Audio Engineering Society

Conference Paper 14

Presented at the AES International Conference on
Machine Learning and Artificial Intelligence for Audio
2025 September 8-10, London, UK

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Challenges in predicting the lyric intelligibility of musical segments for older individuals with hearing loss

William M. Whitmer[1,2], David McShefferty[1], Michael A. Akeroyd[3], Scott C. Bannister[4], Jon P. Barker[5], Trevor J. Cox[6], Bruno M. Fazenda[6], Jennifer Firth[3], Simone N. Graetzer[6], Alinka E. Greasley[4], Gerardo Roa Dabike[5] and Rebecca Vos[6]

[1] *Hearing Sciences – Scottish Section, University of Nottingham*
[2] *School of Health and Wellbeing, University of Glasgow*
[3] *Hearing Sciences, University of Nottingham*
[4] *School of Music, University of Leeds*
[5] *Speech and Hearing Research Group, University of Sheffield*
[6] *Acoustics Research Centre, University of Salford*

Correspondence should be addressed to William M. Whitmer (bill.whitmer@nottingham.ac.uk)

ABSTRACT

Difficulty in understanding lyrics is a major barrier to enjoying music for people with hearing loss. Novel production techniques using machine learning could aid lyric understanding, but objective metrics for training are needed, and these must be informed by the experiences of the target population. Currently, there are no data on the lyric-recall ability of older individuals with hearing loss. Thirteen older participants with mostly mild-sloping hearing loss listened to and recalled 100 vocal segments of popular music that varied in genre, duration and word count. In each trial, participants heard a randomly chosen segment twice with a 5-s interstimulus interval. Reproduction was over headphones at 65 dB(A) using individualised frequency-dependent nonlinear gain to compensate for hearing loss. The proportion of words recalled correctly varied greatly across samples, from 0-100%, and varied as a function of genre, similar to past studies with different populations. Individual intelligibility across segments was correlated with age; segment intelligibility across individuals, however, was not correlated with word count or rate. An objective metric for lyric intelligibility for listeners with hearing loss was developed based on Whisper automatic speech recognition model and individualised hearing-loss simulation. At best, this metric was only able to account for 56% of the variance in listening test results and often underpredicting intelligibility. Improving sung-lyric intelligibility is a different challenge from spoken-speech enhancement for those with hearing loss. Not only does the enhancement need to be considered within the overall enjoyment of the music, but the variation in results – as seen in the current study – reflects variations in genre, orchestration and vocal quality in sung music.

1 Introduction

The joy of music listening can be a key factor for quality of life throughout the lifespan, including older age [1]. For many listeners, lyrics are a key feature in music, with research demonstrating the importance of these for affective experiences [2]-[6]. Critically, people with hearing loss report difficulties with understanding lyrics [7]-[8], introducing a major barrier to the enjoyment of music.

There are numerous factors affecting lyric intelligibility. In a survey of music listeners, Fine and Ginsborg [9] found singer-related (e.g., enunciation,

vibrato), music-related (e.g., accompaniment level/ density, mixing), word-related (e.g., formulaic words), listener-related (e.g., attention, hearing loss), and environmental (e.g., audio-visual, reverberation, reproduction) factors can all affect how well the lyrics are understood. While several studies have looked at the differences between sung and spoken intelligibility, only two studies have looked at the intelligibility of sung lyrics in their original context. Condit-Schultz and Huron [10] asked thirty university music and speech-hearing students (ages not given), who passed a normal-hearing screening questionnaire, to repeat back the lyrics of recorded vocal music segments of differing genres. The results showed a strong genre-dependent effect, with their classical genre excerpts being the least intelligible. In a secondary experiment, a subset of segments was repeated; performance (correct recall) increased on second and slightly on third presentation, but not for additional repeats. Self-reported genre familiarity was not correlated with performance. A later study by Ibrahim et al. [11] included 17 Singaporean participants aged 21-41, who were mostly university students, reported no hearing problems and "spoke some English as a second language" (p. 2). Those participants heard each of 100 recorded segments twice. That study [11] was intended to automatically classify the intelligibility of vocal music for potential use in the teaching of English, hence there was no analysis of the behavioural results, which were considered ground truth to evaluate their classifier.

