International Journal of Audiology

ISSN: 1499-2027 (Print) 1708-8186 (Online) Journal homepage: www.tandfonline.com/journals/iija20

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 & Kevin J. Munro

To cite this article: Mohsen Fatehifar, Josef Schlittenlacher, Ibrahim Almufarrij, David Wong, Tim Cootes & Kevin J. Munro (2025) Applications of automatic speech recognition and text-to-speech technologies for hearing assessment: a scoping review, International Journal of Audiology, 64:6, 537-548, DOI: 10.1080/14992027.2024.2422390

To link to this article: https://doi.org/10.1080/14992027.2024.2422390

INTERNATIONAL JOURNAL OF AUDIOLOGY
2025, VOL. 64, NO. 6, 537–548
https://doi.org/10.1080/14992027.2024.2422390

The British Society of Audiology
The International Society of Audiology
NAS NORDIC AUDIOLOGICAL SOCIETY
Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

OPEN ACCESS Check for updates

Applications of automatic speech recognition and text-to-speech technologies for hearing assessment: a scoping review

Mohsen Fatehifar[a], Josef Schlittenlacher[b], Ibrahim Almufarrij[a,c], David Wong[d], Tim Cootes[e] and Kevin J. Munro[a,f]

[a]Manchester Centre for Audiology and Deafness (ManCAD), School of Health Sciences, University of Manchester, Manchester, UK; [b]Department of Speech, Hearing and Phonetic Sciences, University College London, London, UK; [c]Department of Rehabilitation Sciences, College of Applied Medical Sciences, King Saud University, Riyadh, Saudi Arabia; [d]Leeds Institute of Health Sciences, University of Leeds, Leeds, UK; [e]Centre for Imaging Sciences, University of Manchester, Manchester, UK; [f]Manchester Academic Health Science Centre, Manchester University Hospitals NHS Foundation Trust, Manchester, UK

ABSTRACT

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

Design: Review 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.

ARTICLE HISTORY
Received 18 December 2023
Revised 18 October 2024
Accepted 23 October 2024

KEYWORDS
Automatic speech recognition; hearing assessment; hearing test; hearing aid; hearing in noise test; speech in noise; text to speech

Introduction

According to the World Health Organisation ("World Report on Hearing" 2021), 1.5 billion individuals have hearing loss, with 430 million requiring intervention, the most common of which is the provision of hearing aids (Blazer and Tucci 2019). 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 (Griffiths et al. 2020).

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 (Fagan and Jacobs 2009). Even in developed nations, access to these facilities can prove challenging especially in rural areas, and for elderly or infirm individuals (Planey 2019). 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 (Grasselli et al. 2020; Centers for Disease Control and Prevention 2020).

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 (Hustad and Cahill 2003; Sorin and Thouin-Daniel 1983). Research shows that most people don't come forward to be assessed for hearing aids. Additionally, about half of hearing aids users do not wear them often (Dillon et al. 2020) and poorly fit hearing aid is one of the factors contributing to this problem (McCormack and Fortnum 2013).

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

CONTACT Mohsen Fatehifar mohsen.fatehifar@manchester.ac.uk Manchester Centre for Audiology and Deafness (ManCAD), School of Health Sciences, University of Manchester, Manchester, UK

Supplemental data for this article can be accessed online at https://doi.org/10.1080/14992027.2024.2422390.

progress has been made on this topic (Whitton et al. 2016) and there is evidence that these methods can produce accurate outcomes without human supervision (Almufarrij et al. 2023).

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 (Qayyum et al. 2021). 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 (Cox and de Vries 2021).
- Classifiers that can detect the type and degree of hearing loss from clinical hearing tests (pure tone audiograms) (Taylor and Sheikh 2022).
- Analysing EEG response signals to acoustic stimuli to detect if the person heard the stimuli (Osman and Osman 2021).
- Self-administered speech intelligibility tests using automatic speech recognition (ASR) or text-to-speech techniques (TTS) (Ibelings, Brand, and Holube 2022; Ooster, Tuschen, and Meyer 2023).

The use of ASR and TTS (see Table 1 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.

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. (2022) reviewed automated assessments of hearing but no studies that used ASR or TTS were investigated in their review. Osman and Osman (2021) 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. (2023) 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.

Method

The protocol (inplasy.com/inplasy-2023-1-0029) was submitted to the International Platform of Registered Systematic Review and Meta-Analysis Protocols (Fatehifar et al. 2023) and the review was carried out in accordance with PRISMA scoping review guidelines (Tricco et al. 2018).

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 2 for the complete inclusion and exclusion criteria.

Information sources

Relevant studies were identified through a systematic literature search that was conducted in January of 2023 and updated in June of 2024 in the following electronic databases and preprint

Table 1. 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 (Van and Yanz 1987).
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).

Table 2. Inclusion and exclusion criteria.

Inclusion	TTS used to convert text to acoustic test stimuli.
	ASR used to capture participants' verbal responses.
	ASR used to optimise 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.

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.

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.

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.

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 Figure 1.

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.

Results

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

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 (Ibelings, Brand, and Holube 2022; Kosai et al. 1990; Nuesse et al. 2019; Ooster et al. 2020; Polspoel et al. 2024). 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 Figure 2 and the extracted data is provided in Table 1 of the supplementary material.

