Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Harnessing text-to-speech voice cloning models for improved audiological speech assessment

Lidea Shahidi[1], *Erdem Baha Topbas*[1], *Thu Ngan Dang*[1], *Tobias Goehring*[1]

[1]MRC Cognition and Brain Sciences Unit, University of Cambridge, UK
{Lidea.Shahidi, ErdemBaha.Topbas, ThuNgan.Dang, Tobias.Goehring}@mrc-cbu.cam.ac.uk

Abstract

Conventional audiological speech assessments are limited in their predictive utility, due to the small number of available stimuli and the restricted communications experiences that they represent. To enhance the capabilities of audiological speech assessments, this work evaluates the ability of several text-to-speech voice cloning models on the task of replicating a standard UK open-set speech test used clinically. Models are evaluated using complementary measures of speech perception: speech intelligibility in background noise, speech quality, and speaker discrimination, in a large-scale online study (N = 73). To ensure speech intelligibility measurements are comparable, the psychometric functions are characterized for each model. Results indicate models which accurately and consistently replicate speaker characteristics and produce speech that is similarly intelligible and natural for audiological speech assessment.

Index Terms: speech perception, speech intelligibility, speech quality, speaker identity, speech synthesis, text-to-speech, voice cloning

1. Introduction

Given the importance of speech for everyday communication, speech perception is a primary focus of audiological interventions and evaluations. Conventional speech assessments involve the presentation of speech recordings followed by the listener's word recall. Speech assessments must achieve consistent and well-characterized outcomes, such that multiple measurements made with the same test method are accurate, precise, and comparable.

Due to the time and cost required to develop, record, and calibrate new speech material, speech tests often contain only a few hundred unique speech utterances, thereby limiting how many repeated measures, conditions or sessions can be tested without reusing stimuli. This limits the statistical power and experimental design of the speech assessment while potentially introducing confounds due to imbalanced learning and familiarity effects. Thousands of stimuli would be necessary to fully mitigate these effects, but this amount of speech material would possibly require a prohibitive amount of time and cost to obtain from in-person recordings.

In addition to limitations due to the number of available stimuli, the predictive utility of these assessments has come under scrutiny for their lack of relevance to everyday communication experiences [1, 2] with the assessments themselves explaining limited variability in patient-reported measures of functional hearing and quality of life [3-5]. One shortcoming of existing assessments is the limited diversity of speakers represented in existing corpora. At best, most speech corpora contain speech uttered by just a few speakers (and often only a single speaker), constraining speech perception assessments to the voicing characteristics of the available speakers. When constructing new corpora, speaker characteristics such as linguistic and social identity are largely overlooked [6]. Speaker characteristics affect speech perception on a fundamental level: familiarity with a speaker's accent can improve comprehension [7-11] and is particularly beneficial to older adults with hearing loss [12, 13]. Conversely, competing speech is more disruptive when formed by unfamiliar accented talkers [14]. Given the resources required to produce speech materials, patients or experimental participants from minoritised communities are less likely to be assessed using speech recordings produced by individuals with a familiar language, accent, or sociolect [15, 16], despite facing higher risk and incidence of hearing loss [17-20].

Modern speech synthesis models could address the limitations of existing audiological speech materials. Synthesized speech has shown promise for use in speech assessments, and previous studies have used text-to-speech systems to reproduce speech assessments involving the recall of isolated words [21], phrases [22] and sentences in closed-set [23] and open-set [24] paradigms. These investigations found high test-retest accuracy for measures obtained with synthetic speech and good correspondence with psychometric functions obtained with the original speech assessments, when measuring speech intelligibility as a function of signal-to-noise ratio (SNR) or sound level. Although the synthetic speech samples were altered to match some characteristics (e.g. fundamental frequency or syllable rate) of the original speaker, comprehensive differences in the voicing characteristics of the two speakers under comparison were not accounted for, confounding the interpretation of the results. These evaluations also made use of proprietary models, despite the availability of open-source speech synthesis models. Such proprietary models are typically incompatible with the objectives of replicable scientific inquiry, as the models are not only inaccessible and inscrutable, but may change at any point.