One of the aims of the Cadenza project [12] is to run machine-learning challenges to optimise technology to improve lyric understanding without sacrificing overall enjoyment of the music. To achieve this optimisation, developers need lyric-intelligibility metrics that correlate with the perception of the target population. Currently, there are no data on the ability of older individuals with hearing loss to hear and recall lyrics.

In the current study, we aimed to create more inclusive future objective metrics by applying the same behavioural methodology of Ibrahim et al. to an older population with hearing loss. To ensure that the results were not simply a case of audibility, the stimuli were amplified to compensate for each listener's hearing loss. We then assessed how well a state-of-art machine-learning metric predicts lyric intelligibility for this population, and what is needed to create future lyric-intelligibility metrics that can better predict across listeners.

2 Methods

2.1 Participants

Thirteen adults, aged 58-79 (median 73) years, were recruited from the participant pool of Hearing Sciences – Scottish Section who were all native speakers of British English. Left and right pure-tone thresholds were measured using the modified Hughson-Westlake procedure [13] and are shown in Figure 1. Better-ear four-frequency pure-tone averages (BE4FA), a standard metric of objective hearing ability, were calculated from the left-right minimum average of thresholds at 0.5, 1, 2 and 4 kHz. BE4FAs ranged from 17.5-41.25 (median 25) dB HL. Age and BE4FA were not significantly correlated (Pearson product-moment correlation r = 0.02). All participants provided written informed consent prior to taking part and were paid for their participation.

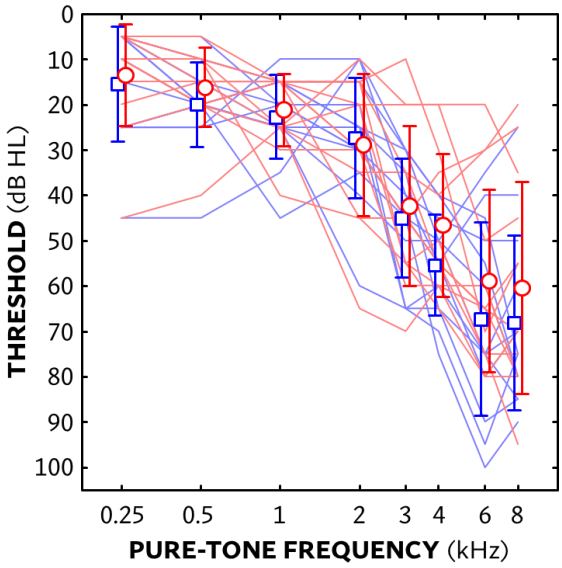


Figure 1. Pure-tone thresholds as a function of frequency for all participants left (blue lines) and right (red lines) ears as well as means (squares and circles for left and right ears, respectively) and standard deviations (error bars) at each frequency.

2.2 Stimuli

Stimuli consisted of 94 segments of stereo vocal music – 2 segments each from 47 pieces of music – that varied in genre, duration and word count. These segments, along with six other segments that were excluded as they were assessed as unpleasant in piloting, were used as they were made available from a previous lyric intelligibility study [11]. The broadly

very defined genres were classical (20 segments, comprised of single-voice, mostly light opera excerpts), folk (20), jazz (20), pop (14) and rhythm & blues (20). Segment durations were 1.65-16 s (median 6 s); some segments began and/or ended abruptly mid phrase. The number of scored words ranged from 3-19 (median 8), and word rates ranged from 0.31-3.64 words/s (median 1.5 words/s). Fidelity varied across segments despite all being originally from commercial recordings, and some segments began or ended abruptly mid phrase. Despite these shortcomings, the corpus was available and allowed direct comparison with a previous dataset.