The first research (Kosai et al. 1990) on TTS for stimuli generation was published in 1990. The researchers used the DECtalk program (Lock and Leong 1989) 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. (2019) used a commercial TTS system

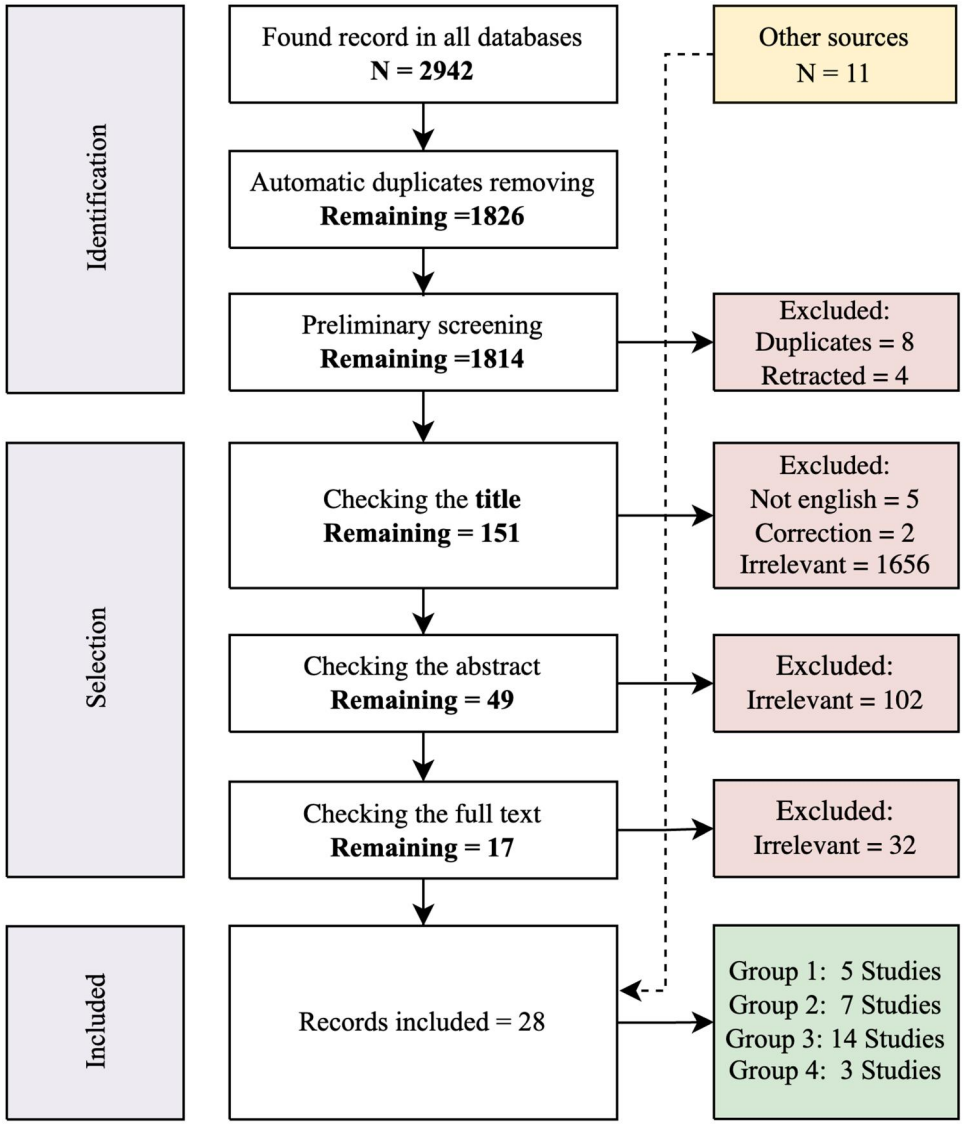

Figure 1. Article selection process. Based on how ASR and TTS were used, the studies were categorised into four groups (Group 1: TTS for generating the acoustic stimuli, Group 2: ASR for capturing the verbal response, Group 3: ASR for estimating speech test performance), Group 4: ASR for configuration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

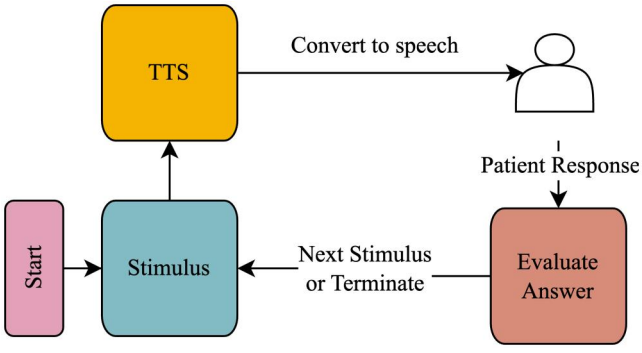

Figure 2. Diagram showing how TTS is used in speech intelligibility tests. A synthetic stimulus is presented and the participant is asked to repeat the stimulus they hear. This procedure is repeated until the predefined stop condition is satisfied.

developed by the Acapela group ("Acapela" 2024) that used a non-uniform unit selection (Bellegarda 2007) for synthesising the German matrix sentence (OLSA) dataset (Wagener, Kühnel, and Kollmeier 1999). 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 (Wagener, Kühnel, and Kollmeier 1999). 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, Brand, and Holube (2022) used a new TTS system from the Acapela group ("Acapela" 2024) to synthesise another German dataset (GöSa (Kollmeier and Wesselkamp 1997)), 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

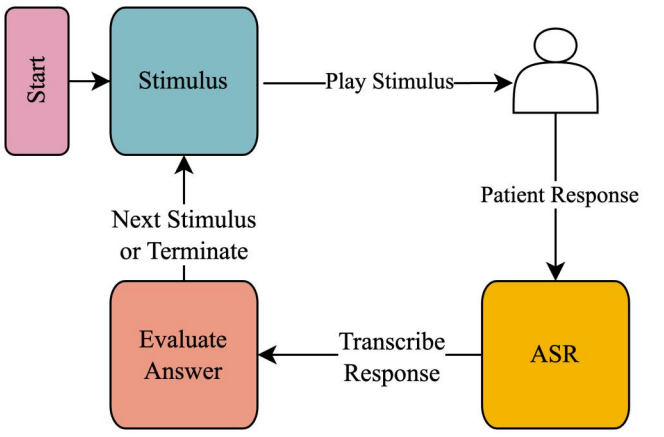

Figure 3. Diagram showing how speech intelligibility tests with ASR works. In this diagram, the ASR transcribe the participant's response and the evaluation of the response is done with the transcribed text.