Voice cloning models that can emulate the articulatory patterns of diverse speakers have recently become available [e.g., 25-28], although their potential has yet to be harnessed in an audiological context. Not only could voice cloning models provide the experimental control needed to evaluate synthetic audiological assessments, but they could be used to generate new assessments with improved ecological relevance, for instance by matching the synthetic speech to the accent or sociolect most often encountered by the listener. As a proof-of-concept, this study evaluates the ability of several open-source text-to-speech voice cloning models to replicate the Bamford-Kowal-Bench (BKB) sentence lists [29], the current standard for speech assessment in research and clinical settings in the UK. To disambiguate the strengths and weaknesses of each model, models are assessed using three complementary

10.21437/Interspeech.2025-2595

measures of speech perception: speech intelligibility, speech quality, and speaker identity. This study provides first evidence towards the use of voice cloning in the development of an audiological assessment, and demonstrates a framework for evaluating speech synthesis for audiological assessment, towards improving the ecological relevance and representation of diverse speakers in audiological assessments.

2. Method

The design and analysis for this study were preregistered.[1]

2.1. Models

This study evaluated three open-source text-to-speech voice cloning models: VALL-E X[2] [26], XTTSv2[3] [27], and E2[4] [28], selected using an informal evaluation of synthesised speech quality across a range of speakers. To aid experimental control, implementation, and assessment, all models were pre-trained zero-shot models.

2.2. Speech Material

This study assessed replication of the BKB speech corpus [29], a set of speech materials that is used extensively in the UK to assess speech outcomes in research and clinical settings. The BKB corpus is spoken by a male speaker with a Southern Standard British English accent and contains 270 sentences in total (18 lists of 15 sentences each). Models were provided with recordings from lists 1-5 to specify the acoustic characteristics of the target speaker. Models were then prompted to produce sentences with equivalent semantic context as lists 6-18, and the resulting utterances were evaluated in a speech perception study.

2.3. Participants

Seventy-three participants were recruited through the Prolific online platform (www.prolific.com). Participants were excluded from the study if they were not native English speakers raised in Britain, did not have normal or corrected-to-normal vision, or had a history of hearing loss, a speech or language disorder, or a neurological disorder. Each participant provided informed consent and received reimbursement for completing the study. This study was approved by the Cambridge Psychology Research Ethics Committee at the University of Cambridge.

2.4. Procedure

Participants first set the volume to a comfortable listening level and then completed a headphone test [30]. Participants then assessed the intelligibility, quality, and speaker identity of the speech samples from each speech source in three separate tasks. The order of tasks was consistent across participants.

2.4.1. Speech intelligibility

Intelligibility was measured as the percentage of words correctly identified in each sentence. Sentences were presented in a multi-talker babble noise at nine SNRs: -12, -10, -8, -6, -4, -2, 0, +2, and +4 dB. The babble noise was comprised of a mixture of competing speech from twenty English talkers (8 male, 12 female) and was recorded at Auditec, Inc. (St. Louis, MO, USA). The SNR was manipulated by varying the level of the noise relative to the target sentence. Each participant completed 3 repetitions for each combination of speech source (VALL-E X, XTTSv2, E2, and Original) and SNR, totalling 108 trials presented in a random order.

2.4.2. Speech quality

Quality was measured using the MOS-X2 scale, developed for the assessment of synthetic speech [31]. The scale measures quality in four dimensions: intelligibility, naturalness, prosody, and social impression, and the aggregate MOS-X2 score is obtained by averaging the scores obtained across the four dimensions. To provide meaningful perceptual anchors, two additional conditions were included: speech processed by a vocoder that simulates the electrical current spread that occurs in a cochlear implant (using a spread of -16 dB/octave, 22 channels, and 80 carriers) [32], and speech generated by a hidden Markov model-based speech synthesis system [5] (Festival) [33]. Participants performed 5 repetitions for each condition (Vocoded, Festival, VALL-E X, XTTSv2, E2, and Original), totalling 30 trials presented in a random order.

2.4.3. Speaker identity

Speaker identity was measured using a two-alternative forced choice speaker discrimination task in which participants were presented with two consecutive speech samples and asked whether the samples came from the same speaker or from different speakers. Seven comparisons were performed across the speech sources. Samples from each speech source were compared to samples from the original corpus, to measure replication of target speaker characteristics (yielding 4 comparisons), and samples from each model were compared to samples from the same model, to measure the consistency in the speaker characteristics generated by each model (yielding 3 comparisons).