Segments were all RMS normalised (without and frequency weighting) and processed using a simulation of hearing-aid non-linear amplification. Each channel (L/R) of each segment was first split into octave bands using cascaded Butterworth filters with centre frequencies of 250, 500, 1000, 2000, 4000 and 8000 Hz. Gains for each participants' ears were based on an A-weighted 65 dB presentation level using an implementation of the CAMFIT nonlinear prescription [14]-[15]. These individualised frequency-dependent gains were then applied to the filtered-band L/R signals. Compression ratios for each participants' ears were determined by the difference in prescribed outputs at 55- and 85-dB input [16] with an upper ratio limit of 6:1 imposed to avoid extreme values. Compression attack and release times (11-14 and 80-100 ms, respectively, across centre frequencies) were based on average values measured using IEC 60118 procedures in conventional nonlinear hearing aids [17]. Individualised compression was then applied to the filtered-band L/R signals. The filtered, amplified and compressed stimuli were then summed. Stimuli were presented through an external soundcard (RME Babyface) to circumaural headphones (AKG702) in a sound-dampened booth. Participant responses were monitored and recorded using a lapel microphone.

2.3 Procedure

In each trial, participants heard a segment that was presented twice with a 5-s interstimulus interval over headphones at an A-weighted level of 65 dB plus individualised frequency-dependent gain (CAMFIT-nonlinear). Participants could begin recalling lyrics after the first presentation. Segments were presented in a randomised order for each participant. Participant responses were recorded as well as marked in real time, then checked again after the experiment by the authors for marking accuracy. Lyrics for each segment were initially provided by [11] and then validated by the authors.

2.4 Prediction

Each music segment used in the study was processed independently for the left and right channels for each participant. First, gain and compression were applied as described in the Methods above. The amplified signals were then passed through a Python implementation of a gammatone-filterbank hearing-loss model [18]-[19] that simulates decreased audibility, dynamic range, and temporal and frequency resolution using solely the participants' pure-tone thresholds as input. Each processed signal was then transcribed using an automatic speech recognition system trained on a large multilingual dataset in various contexts including background music (Whisper; [20]). The Whisper model was calibrated to produce deterministic results (i.e., avoid hallucinations) by setting its temperature parameter to zero.

3. Results

3.1 Behavioural results

Performance was measured as the proportion of words recalled correctly [P(C)]. Performance varied greatly across music segments, from 0-100%, and varied as a function of genre, as shown in Figure 2. There was a significant effect of genre on individual mean performance [repeated-measure ANOVA $F_{(4,24.36)} = 142.73$; $p << 0.001$; $\eta^2 = 0.92$]. In comparison to previous studies [10]-[11], the overall P(C) was significantly lower than either of the previous studies: 0.65 vs. 0.72 in both studies [Student's $t_{(12)} = 3.07$; $p = 0.0098$].

By genre, the current mean results were significantly less than both previous studies for the pop and R&B genres [$t_{(12)} = 3.15$-6.19; all $p_{adj} < 0.01$], and also significantly less than [11] for the classical genre, and [10] for the jazz genre (n.b., which was near perfect performance). The mean P(C) for the jazz genre was significantly greater than it was in [11] using the same segments, but this was partially due to a particular stimulus (a segment of Nat King Cole singing "Rambling Rose") for which the previous study's participants recalled zero correct words, whereas the current participants on average were very accurate [P(C) = 0.94]. Considering that previous study [11] used the same stimuli with a different population, the results were reasonably well correlated ($r = 0.72$; Spearman's $\rho = 0.70$; $p <$

0.001), although there were several more segments like the one noted above with large discrepancies (e.g., 12 segments with interstudy differences > 0.33), and a majority of segments (70%) were more intelligible [ΔP(C) > 0] in the previous study. Individual mean results in the current study were significantly correlated with age [r = 0.83; p < 0.001 (ρ = 0.73; p = 0.0045)] but not BE4FA (r = -0.48; p = 0.099), although this may be due to the small n (i.e., underpowered).