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. (2020) 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. (2019) and a pretrained ASR system from Amazon. To evaluate the method, OLSA (Wagener, Kühnel, and Kollmeier 1999) 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. (2024) 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. (1990), 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. (2019) and Ibelings, Brand, and Holube (2022) 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 (Ibelings, Brand, and Holube 2022; Nuesse et al. 2019; Polspoel et al. 2024) 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 (Willberg et al. 2020; Schlueter et al. 2016).

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 (Ooster, Tuschen, and Meyer 2023; Ooster et al. 2020; Meyer, Kollmeier, and Ooster 2015; Ooster et al. 2018; Nisar et al. 2019; Bruns et al. 2022; Araiza-Illan et al. 2024). 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 Figure 3 and the extraction table is provided in Table 2 of the supplementary materials.

In 2015, Meyer, Kollmeier, and Ooster (2015) built an ASR system with a Hidden Markov Model (HMM) (Rabiner 1989) and Mel-frequency cepstrum coefficients(MFCC) (Davis and Mermelstein 1980) trained on a dataset of 23.2 hours of speech. To assess their method, they used the OLSA dataset (Wagener, Kühnel, and Kollmeier 1999) 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 (Ooster et al. 2018) built upon their own previous study (Meyer, Kollmeier, and Ooster 2015). They improved the training dataset of the ASR 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 (Wagener, Kühnel, and Kollmeier 1999). 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 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. (2020) 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, Tuschen, and Meyer (2023) aimed to enhance the accuracy of the ASR system by using a new time-delay neural network (Peddinti et al. 2018) with MFCCs (Davis and Mermelstein 1980) and an HMM (Rabiner 1989). 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, Kollmeier, and Ooster (2015) and another dataset with 18 hours of speech ("King-ASR-L-092", "King-ASR-L-182). 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 (Wagener, Kühnel, and Kollmeier 1999). 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. (2019). 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. (2022) developed a fully remote speech intelligibility test. To implement the ASR system, they adopted the model proposed by Peddinti et al. (2018), which used a deep neural network with MFCC feature extractor (Davis and Mermelstein 1980) 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 (Wagener, Kühnel, and Kollmeier 1999) 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. (2024) proposed a self-supervised digits-in-noise test using an ASR system trained on 1000 hours of Dutch speech (Oostdijk, et al. 2000). 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 (Ooster et al. 2020) used commercial ASRs, while others (85%) trained their model using HMM (Rabiner 1989) and MFCCs (Davis and Mermelstein 1980) based model. Nisar et al. (2019) 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 (Wagener, Kühnel, and Kollmeier 1999) as the test stimuli and reported SRT bias of approximately 1 dB. However, Meyer, Kollmeier, and Ooster (2015) did not compare their method with the clinical SRT. Nisar et al. (2019) 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%).

ASR for estimating speech test performance

The 14 (50%) studies in this category predicted speech intelligibility (Fontan, Le Coz, et al. 2020; Roßbach, Kollmeier, et al. 2022; Brochier et al. 2022; Tu et al., 2022a; Mawalim, Titalim, and Unoki 2022; Tu et al., 2022b; Zezario et al. 2022; Kamo et al. 2022; Roßbach, Kollmeier, et al. 2022; Cuervo and Marxer 2023; Huckvale and Hilkhuysen 2023; Mogridge et al. 2023; Tu, Ma, and Barker 2023; Zezario et al. 2023). 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 into various situations that can affect speech intelligibility. The overall flow of simulating speech intelligibility tests with ASR is presented in Figure 4 and the extraction table is provided in Tables 3, 4 and 5 of the supplementary materials.

Fontan, Le Coz, et al. (2020) trained the ASR model using SPHINX-3 (Seymore et al. 1998) on 31 hours of French radio broadcast recordings (Galliano, Gravier, and Chaubard 2009). They evaluated the ASR system on three types of inputs; pseudo-word (Dodelé and Dodelé 2000) (87.4% accuracy), words (Fournier 1951) (98.3% accuracy) and sentences (Vaillancourt et al. 2005) (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 (Nejime and Moore 1997) 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, Kollmeier, et al. (2022). They trained a deep learning-based ASR system based on the architecture proposed by (Vesely et al. 2013) on 10 hours of speech from 20 speakers in the shape of the OLSA dataset (Wagener, Kühnel, and Kollmeier 1999). 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 (Schubotz et al. 2016). The hearing loss simulation was done by

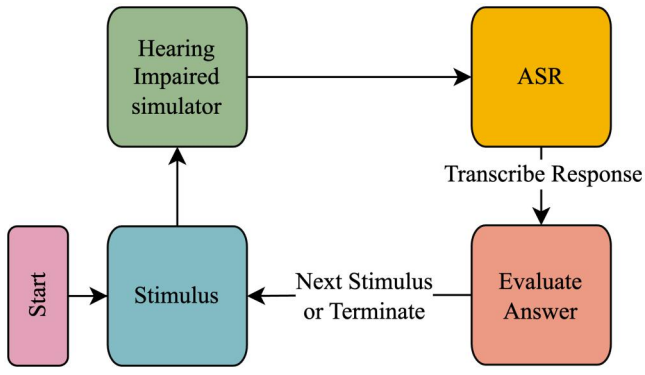

Figure 4. Diagram showing how ASR is used to simulate speech intelligibility tests. In this system, an ASR and a hearing loss simulator replace a person with hearing loss, and it should achieve a result close to the person it is modelling.