The speaker identity task was performed twice: once in quiet and once in degraded listening conditions, to simulate an everyday listening scenario in which speaker identity may not be as easily discernible. The degraded condition simulated information loss due to telecommunications transmission by convolving the signal with a microphone impulse response, applying a bandpass filter (lower and upper cut-offs of 300 Hz and 3 kHz, respectively; implemented using a 6th order Butterworth filter), convolving with an impulse response of a telephone horn and combining with speech-shaped noise at +15 dB SNR. Participants performed 5 repetitions for each comparison in both quiet and degraded listening conditions, totalling 70 trials. The order of comparisons and repetitions were randomized throughout the quiet and degraded tasks, with participants completing the quiet task before the degraded task.

2.5. Analysis

A one-way repeated measures ANOVA was performed to test for an effect of speech source on each dependent variable. For analyses where a main effect of source was observed, pairwise comparisons between sources were performed using *post hoc*

[1] https://osf.io/etnvh
[2] https://github.com/Plachtaa/VALL-E-X
[3] https://github.com/coqui-ai/TTS
[4] https://github.com/SWivid/F5-TTS
[5] http://www.cstr.ed.ac.uk/projects/festival.html

paired t-tests with Bonferroni corrections for multiple comparisons. Significance was tested for at the $p < 0.05$ level.

For the speech intelligibility task, a logistic function was fit to the percent correct scores for each participant (averaged over repetitions within participants) and the resulting threshold (the SNR at which 50% of words are recalled correctly) and slope were used as dependent measures. For the speech quality task, repetitions were averaged within participants and the aggregate MOS-X2 score was used as a dependent measure. For the speaker identity task, responses of "same speaker" or "different speakers" were assigned a score of 1 or 0, respectively, averaged across repetitions within each participant to obtain the proportion of same speaker responses for each participant. Proportions were then transformed using the arcsine transformation and used as a dependent measure.

3. Results

3.1. Speech intelligibility

Psychometric functions obtained in the speech intelligibility task are presented in Figure 1. Speech produced by VALL-E X was generally harder to understand than that produced by the other models, particularly above -8 dB SNR. XTTSv2 produced speech that was the most robust to the added babble noise, followed by the original speech samples and samples produced by the E2 model.

A significant effect of speech source was observed on both psychometric function thresholds ($F(3,213) = 11.86$, $p < 0.0001$) and slopes ($F(3,213) = 3.27$, $p = 0.038$). Post hoc tests indicated significant differences in the thresholds obtained for speech from the original corpus and XTTSv2 ($t(71) = 3.86$, $p < 0.001$), with an estimated difference in threshold of -0.58 dB, suggesting that speech produced by XTTSv2 was more easily understood at lower SNRs. Differences between the thresholds obtained for the original speech and VALL-E X (of 0.05 dB) and E2 (of 0.41 dB) were not significant (VALL-E X: $t(71) = -0.02$, $p = 1.00$; E2: $t(71) = 2.69$, $p = 0.053$). There were no significant differences in the slopes of the psychometric functions obtained with the original corpus and any of the models (VALL-E X: $t(71) = 2.44$, $p = 0.102$; XTTSv2: $t(71) = 2.37$, $p = 0.124$; E2: $t(71) = 0.68$, $p = 1.00$) indicating that the precision of the threshold measurement was not impacted.

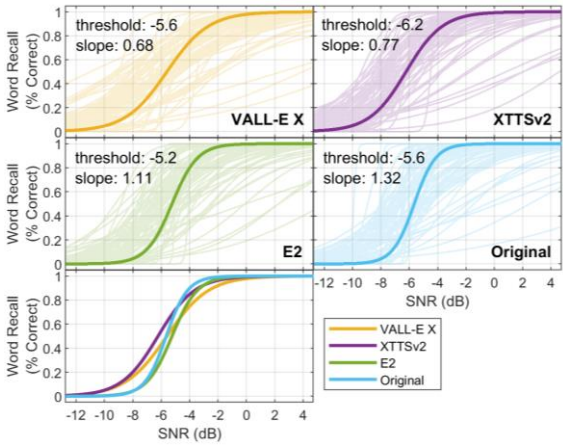

Figure 1: *Psychometric functions of word recall by SNR obtained for each speech source. Logistic functions fit to individual participant data are plotted with thin lines while the mean logistic function for each model is plotted with thick lines.*

3.2. Speech quality

Speech quality ratings acquired for each speech source are presented in Figure 2. Samples produced by XTTSv2 and E2 were rated as being similarly intelligible as the original samples, although E2 generally earned lower scores than XTTSv2 on the naturalness, prosody, and social impression ratings. VALL-E X earned the lowest ratings of the three models, but outperformed the HMM-based model, Festival, in ratings of naturalness, prosody, and social impression.