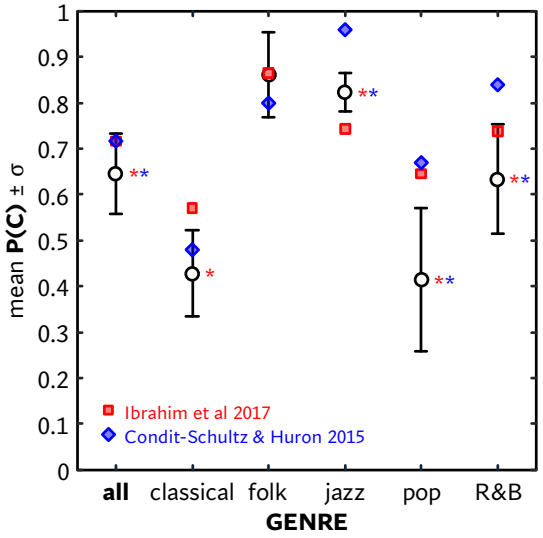


Figure 2. Mean proportion correct [P(C)] of lyrics recognised as a function of genre as compared to results for the same samples from Ibrahim et al. [11] (red squares) and different samples categorized into the same genres from Condit-Schultz and Huron [10] (blue diamonds). Error bars show ±1 standard deviation. Asterisks indicate significant differences from past studies. Note that the "pop" genre mean is based on 14 samples, all other genre means based on 20 samples.

Performance [P(C)] was generally not correlated with segment attributes: duration, word count nor word rate across all segments. However, as shown in Figure 3, P(C) was correlated with word rate within the pop genre. As highlighted by the filled symbols in Figure 3, there were some large P(C) differences between excerpts of the same song that had similar attributes. For example, there was a 59% difference in mean performance on the two "classical"-genre segments from Aaron Copland's The Tender Land (filled red circles). Both segments feature the same singer, but there are audible music-related differences: the highly intelligible segment has a sustained string background, whereas the much less intelligible segment has a more pronounced accompaniment that changes with each syllable, masking the consonants. Another example is the disparity between two "jazz"-genre segments of "This Masquerade Is Over" sung by Marvin Gaye (filled green downward triangles). The disparity in performance can be attributed to multiple factors: singer-related changes in enunciation, as well as word-related differences in between these samples that are both singer-related changes in enunciation and word-related changes in prevalence ("and so is love my darling" vs. "learn to laugh like Pagliacci with tears" for the more and less intelligible segments, respectively).

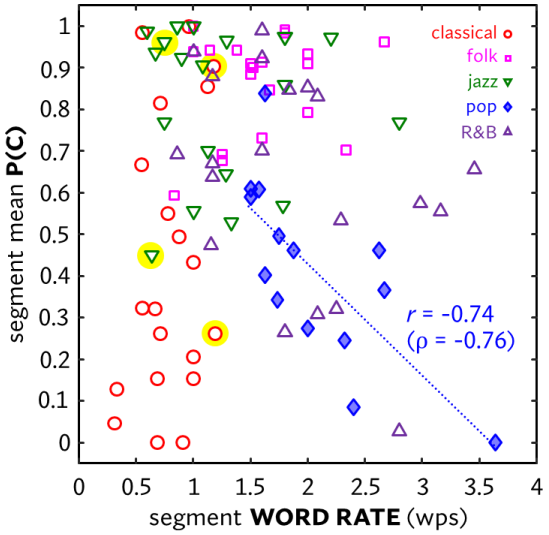


Figure 3. Mean proportion correct as a function of word rate for each segment with symbols indicating genre. Dotted (blue) line indicates linear regression for the pop genre (half-filled blue diamonds). The large differences between segments from the same recording described in the text are highlighted with yellow circles.