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. (2022) 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 (McKay and McDermott 1993; Munson et al. 2003) 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) (Barker et al. 2022; Barker et al. 2024). 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 (Barker et al. 2022) 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 4 and 5 of the supplementary materials.

In CPC1, intrusive systems, on average, performed better as they had access to the clean reference data (Tu et al., 2022a; Mawalim, Titalim, and Unoki 2022; Kamo et al. 2022). The winner of the CPC1 (Huckvale and Hilkhuysen 2022) was an intrusive system, but they did not use ASR. The best ASR based system (Tu et al., 2022a) 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 (Huckvale and Hilkhuysen 2022) by using Wav2Vec (Baevski et al. 2020) and fine-tuning it on the Cambridge read news dataset (Robinson et al. 1995), achieving a correlation score of 0.78. Tu, Ma, and Barker (2023) also extended on previous entries to CPC1 (Tu et al., 2022a; Tu et al., 2022b). 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 (Cuervo and Marxer 2023) of this challenge was a non-intrusive system that used pre-trained WavLM (Chen et al. 2022) and Whisper (Gong et al. 2023) 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 (Gong et al. 2023) to extract features from the input speech and use another machine learning algorithm to map the extracted features to intelligibility score (Mogridge et al. 2023; Zezario et al. 2023).

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%) (Brochier et al. 2022) 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 byRoßbach, Kollmeier, et al. (2022) 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.

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 (Fontan, Le Coz, et al. 2020; Gonçalves Braz et al. 2022; Fontan et al. 2022). 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 Figure 5 and the extraction table is provided in Table 6 of the supplementary materials.

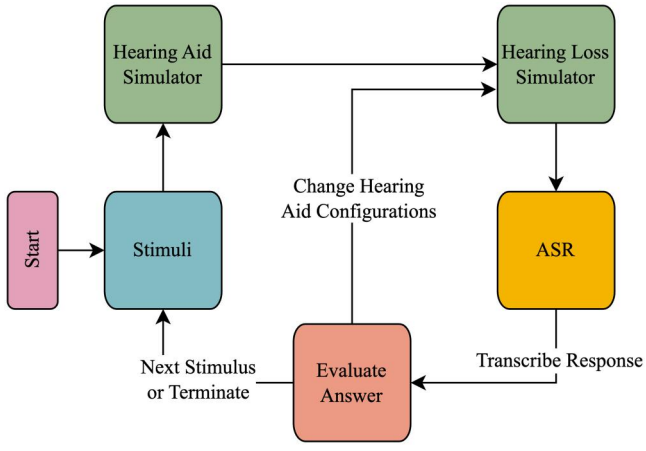

Figure 5. Diagram showing how ASR is used to fit a hearing aid. In this diagram, an ASR is used to evaluate a particular hearing aid configuration and change the parameters to achieve the best results.

Fontan, Le Coz, et al. (2020) used the SPHINX-3 (Seymore et al. 1998) ASR system and trained it on 31 hours of French radio broadcast recordings (Galliano, Gravier, and Chaubard 2009). 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 (Moore, Glasberg, and Stone 2010), 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 (Fournier 1951) and 40 sentences from the French hearing in noise test (Vaillancourt et al. 2005) 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 (Moore, Füllgrabe, and Stone 2011).

Gonçalves Braz et al. (2022) expanded Fontan's work by using genetic algorithms (Katoch, Chauhan, and Kumar 2021) 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 (Lee and Kawahara 2009) and trained it on 100 hours of French radio broadcasting recordings (Galliano, Gravier, and Chaubard 2009; Galliano et al. 2006). To evaluate the model, they used 60 disyllabic nouns (Fournier 1951) 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 (Moore, Glasberg, and Stone 2010). 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 (Fontan et al. 2022) was a continuation of their previous research (Gonçalves Braz et al. 2022) 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 (Gonçalves Braz et al. 2022). While they reported an increase in the ASR intelligibility score compared to CAM2 (Moore, Glasberg, and Stone 2010) fitted hearing aid (92% compared to 88%), the results showed no improvement from the configuration obtained from (Gonçalves Braz et al. 2022). 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 (Nejime and Moore 1997) 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 (Vikhar 2016) to find the optimal values faster and by doing so, they were able to expand their search space.

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.

With the exception of the Clarity challenge, the dominant language 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 seven 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.

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 pre-recorded speech. The current studies are mostly on German datasets with a limited vocabulary (OLSA) and one study synthesised English and Dutch digits (Polspoel et al. 2024). 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 (Willberg et al. 2020). 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.

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 (Brand and Kollmeier 2002) and 0.9 dB in hearing impaired (Wagener and Brand 2005) 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 investigating 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.

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, Le Coz, et al. (2020) using French stimuli of varying linguistic complexity to predict word identification scores. Brochier et al. (2022) 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 (Gong et al. 2023) 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 (Zhou et al. 2023; Mawalim et al. 2023; Mogridge et al. 2024) 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.