MOS-X2 ratings were lowest for the vocoded condition, followed by Festival, VALL-E X, E2, XTTSv2, and the original speaker, with the analysis indicating a significant main effect of speech source ($F(5,360) = 571.36$, $p < 0.0001$). All sources obtained significantly lower MOS-X2 ratings from those obtained by the original speech (Original-Vocoded: $t(72) = 41.73$, $p < 0.0001$; Original-Festival: $t(72) = 13.98$, $p < 0.0001$; Original-VALL-E X: $t(72) = 12.98$, $p < 0.0001$; Original-XTTSv2: $t(72) = 3.82$, $p < 0.001$; Original-E2: $t(72) = 8.08$, p

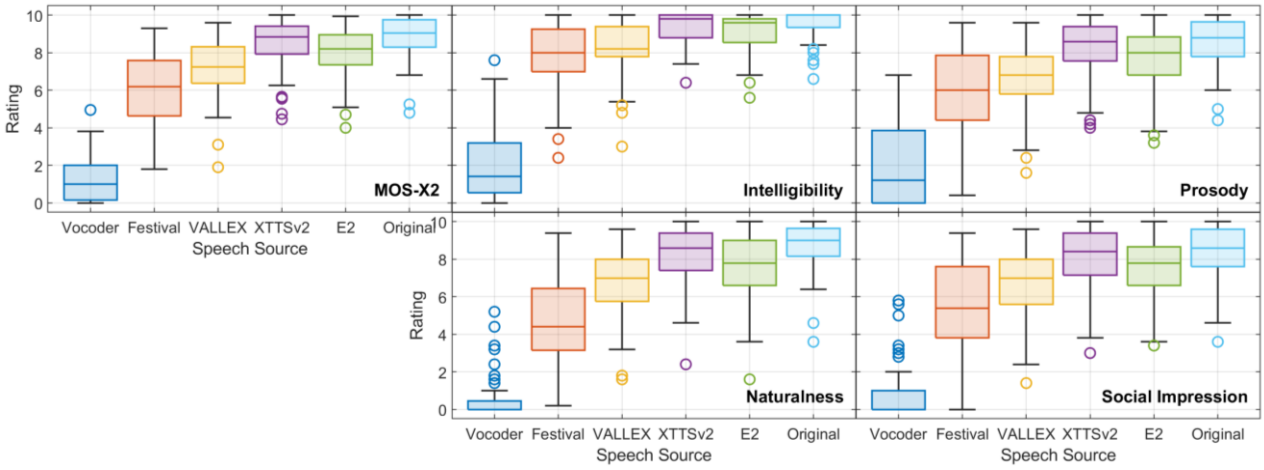

Figure 2: *Speech quality ratings obtained across speech sources, for the aggregate MOS-X2, and the individual ratings of intelligibility, prosody, naturalness, and social impression. Boxplots display the median, and upper and lower quartiles, with outliers indicated by circular markers.*

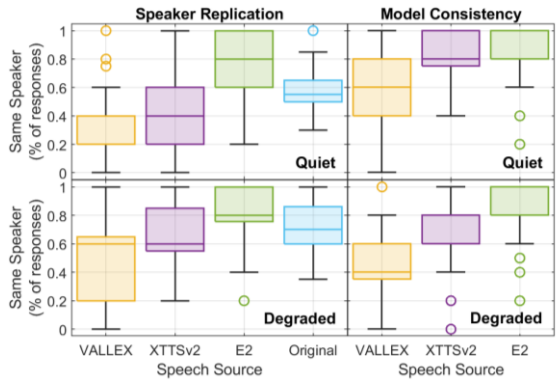

Figure 2: *Proportion of samples rated as being produced by the same speaker, as a measure of speaker replication (left column) or model consistency (right column), in quiet (top row) or degraded (bottom row) conditions. Boxplots display the median, and upper and lower quartiles, with outliers indicated by circular markers.*

< 0.0001), suggesting that participants were able to discern the qualitative differences between utterances from the original corpus and the different sources.