3.2 Prediction results

From the Whisper transcription, the number of correctly recognised words was computed without applying penalties for substitutions, insertions or deletions. The better score (proportion of correctly transcribed words) from the left and right channels was then used as the predicted performance for that participant. The use of the better-ear score emulated better-ear listening, well demonstrated in speech intelligibility but unknown for lyric intelligibility, as well as optimised the correlation with the listening test results.

Model results are shown in Figure 4. Predicted mean scores for each segment were well correlated with actual scores, both across and within genres except for the pop genre, for which intelligibility decreased

with increasing word rate. The model, however, underpredicted performance for the vast majority (88%) of the segments; the mean P(C) difference between actual and predicted mean segment scores was +0.27. By genre, the underprediction was greatest for the "jazz" genre (mean difference = +0.35) and least for the "classical" genre (mean difference = +0.19). Actual-predicted score differences were not correlated with segment attributes (i.e., duration, word count or rate; r^2 = 0.0061-0.015). By individual participant, predicted scores, averaged across segments, were not correlated with actual scores (r^2 = 0.007). Actual-predicted differences were negatively correlated with BE4FA (ρ = -0.60; p = 0.032), indicating that the prediction was generally worse for those with milder losses.

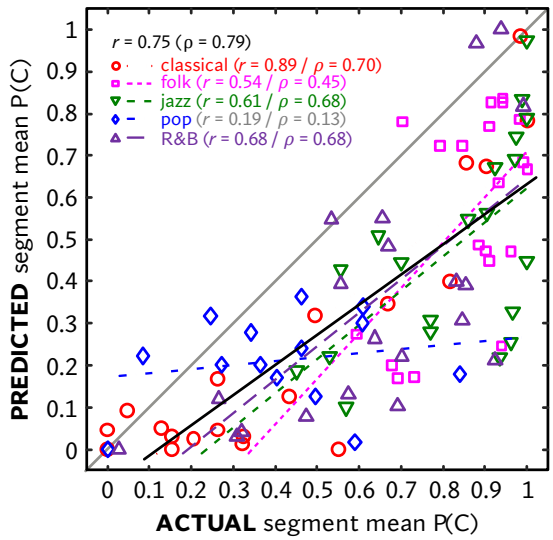


Figure 4. Predicted segment mean proportions correct [P(C)] as a function of actual segment mean P(C) for each segment. Genres are indicated by different symbols and colours. Linear regressions are shown for all genres; the solid black line showing the regression for all segments. The axes diagonal (grey line) is provided to better visualise the disparities from perfect prediction.

To better understand the potential sources of misprediction, we repeated the Whisper prediction (1) with only the individual hearing-loss simulation (i.e., without the nonlinear gain stage), and (2) without the hearing-loss simulation (i.e., without any processing). Overall results for these predictions along with the behavioural and fully processed predictions are shown in Table 1. Removing the gain stage, which should have partially compensated for reduced audibility introduced in the following hearing-loss simulation stage, only decreased the predicted mean performance by 1%. Removing the hearing-loss simulation processing – that is, the Whisper prediction for a listener without hearing loss – predicted poorer performance than what was achieved by the 13 individuals with hearing loss in the current study.

Table 1. Mean results showing changes in mean proportion correct P(C) and Pearson product-moment correlation coefficients (r^2) between actual and predicted mean scores for each segment.

data	mean P(C)	r^2
actual	0.65	--
predicted (gain + HL)	0.38	0.56
predicted (HL only)	0.37	0.64
predicted (no processing)	0.60	0.52