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, Le Coz, et al. (2020) also used a French speech-in-noise test (Vaillancourt et al. 2005). 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 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, Nielsen, and Larsen (2014) 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.

Conclusion

ASR and TTS have been used 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.

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.

Disclosure statement

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

Funding

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.

ORCID

Mohsen Fatehifar http://orcid.org/0000-0001-6256-490X
Josef Schlittenlacher http://orcid.org/0000-0002-3350-3355
Ibrahim Almufarrij http://orcid.org/0000-0003-4043-7234
David Wong http://orcid.org/0000-0001-8117-9193
Tim Cootes http://orcid.org/0000-0002-2695-9063
Kevin J. Munro http://orcid.org/0000-0001-6543-9098

References

"Acapela," 2024. https://www.acapela-group.com/.

Almufarrij, I., H. Dillon, P. Dawes, D. R. Moore, W. Yeung, A.-P. Charalambous, C. Thodi, and K. J. Munro. 2023. "Web- and App-Based Tools for Remote Hearing Assessment: A Scoping Review." *International Journal of Audiology* 62 (8):699–712. https://doi.org/10.1080/14992027.2022.2075798.

Araiza-Illan, G., L. Meyer, K. P. Truong, and D. Başkent. 2024. "Automated Speech Audiometry: Can It Work Using Open-Source Pre-Trained Kaldi-NL Automatic Speech Recognition?" *Trends in Hearing* 28: 23312165241229057. https://doi.org/10.1177/23312165241229057.

Baevski, A., Y. Zhou, A. Mohamed, and M. Auli. 2020. "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." *Advances in Neural Information Processing Systems* 33:12 449–12 460.

Barker, J., M. A. Akeroyd, W. Bailey, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor. 2024. "The 2nd Clarity Prediction Challenge: A Machine Learning Challenge for Hearing Aid Intelligibility Prediction." In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ICASSP, April, 11 551–11 555.

Barker, J., M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, et al. 2022. "The 1st Clarity Prediction Challenge: A Machine Learning Challenge for Hearing Aid Intelligibility Prediction." In Interspeech 2022. ISCA, September, 3508–3512. https://doi.org/10.21437/Interspeech.2022-10821.

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

Blazer, D. G., and D. L. Tucci. 2019. "Hearing Loss and Psychiatric Disorders: A Review." *Psychological Medicine* 49 (6):891–897. Apr. https://doi.org/10.1017/S0033291718003409.

Brand, T., and B. Kollmeier. 2002. "Efficient Adaptive Procedures for Threshold and Concurrent Slope Estimates for Psychophysics and Speech Intelligibility Tests." *The Journal of the Acoustical Society of America* 111 (6):2801–2810. https://doi.org/10.1121/1.1479152.

Brochier, T., J. Schlittenlacher, I. Roberts, T. Goehring, C. Jiang, D. Vickers, and M. Bance. 2022. "From Microphone to Phoneme: An End-to-End Computational Neural Model for Predicting Speech Perception with Cochlear Implants." *IEEE Transactions on Bio-Medical Engineering* 69 (11):3300–3312. https://doi.org/10.1109/TBME.2022.3167113.

Bruns, T., J. Ooster, M. Stennes, and J. Rennies. 2022. "Automated Speech Audiometry for Integrated Voice Over Internet Protocol Communication Services." *American Journal of Audiology* 31 (3S):980–992. https://doi.org/10.1044/2022_AJA-21-00217.

Centers for Disease Control and Prevention 2020., "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)."

Chen, S., C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al. 2022. "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing." *IEEE Journal of Selected Topics in Signal Processing* 16 (6):1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113.

Cox, M., and B. de Vries. 2021. "Bayesian Pure-Tone Audiometry Through Active Learning Under Informed Priors." *Frontiers in Digital Health* 3: 723348. https://doi.org/10.3389/fdgth.2021.723348.

Cuervo, S., and R. Marxer. 2023. "Temporal-Hierarchical Features from Noise-Robust Speech Foundation Models for Non-Intrusive Intelligibility Prediction." In Proc. ISCA Clarity-2023.

Davis, S., and P. Mermelstein. 1980. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28 (4):357–366. https://doi.org/10.1109/TASSP.1980.1163420.

Dillon, H., J. Day, S. Bant, and K. J. Munro. 2020. "Adoption, Use and Non-Use of Hearing Aids: A Robust Estimate Based on Welsh National Survey Statistics." *International Journal of Audiology* 59 (8):567–573. https://doi.org/10.1080/14992027.2020.1773550.

Dodelé, L., and D. Dodelé. 2000. "L'audiométrie vocale en présence de bruit et filetest AVfB." *Cahiers de l'Audition* 13 (6):15–22.

Fagan, J., and M. Jacobs. 2009. "Survey of ENT Services in Africa: Need for a Comprehensive Intervention." *Global Health Action* 2 (1):1932. https://doi.org/10.3402/gha.v2i0.1932.

Fatehifar, M., J. Schlittenlacher, D. Wong, and K. Munro. 2023. "Applications Of Automatic Speech Recognition And Text-To-Speech Models To Detect Hearing Loss: A Scoping Review Protocol." *Inplasy Protocol*: 202310029.

Fontan, L., T. Cretin-Maitenaz, and C. Füllgrabe. 2020. "Predicting Speech Perception in Older Listeners with Sensorineural Hearing Loss Using Automatic Speech Recognition." *Trends in Hearing* 24:2331216520914769. https://doi.org/10.1177/2331216520914769.

Fontan, L., L. Gonçalves Braz, J. Pinquier, M. A. Stone, and C. Füllgrabe. 2022. "Using Automatic Speech Recognition to Optimize Hearing-Aid Time Constants." *Frontiers in Neuroscience* 16:779062. https://doi.org/10.3389/fnins.2022.779062.

