of electrode deactivation on speech recognition in multichannel cochlear implant recipients," *Cochlear Implants International*, vol. 18, no. 6, pp. 324–334, Nov. 2017.

[37] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Interspeech 2009*. ISCA, Sep. 2009, pp. 2583–2586.

[38] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, Aug. 1990.

[39] Z. Tu, N. Ma, and J. Barker, "Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-impaired Listeners," Jul. 2022.

[40] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," in *2018 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*. Calgary, AB: IEEE, Apr. 2018, pp. 5884–5888.

[41] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*, Apr. 2015, pp. 5206–5210.

[42] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. V. Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 3508–3512.

[43] Z. Tu, N. Ma, and J. Barker, "Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-intrusive Speech Intelligibility Prediction," Jul. 2022.

[44] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 686–690.

[45] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," Nov. 2017.

[46] C. O. Mawalim, B. A. Titalim, and M. Unoki, "CPC1 E031 system description," in *Proc. 2nd Clarity Workshop Mach. Learn. Chall. Hear. Aids Clarity-2022 Online*, 2022.

[47] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids," Aug. 2022.

[48] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep Learning-Based Non-Intrusive Multi-Objective Speech Assessment Model With Cross-Domain Features," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 54–70, 2023.

[49] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[50] N. Kamo, K. Arai, A. Ogawa, S. Araki, T. Nakatani, K. Kinoshita, M. Delcroix, T. Ochiai, and T. Irino, "Conformer-based fusion of text, audio, and listener characteristics for predicting speech intelligibility of hearing aid users," in *Proc. 2nd Clarity Workshop Mach. Learn. Chall. Hear. Aids*, 2022.

[51] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," May 2020.

[52] J. Roßbach, R. Huber, S. Röttges, C. F. Hauth, T. Biberger, T. Brand, B. T. Meyer, and J. Rennies, "Speech intelligibility prediction for hearing-impaired listeners with the LEAP model." in *INTERSPEECH*, 2022, pp. 3498–3502.

[53] R. Huber, A. Pusch, N. Moritz, J. Rennies, H. Schepker, and B. T. Meyer, "Objective assessment of a speech enhancement scheme with an automatic speech recognition-based system," in *Speech Commun. 13th ITG-Symp.* VDE, 2018, pp. 1–5.

[54] S. Cuervo and R. Marxer, "Temporal-hierarchical features from noise-robust speech foundation models for non-intrusive intelligibility prediction," in *Proc. ISCA Clarity-2023*, 2023.

[55] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," in *INTERSPEECH 2023*, Aug. 2023, pp. 2798–2802.

[56] J. Barker, M. A. Akeroyd, W. Bailey, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor, "The 2nd Clarity Prediction Challenge: A Machine Learning Challenge for Hearing Aid Intelligibility Prediction," in *ICASSP 2024 - 2024 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*, Apr. 2024, pp. 11 551–11 555.

[57] M. Huckvale and G. Hilkhuysen, "Combining acoustic phonetic linguistic and audiometric data in an intrusive intelligibility metric for hearing-impaired listeners," in *Proc. ISCA Clarity-2023*, 2023.

[58] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Adv. Neural Inf. Process. Syst.*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[59] Robinson, Tony, Fransen, Jeroen, Pye, David, Foote, Jonathan, Renals, Steve, Woodland, Phil, and Young, Steve, "WSJCAM0 Cambridge Read News," p. 3670016 KB, 1995.

[60] R. Mogridge, G. Close, R. Sutherland, S. Goetze, and A. Ragni, "Pre-trained intermediate ASR features and Human memory simulation for non-intrusive speech intelligibility prediction in the Clarity Prediction Challenge 2," in *Proc. ISCA Clarity-2023*, 2023.

[61] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov./1997.

[62] Z. Tu, N. Ma, and J. Barker, "Intelligibility prediction with a pretrained noise-robust automatic speech recognition model," Oct. 2023.

[63] L. Fontan, M. Le Coz, C. Azzopardi, M. A. Stone, and C. Füllgrabe, "Improving hearing-aid gains based on automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 148, no. 3, pp. EL227–EL233, Sep. 2020.

[64] L. Gonçalves Braz, L. Fontan, J. Pinquier, M. A. Stone, and C. Füllgrabe, "OPRA-RS: A Hearing-Aid Fitting Method Based on Automatic Speech Recognition and Random Search," *Front. Neurosci.*, vol. 16, p. 779048, Feb. 2022.

[65] L. Fontan, L. Gonçalves Braz, J. Pinquier, M. A. Stone, and C. Füllgrabe, "Using Automatic Speech Recognition to Optimize Hearing-Aid Time Constants," *Front. Neurosci.*, vol. 16, 2022.

[66] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proc. APSIPA ASC 2009 Asia-Pac. Signal Inf. Process. Assoc. 2009 Annu. Summit Conf.* Asia-Pacific Signal and Information Processing Association, 2009, pp. 131–137.

[67] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Proc. Fifth Int. Conf. Lang. Resour. Eval. LREC06*. Genoa, Italy: European Language Resources Association (ELRA), May 2006.

[68] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: Past, present, and future," *Multimed. Tools Appl.*, vol. 80, no. 5, pp. 8091–8126, Feb. 2021.