4 Discussion

Predicting the ability to recall lyrics for those with hearing loss poses myriad difficulties. In the current study, there was an absence of any general relationship between segment features and lyric intelligibility. This disparity is not limited to listeners with hearing loss; [10] only found a correlation between performance and log-duration of their excerpts; no other features correlated with performance in their study nor in [11]. Curiously, the one genre – "pop" – that showed any relationship to attributes (segment word rate; Figure 3) in the current study was also the only genre for which the Whisper-based prediction was poor. In [11] and a follow-up study [21], broad categories of (actual) intelligibility scores were used to inform and initially assess classifiers based on acoustic attributes. With the current results, such broad categories would be problematic due to the scale of underprediction [e.g., 38% of the segments had an actual P(C) > 0.5 and predicted P(C) < 0.5 whereas only 31% had an actual and predicted P(C) < 0.5]. In machine-learning endeavours, we want to include rather than exclude – to generalise our predictive capacity for real-world benefit. The current study, however, highlights the inherent difficulty in prediction when numerous factors (i.e., singer-, music-, word- and environment-related [9]) are involved. Homogeneity in sampling, however, is a challenge unto itself. Focusing on a single genre may not be enough to disentangle these factors, as here lyric intelligibility varied within song as well as within genre, and many features are shared across genres [22].

Another potential factor in this underprediction may be the hearing loss simulator, especially considering

that gain compensation was applied to the signals to partially account for loss in audibility. One potential explanation is that Whisper ASR is unlikely to have been trained on audio degraded with a hearing loss simulator, and consequently some of the underprediction could be an issue of out-of-domain generalisation. Another potential explanation is that Whisper's predictions may not consider the reduced sensation level due to hearing loss; that is, it normalises the input. This normalisation would then at least partially explain the lack of change due to applying nonlinear gain, but then the difference between predicted results with and without the HL simulator (see Table 1) would only be the result of Yet another consideration is that current hearing-loss models only use pure-tone thresholds as parameters, hence can only be expected to accurately simulate elevated thresholds, not supra-threshold abilities [23]. To better understand the role of the hearing-loss simulator in lyric intelligibility prediction, future work could present hearing-loss simulator processed stimuli to young normal-hearing listeners to compare performance to those with hearing loss. Another important facet for prediction is taking into consideration the age of the listener, although the negative correlation between age and performance may just be a result of the difficulty of the task (e.g., the short-term memory resources involved in having to recall sung words from a musical segment presented, albeit twice, out of context).

The current study was limited by having a small number of participants that were all older with varying but generally milder forms of hearing loss. There were no age-matched controls per se; however, the comparison with previous studies using self-reported normal-hearing students (Figure 2) shows that variance by participant group – an overall mean difference of 7% – is far less than the differences due to individual segments or genre. Another distinction of the current study was the older age of the participants. Within the older population, age was correlated with overall performance. The potential role of additional age-related factors in lyric intelligibility as separate from the measurement of intelligibility (i.e., the short-term memory demands of recall) warrant further study. The correlation with word count (Figure 3) and lack of prediction (Figure 4) for the "pop" genre, consisting of more recent recorded material also warrants further exploration as to the interaction between performance attributes (e.g., lyrical word rate), familiarity – both broadly (style) and specifically (performer) – and listener age and cognitive capacity.

5 Conclusion

The current study measured lyric intelligibility in a small sample of adults with hearing loss. An objective metric for lyric intelligibility for people with hearing loss was developed that combined hearing-loss simulation with Whisper ASR. The metric was developed for the Cadenza challenges, to allow researchers to train machine learning systems to improve the intelligibility of music lyrics. Such improvements could provide novel solutions for some of the problems faced by those with hearing loss when listening to sung music.

The objective metric did correlate with listening test results, accounting for 56% of the variability in segment intelligibility, but there was significant underprediction. To overcome the challenges in predicting lyric intelligibility, focus may need to shift from the intelligibility as defined as words correctly recalled, to the subjective impression of intelligibility – a song's lyrical clarity.

Acknowledgements

This work was supported by funding from the Engineering and Physical Sciences Research Council (EP/W019434/1) and Medical Research Council (MR/X003620/1).