Fontan, L., M. Le Coz, C. Azzopardi, M. A. Stone, and C. Füllgrabe. 2020. "Improving Hearing-Aid Gains Based on Automatic Speech Recognition." *The Journal of the Acoustical Society of America* 148 (3):EL227–EL233. https://doi.org/10.1121/10.0001866.

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

Galliano, S., E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri. May 2006. "Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).

Galliano, S., G. Gravier, and L. Chaubard. 2009. "The Ester 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts." In Interspeech 2009. ISCA, September, 2583–2586. https://doi.org/10.21437/Interspeech.2009-680.

Gonçalves Braz, L., L. Fontan, J. Pinquier, M. A. Stone, and C. Füllgrabe. 2022. "OPRA-RS: A Hearing-Aid Fitting Method Based on Automatic Speech Recognition and Random Search." *Frontiers in Neuroscience* 16:779048. https://doi.org/10.3389/fnins.2022.779048.

Gong, Y., S. Khurana, L. Karlinsky, and J. Glass. 2023. "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers." In INTERSPEECH 2023, August, 2798–2802. https://doi.org/10.21437/Interspeech.2023-2193.

Grasselli, G., A. Zangrillo, A. Zanella, M. Antonelli, L. Cabrini, A. Castelli, D. Cereda, A. Coluccello, G. Foti, R. Fumagalli, COVID-19 Lombardy ICU Network, et al. 2020. "Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy." *JAMA* 323 (16):1574–1581. https://doi.org/10.1001/jama.2020.5394.

Griffiths, T. D., M. Lad, S. Kumar, E. Holmes, B. McMurray, E. A. Maguire, A. J. Billig, and W. Sedley. 2020. "How Can Hearing Loss Cause Dementia?" *Neuron* 108 (3):401–412. https://doi.org/10.1016/j.neuron.2020.08.003.

Huckvale, M., and G. Hilkhuysen. 2022. "ELO-SPHERES Intelligibility Prediction Model for the Clarity Prediction Challenge 2022." In Interspeech 2022. ISCA, September, 3934–3938. https://doi.org/10.21437/Interspeech.2022-10521.

Huckvale, M., and G. Hilkhuysen. 2023. "Combining Acoustic Phonetic Linguistic and Audiometric Data in an Intrusive Intelligibility Metric for Hearing-Impaired Listeners." In Proc. ISCA Clarity-2023.

Hustad, K. C., and M. A. CahillMay 2003. "Effects of Presentation Mode and Repeated Familiarization on Intelligibility of Dysarthric Speech." *American Journal of Speech-Language Pathology* 12 (2):198–208. https://doi.org/10.1044/1058-0360(2003/066).

Ibelings, S., T. Brand, and I. Holube. 2022. "Speech Recognition and Listening Effort of Meaningful Sentences Using Synthetic Speech." *Trends in Hearing* 26:23312165221130656. https://doi.org/10.1177/23312165221130656.

Kamo, N., K. Arai, A. Ogawa, S. Araki, T. Nakatani, K. Kinoshita, M. Delcroix, T. Ochiai, and T. Irino. 2022. "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,

Katoch, S., S. S. Chauhan, and V. Kumar. 2021. "A review on Genetic Algorithm: Past, Present, and Future." *Multimedia Tools and Applications* 80 (5):8091–8126. https://doi.org/10.1007/s11042-020-10139-6.

Kollmeier, B., and M. Wesselkamp. 1997. "Development and Evaluation of a German Sentence Test for Objective and Subjective Speech Intelligibility Assessment." *The Journal of the Acoustical Society of America* 102 (4):2412–2421. https://doi.org/10.1121/1.419624.

Kosai, M., J. Kohda, J. Udaka, and Y. Koike. 1990. "Speech Audiometric Trial Using Synthetic Vowels Produced with DECtalk." *Medical Informatics* 15 (4):309–318. Jan. https://doi.org/10.3109/14639239009025279.

Lee, A., and T. Kawahara. 2009. "Recent Development of Open-Source Speech Recognition Engine Julius." In *Proceedings: APSIPA ASC 2009: Asia-Pacific signal and information processing association, 2009 annual summit and conference*, 131–137. Sapporo: Asia-Pacific Signal and Information Processing Association, October, 2009.

Lock, S., and C. K. Leong. 1989. "Program Library for DECtalk Text-to-Speech System." *Behavior Research Methods Instruments & Computers* 21 (3):394–400. https://doi.org/10.3758/BF03202806.

Mawalim, C. O., B. A. Titalim, S. Okada, and M. Unoki. 2023. "Non-Intrusive Speech Intelligibility Prediction Using an Auditory Periphery Model with Hearing Loss." *Applied Acoustics* 214:109663. https://doi.org/10.1016/j.apacoust.2023.109663.

Mawalim, C. O., B. A. Titalim, and M. Unoki. 2022. "CPC1 E031 System Description." In *Proceedings of the 2nd Clarity Workshop on Machine Learning Challenges for Hearing Aids (Clarity-2022)*.

McCormack, A., and H. Fortnum. May 2013. "Why Do People Fitted with Hearing Aids Not Wear Them?" *International Journal of Audiology* 52 (5):360–368. https://doi.org/10.3109/14992027.2013.769066.

McKay, C. M., and H. J. McDermott. 1993. "Perceptual Performance of Subjects with Cochlear Implants Using the Spectral Maxima Sound Processor (SMSP) and the Mini Speech Processor (MSP)." *Ear and Hearing* 14 (5):350–367. https://doi.org/10.1097/00003446-199310000-00006.