3.3. Speaker identity

Scores obtained for the speaker identity task are presented in Figure 3. For replication of the BKB speaker characteristics, there was a significant effect of speech source on the proportion of trials in which participants indicated that the speech sample was spoken by the original BKB speaker, both in quiet ($F(3,216) = 135.9$, $p < 0.0001$) and degraded ($F(3,216) = 67.35$, $p < 0.0001$) listening conditions. In quiet, pairwise comparisons indicated that VALL-E X and XTTSv2 received fewer "same speaker" responses when presented with samples from the original speaker, when compared to the proportion of "same speaker" responses received when two samples from the original speaker were compared (Original-VALL-E X: $t(72) = 18.23$, $p < 0.0001$; Original-XTTSv2: $t(72) = 13.12$, $p < 0.0001$). In contrast, E2 received a higher proportion of "same speaker" responses when presented with a sample from the original speaker, compared to the proportion of "same speaker" responses received by the original speaker samples (Original-E2: $t(72) = -5.96$, $p < 0.0001$), indicating that samples from the E2 model were perceived as more similar to the original samples than the original samples themselves. In degraded conditions, the trends across models remained (Original-VALL-E X: $t(72) = 12.98$, $p < 0.0001$; Original-XTTSv2: $t(72) = 7.48$, $p < 0.0001$) but the significant difference between E2 and the original samples was resolved by the more challenging listening condition (Original-E2: $t(72) = -2.55$, $p = 0.078$).

For the consistency in the voice generated by each model, there was again a main effect of model both in quiet ($F(2,144) = 69.82$, $p < 0.0001$) and degraded ($F(2,144) = 63.07$, $p < 0.0001$) listening conditions. Pairwise comparisons indicated that E2 achieved greater consistency than XTTSv2 and VALL-E X both in quiet (E2-XTTSv2: $t(72) = 4.90$, $p < 0.0001$; E2-VALL-E X: $t(72) = 12.67$, $p < 0.0001$) and degraded (E2-XTTSv2: $t(72) = 3.03$, $p = 0.01$; E2-VALL-E X: $t(72) = 11.61$, $p < 0.0001$) conditions. XTTSv2 also achieved greater consistency than VALL-E X in quiet (XTTSv2-VALL-E X: $t(72) = 5.73$, $p < 0.0001$) and degraded (XTTSv2-VALL-E X: $t(72) = 8.19$, $p < 0.0001$) conditions. Taken together, these results indicate that the models reproduce the speaker characteristics of the original BKB corpus to varying degrees, with some models producing speech samples from a distinct yet consistent vocal identity.

4. Discussion

This study evaluated three text-to-speech models (VALL-E X, XTTSv2, and E2) on the task of replicating the BKB speech corpus, and found that each model demonstrated different strengths and shortcomings in reproducing speech outcomes obtained with the target speech material.

Psychometric functions obtained with speech produced by VALL-E X demonstrated comparable thresholds and slopes to those obtained with the original corpus, although deviations of the psychometric function above -6 dB SNR were noted. Speech perception at moderately low to high SNRs is particularly pertinent to hearing impaired listeners, calling into question the use of VALL-E X-produced speech in assessments relevant to this population. Poor speech intelligibility at these SNRs may have been due to variability in the acoustic and linguistic quality of the generated speech material, as indicated by low speech quality and model consistency scores, and would suggests poor test-retest reliability if deployed in an audiological assessment.

In contrast, XTTSv2 consistently produced speech that was more intelligible than speech produced by the other models, as indicated by significantly lower thresholds of the psychometric functions obtained with XTTSv2. XTTSv2 also earned the highest speech quality ratings of all models under test and high scores for model consistency. These outcomes suggest that the voice produced by XTTSv2 was not only natural and intelligible but conferred a consistent speaker identity. XTTSv2's primary limitation was its inability to reproduce the speaker characteristics of the original speaker, as evidenced by significantly lower scores on the speaker replication task than those obtained with the original speech.

Psychometric functions obtained with E2 were comparable to those obtained with the original corpus, providing sufficient measurement precision for audiological assessment. Although E2 earned lower quality ratings than the original corpus, it produced a highly consistent vocal identity that was often rated as being more similar to the original speaker than the original speaker himself. These outcomes suggest that E2 could be used to extend the capabilities of the BKB corpus and generate new corpora containing speech from a diverse and representative sample of speakers, enhancing the predictive utility of audiological assessments. Towards this end, the speaker replication fidelity of E2 must still be rigorously evaluated with speakers with a wide range of accents and speaking styles.

5. Conclusions

Several voice cloning text-to-speech models were evaluated on their ability to reproduce the speech perception outcomes and speaker characteristics of the BKB speech corpus. Measures included speech intelligibility in noise, speech quality, and speaker discrimination. While models were able to replicate the psychometric functions obtained with the original speech corpus, differences in models arose in the quality of the speech produced, the consistency of the voice generated by each model, and the fidelity with which the voicing characteristics of the original speaker were reproduced. Models were identified that can produce speech suitable for audiological testing, with implications for the development of synthetic audiological speech materials that represent a range of lived experiences.

6. Acknowledgements

The authors would like to thank the participants for their time and enthusiasm and the members of the Cambridge Hearing Group for their input on this project. This work was funded by an award from Cambridge Language Sciences and AI@Cam at the University of Cambridge.

7. References

[1] G. Keidser, G. Naylor, D. S. Brungart, A. Caduff, J. Campos, S. Carlile, et al. "The quest for ecological validity in hearing science: What it is, why it matters, and how to advance it," *Ear and hearing*, vol. 41, pp. 5S-19S, 2020.

[2] M. B. Winn and R. A. Wright, "Reconsidering commonly used stimuli in speech perception experiments." *The Journal of the Acoustical Society of America*, vol. 152, no. 3, 1394-1403, 2022.

[3] T. C. Walden and B. E. Walden, "Predicting success with hearing aids in everyday living," *Journal of the American Academy of Audiology*, vol. 15, no. 05, 342-352, 2004.

[4] , J. R. Dornhoffer, T. A. Meyer, J. R. Dubno, and T. R. McRackan, "Assessment of hearing aid benefit using patient-reported outcomes and audiologic measures," *Audiology and Neurotology*, vol. 25, no. 4, pp. 215-223, 2020.

[5] A. Davidson, N. Marrone, B. Wong, and F. Musiek, "Predicting hearing aid satisfaction in adults: A systematic review of speech-in-noise tests and other behavioral measures," *Ear and hearing*, vol. 42, no. 6, pp. 1485-1498, 2021.

[6] Y. D. Hyter and M. B. Salas-Provance, "Culturally responsive practices in speech, language, and hearing sciences," Plural Publishing, 2021.

[7] S. H. Ferguson, A. Jongman, J. A. Sereno, and K. A. Keum, "Intelligibility of foreign-accented speech for older adults with and without hearing loss," *Journal of the American Academy of Audiology*, vol. 21, no. 03, pp. 153-162, 2010.

[8] S. Gordon-Salant, G. H. Yeni-Komshian, and P. J. Fitzgibbons, "Recognition of accented English in quiet and noise by younger and older listeners," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3152-3160, 2010.

[9] A. Cristia, A. Seidl, C. Vaughn, R. Schmale, A. Bradlow, and C. Floccia, "Linguistic processing of accented speech across the lifespan." *Frontiers in psychology*, vol. 3, no. 479, 2012.

[10] K. J. Van Engen and J. E. Peelle, "Listening effort and accented speech," *Frontiers in human neuroscience*, vol. 8, no. 577, 2014.

[11] L. Stringer and P. Iverson, "Accent intelligibility differences in noise across native and nonnative accents: Effects of talker–listener pairing at acoustic–phonetic and lexical levels," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 7, pp. 2213-2226, 2019.

[12] D. Strori, A. R. Bradlow, and P. E. Souza, "Recognition of foreign-accented speech in noise: The interplay between talker intelligibility and linguistic structure," *The Journal of the Acoustical Society of America*, vol. 147, no. 6, pp. 3765-3782, 2020.

[13] S. Schilaty, S. H. Ferguson, S. D. Morgan, and C. Champougny, "Intelligibility of British-and American-Accented Sentences for American Younger and Older Listeners With and Without Hearing Loss," *Journal of the American Academy of Audiology*, vol. 32, no. 01, pp. 010-015, 2021.