Meyer, B. T., B. Kollmeier, and J. Ooster. 2015. "Autonomous Measurement of Speech Intelligibility Utilizing Automatic Speech Recognition." In Interspeech 2015. ISCA, September, 2982–2986. https://doi.org/10.21437/Interspeech.2015-617.

Mogridge, R., G. Close, R. Sutherland, S. Goetze, and A. Ragni. 2023. "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.

Mogridge, R., G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni. 2024. "Non-Intrusive Speech Intelligibility Prediction for Hearing-Impaired Users Using Intermediate ASR Features and Human Memory Models." In I*CASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*s, 306–310. ICASSP. Seoul, Republic of Korea: IEEE, April.

Moore, B. C. J., C. Füllgrabe, and M. A. Stone. 2011. "Determination of Preferred Parameters for Multichannel Compression Using Individually Fitted Simulated Hearing Aids and Paired Comparisons." *Ear and Hearing* 32 (5):556–568. https://doi.org/10.1097/AUD.0b013e31820b5f4c.

Moore, B. C. J., B. R. Glasberg, and M. A. Stone. 2010. "Development of a New Method for Deriving Initial Fittings for Hearing Aids with Multi-Channel Compression: CAMEQ2-HF." *International Journal of Audiology* 49 (3):216–227. https://doi.org/10.3109/14992020903296746.

Munson, B., G. S. Donaldson, S. L. Allen, E. A. Collison, and D. A. Nelson. 2003. "Patterns of Phoneme Perception Errors by Listeners with Cochlear Implants as a Function of Overall Speech Perception Ability." *The Journal of the Acoustical Society of America* 113 (2):925–935. https://doi.org/10.1121/1.1536630.

Nejime, Y., and B. C. J. Moore. 1997. "Simulation of the Effect of Threshold Elevation and Loudness Recruitment Combined with Reduced Frequency Selectivity on the Intelligibility of Speech in Noise." *The Journal of the Acoustical Society of America* 102 (1):603–615. https://doi.org/10.1121/1.419733.

Nielsen, J., J. Nielsen, and J. Larsen. 2014. "Perception-based Personalization of Hearing Aids using Gaussian Processes and Active Learning." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23 (1):1–1. https://doi.org/10.1109/TASLP.2014.2377581.

Nisar, S., M. Tariq, A. Adeel, M. Gogate, and A. Hussain. 2019. "Cognitively Inspired Feature Extraction and Speech Recognition for Automated Hearing Loss Testing." *Cognitive Computation* 11 (4):489–502. https://doi.org/10.1007/s12559-018-9607-4.

Nuesse, T., B. Wiercinski, T. Brand, and I. Holube. 2019. "Measuring Speech Recognition With a Matrix Test Using Synthetic Speech." *Trends in Hearing* 23:2331216519862982. https://doi.org/10.1177/2331216519862982.

Oostdijk, N, et al. 2000. "The Spoken Dutch Corpus. Overview and First Evaluation." In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, 2: 887–894. Athens, Greece.

Ooster, J., R. Huber, B. Kollmeier, and B. T. Meyer. 2018. "Evaluation of an Automated Speech-Controlled Listening Test with Spontaneous and Read Responses." *Speech Communication* 98:85–94. https://doi.org/10.1016/j.specom.2018.01.005.

Ooster, J., M. Krueger, J.-H. Bach, K. C. Wagener, B. Kollmeier, and B. T. Meyer. 2020. "Speech Audiometry at Home: Automated Listening Tests via Smart Speakers with Normal-Hearing and Hearing-Impaired Listeners." *Trends in Hearing* 24:2331216520970011. https://doi.org/10.1177/2331216520970011.

Ooster, J., L. Tuschen, and B. T. Meyer. 2023. "Self-Conducted Speech Audiometry Using Automatic Speech Recognition: Simulation Results for

Listeners with Hearing Loss." *Computer Speech & Language* 78:101447. https://doi.org/10.1016/j.csl.2022.101447.
Osman, R. A., and H. A. Osman. 2021. "On the Use of Machine Learning for Classifying Auditory Brainstem Responses: A Scoping Review." *IEEE Access*. 9:110592–110600. https://doi.org/10.1109/ACCESS.2021.3102096.
Peddinti, V., Y. Wang, D. Povey, and S. Khudanpur. 2018. "Low Latency Acoustic Modeling Using Temporal Convolution and LSTMs." *IEEE Signal Processing Letters* 25 (3):373–377. https://doi.org/10.1109/LSP.2017.2723507.
Planey, A. M. 2019. "Audiologist Availability and Supply in the United States: A Multi-Scale Spatial and Political Economic Analysis." *Social Science & Medicine (1982)* 222:216–224. https://doi.org/10.1016/j.socscimed.2019.01.015.
Polspoel, S., D. R. Moore, D. W. Swanepoel, S. E. Kramer, and C. Smits. 2024. "Global Access to Speech Hearing Tests." *medRxiv* (2024): 2024–06.
Qayyum, A., J. Qadir, M. Bilal, and A. Al-Fuqaha. 2021. "Secure and Robust Machine Learning for Healthcare: A Survey." *IEEE Reviews in Biomedical Engineering* 14:156–180. https://doi.org/10.1109/RBME.2020.3013489.
Rabiner, L. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77 (2):257–286. https://doi.org/10.1109/5.18626.
Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S., Woodland, P., and Young S. 1995. "WSJCAM0 Cambridge Read News." p. 3670016 KB,
Roßbach, J., B. Kollmeier, and B. T. Meyer. 2022. "A Model of Speech Recognition for Hearing-Impaired Listeners Based on Deep Learning." *The Journal of the Acoustical Society of America* 151 (3):1417–1427. https://doi.org/10.1121/10.0009411.
Roßbach, J., R. Huber, S. Röttges, C. F. Hauth, T. Biberger, T. Brand, B. T. Meyer, and J. Rennies. 2022. "Speech Intelligibility Prediction for Hearing-Impaired Listeners with the LEAP Model." In proceedings *INTERSPEECH* 2022 :3498–3502. https://doi.org/10.21437/Interspeech.2022.
Schlueter, A., U. Lemke, B. Kollmeier, and I. Holube. 2016. "Normal and Time-Compressed Speech: How Does Learning Affect Speech Recognition Thresholds in Noise?" *Trends Hear* 20:233121651666988.
Schubotz, W., T. Brand, B. Kollmeier, and S. D. Ewert. 2016. "Monaural Speech Intelligibility and Detection in Maskers with Varying Amounts of Spectro-Temporal Speech Features." *The Journal of the Acoustical Society of America* 140 (1):524–540. https://doi.org/10.1121/1.4955079.
Seymore, K., S. Chen, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, et al. 1998. "The 1997 CMU Sphinx-3 English Broadcast News Transcription System." *Carnegie Mellon University*. https://doi.org/10.1184/R1/21709784.v1.
Sorin, C., and C. Thouin-Daniel. 1983. "Effects of Auditory Fatigue on Speech Intelligibility and Lexical Decision in Noise." *The Journal of the Acoustical Society of America* 74 (2):456–466. https://doi.org/10.1121/1.389839.
Taylor, K., and W. Sheikh., 2022. "Automated Hearing Impairment Diagnosis Using Machine Learning." In *2022 Intermountain Engineering, Technology and Computing. IETC*, 1–6. IEEE: Orem, UT.
Tricco, A. C., E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. J. Peters, T. Horsley, L. Weeks, et al. 2018. "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation." *Annals of Internal Medicine* 169 (7):467–473. https://doi.org/10.7326/M18-0850.
Tu, Z., N. Ma, and J. Barker. 2022a. "Exploiting Hidden Representations from a DNN-Based Speech Recogniser for Speech Intelligibility Prediction in Hearing-impaired Listeners."
Tu, Z., N. Ma, and J. Barker. 2022b. "Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-intrusive Speech Intelligibility Prediction."
Tu, Z., N. Ma, and J. Barker. 2023. "Intelligibility Prediction with a Pretrained Noise-Robust Automatic Speech Recognition Model."
Vaillancourt, V., C. Laroche, C. Mayer, C. Basque, M. Nali, A. Eriks-Brophy, S. D. Soli, and C. Giguère. 2005. "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." *International Journal of Audiology* 44 (6):358–369. https://doi.org/10.1080/14992020500060875.
Van, T. D. J., and J. L. Yanz. 1987. "Speech Recognition Threshold in Noise." *Journal of Speech, Language, and Hearing Research* 30 (3):377–386. https://doi.org/10.1044/jshr.3003.377.
Vesely, K., A. Ghoshal, L. Burget, and D. Povey. 2013. "Sequence-Discriminative Training of Deep Neural Networks." *Interspeech* 2013: 2345–2349.
Vikhar, P. A. 2016. "Evolutionary algorithms: A critical review and its future prospects." In *2016 International conference on global trends in signal processing, information computing and communication. ICGTSPICC*, 261–265. Jalgaon, India: IEEE.
Wagener, K. C., and T. Brand. 2005. "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." *International Journal of Audiology* 44 (3):144–156. https://doi.org/10.1080/14992020500057517.
Wagener, K., V. Kühnel, and B. Kollmeier. 1999. "Development and Evaluation of a German Sentence Test I: Design of the Oldenburg Sentence Test." *Z. Audiol* 38:4–15.
Wasmann, J.-W., L. Pragt, R. Eikelboom, and D. W. Swanepoel. 2022. "Digital Approaches to Automated and Machine Learning Assessments of Hearing: Scoping Review." *Journal of Medical Internet Research* 24 (2): e32581. https://doi.org/10.2196/32581.
Whitton, J. P., K. E. Hancock, J. M. Shannon, and D. B. Polley. 2016. "Validation of a Self-Administered Audiometry Application: An Equivalence Study: Equivalence of Mobile and Clinic-Based Tests." *The Laryngoscope* 126 (10):2382–2388. https://doi.org/10.1002/lary.25988.
Willberg, T., V. Sivonen, S. Hurme, A. A. Aarnisalo, H. Löppönen, and A. Dietz. 2020. "The Long-Term Learning Effect Related to the Repeated use of the Finnish Matrix Sentence Test and the Finnish Digit Triplet Test." *International Journal of Audiology* 59 (10):753–762. https://doi.org/10.1080/14992027.2020.1753893.
"World Report on Hearing," 2021. https://www.who.int/publications-detail-redirect/9789240020481.
Zezario, R. E., F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao. 2022. "MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids."
Zezario, R. E., S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao. 2023. "Deep Learning-Based Non-Intrusive Multi-Objective Speech Assessment Model with Cross-Domain Features." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31:54–70. https://doi.org/10.1109/TASLP.2022.3205757.
Zhou, X., C. O. Mawalim, B. A. Titalim, and M. Unoki. 2023. "Incorporating the Digit Triplet Test in A Lightweight Speech Intelligibility Prediction for Hearing Aids." In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 1593–1600. Taipei, Taiwan: IEEE.