[14] B. Senior and M. Babel, "The role of unfamiliar accents in competing speech," *The Journal of the Acoustical Society of America*, vol. 143, no. 2, pp. 931-942, 2018.

[15] I. Ramkissoon, "Speech recognition thresholds for multilingual populations," *Communication Disorders Quarterly*, vol. 22, no. 3, pp. 158-162, 2001.

[16] M. J. Suh, J. Lee, W. H. Cho, I. K. Jin, T. H. Kong, S. H. Oh, et al. "Improving accuracy and reliability of hearing tests: an exploration of international standards," *Journal of Audiology & Otology*, vol. 27, no. 4, pp. 169, 2023.

[17] P. Dawes, H. Fortnum, D. R. Moore, R. Emsley, P. Norman, K. Cruickshanks, et al. "Hearing in middle age: a population snapshot of 40-to 69-year olds in the United Kingdom," *Ear and hearing*, vol. 35, no. 3, pp.e44-e51, 2014.

[18] M. Lor, S. Thao, and S. M. Misurelli, "Review of hearing loss among racial/ethnic minorities in the United States," *Western Journal of Nursing Research*, vol. 43, no. 9, pp. 859-876, 2021.

[19] H. Zhou, Y. Zhou, H. Zhang, A. Yu, B. Zhu, and L. Zhang, £Socio-economic disparity in the global burden of occupational noise-induced hearing loss: an analysis for 2017 and the trend since 1990," *Occupational and Environmental Medicine*, vol. 78, no. 2, pp. 125-128, 2021.

[20] J. Allum, R. Meredith, K. Uus, J. J. Kirkham, and P. Dawes, "Association between hearing loss and deprivation among Welsh adults: a cross-sectional observational study," *BMJ open*, vol. 13, no. 5, pp. e066419, 2023.

[21] S. Polspoel, F. S. Holtrop, A. J. Bosman, S. E. Kramer, and C. Smits, "Measurement and optimisation of the perceptual equivalence of the Dutch consonant-vowel-consonant (CVC) word lists using synthetic speech and list pairs," *International Journal of Audiology*, pp. 1–8, 2024.

[22] S. Ibelings, T. Brand, E. Ruigendijk, and I. Holube, "Development of a Phrase-Based Speech-Recognition Test Using Synthetic Speech," *Trends in Hearing*, vol. 28, 2024.

[23] T. Nuesse, B. Wiercinski, T. Brand, I. Holube, "Measuring Speech Recognition With a Matrix Test Using Synthetic Speech," *Trends Hear*, vol. 23, pp.1-14, 2019.

[24] B. Herrmann, "The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults," *International Journal of Speech Technology*, vol. 26, pp. 395-415, 2023.

[25] F. Lux, S. Meyer, L. Behringer, F. Zalkow, P. Do, M. Coler, E. A. P. Habets, N. T. Vu, "Meta Learning Text-to-Speech Synthesis in over 7000 Languages," *Interspeech*, Greece, pp. 4958-4962, 2024.

[26] L. Meng, L. Zhou, S. Liu, S. Chen, B. Han, S. Hu, et al. "Autoregressive Speech Synthesis without Vector Quantization," ArXiv, abs/2407.08551, 2024.

[27] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, et al. "XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model," ArXiv, abs/2406.04904, 2024.

[28] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C. Tsai, Z. Xiao, et al. "E2 TTS: Embarrassingly Easy Fully Non-Autoregressive Zero-Shot TTS," *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 682-689, 2024.

[29] J. Bench and J. Bamford, "Speech hearing tests and the spoken language of hearing-impaired children," London: Academic Press, 1979.

[30] K. J. P. Woods, M. H. Siegel, J. Traer and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, pp. 2064–2072, 2017.

[31] J. R. Lewis, "Investigating MOS-X ratings of synthetic and human voices," *Voice Interaction Design*, vol. 2, pp. 1–22, 2018.

[32] J. A. Grange, J. F. Culling, N. S. L. Harris, and S. Bergfeld, "Cochlear implant simulator with independent representation of the full spiral ganglion," The Journal of the Acoustical Society of America, vol. 142, no. 5, pp. EL484, 2017.

[33] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," *Proc. 6th ISCA Workshop on Speech Synthesis*, pp. 294-299, 2007.