References

[1] Creech A, Hallam S, McQueen H, Varvarigou M (2013) The power of music in the lives of older adults. *Res Studies Mus Ed* 35: 87–102. DOI: 10.1177/1321103X13478862.

[2] Ali SO, Peynircioğlu ZF (2006) Songs and emotions: Are lyrics and melodies equal partners? *Psych Mus* 34(4): 511–534. DOI: 10.1177/0305735606067168.

[3] Barradas GT, Sakka LS (2022) When words matter: A cross-cultural perspective on lyrics and their relationship to musical emotions. *Psych Mus* 50(2): 650–669. DOI: 10.1177/03057356211013390.

[4] Brattico E, Alluri V, Bogert B, Jacobsen T, Vartiainen N, Nieminen S, Tervaniemi MA (2011) Functional MRI study of happy and sad emotions in music with and without lyrics. *Frontiers Psych* 2: 308. DOI: 10.3389/fpsyg.2011.00308.

[5] Gabrielsson A (2011) Strong experiences with music: Music is much more than just music. Oxford, UK: Oxford University Press.

[6] Lamont A (2011). University students' strong experiences of music: Pleasure, engagement, and meaning. *Musicae Scientiae* 15(2): 229–249. DOI: 10.1177/1029864911403368

[7] Madsen SMK, Moore BCJ (2014) Music and hearing aids. *Trends Hear* 18. DOI: 10.1177/2331216514558271.

[8] Greasley A, Crook H, Fulford R (2020) Music listening and hearing aids: Perspectives from audiologists and their patients. *Int J Audiol* 59(9): 694-706. DOI: 10.1080/14992027.2020.1762126.

[9] Fine PA, Ginsborg J (2014) Making myself understood: perceived factors affecting the intelligibility of sung text. *Frontiers Psych* 5: 809. DOI: 10.3389/fpsyg.2014.00809.

[10] Condit-Schultz N, Huron D (2015) Catching the lyrics: Intelligibility in twelve song genres. *Music Perc* 32(5): 470-483.

[11] Ibrahim KM, Grunberg D, Agres K, Gupta C, Wang Y (2017) Intelligibility of sung lyrics: a pilot study. *Proc 18th ISMIR*, pp. 686-693.

[12] https://cadenzachallenge.org/.

[13] British Society of Audiology (2018) *Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking*. Retrieved from https://www.thebsa.org.uk/wp-content/uploads/2023/10/OD104-32-Recommended-Procedure-Pure-Tone-Audiometry-August-2018-FINAL-1.pdf.

[14] Moore BCJ, Alcántara JI, Stone MA, Glasberg BR (1999a) Use of a loudness model for hearing aid fitting II. Hearing aids with multi-channel compression. *Br J Audiol* 33(3): 157-170. DOI: 10.3109/03005369909090095.

[15] Moore BCJ, Glasberg BR, Stone MA (1999b) Use of a loudness model for hearing aid fitting III. A general method for deriving initial fittings for hearing aids with multi-channel compression. *Br J Audiol* 33(4): 241-258. DOI: 10.3109/03005369909090105.

[16] International Electrotechnical Commission (IEC) (2022) Electroacoustics - Hearing aids - Part 0: Measurement of the performance characteristics of hearing aids.

[17] Whitmer WM, Akeroyd MA (2011) Level discrimination of speech sounds by hearing-impaired individuals with and without hearing amplification. *Ear Hear* 32(3): 391-398. DOI: 10.1097/AUD.0b013e318202b620.

[18] Nejime Y, Moore BCJ (1997) Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise J Acoust Soc Am 102(1): 603–615.

[19] https://github.com/clarity challenge/clarity

[20] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I (2023) Robust speech recognition via large-scale weak supervision. *Proc Int Conf Machine Learning*, pp. 28492–28518.

[21] Sharma B, Wang Y (2020) Automatic evaluation of song intelligibility using singing adapted STOI and vocal-specific features. *IEEE/ACM Trans Audio Sp Lang Proc* 28: 319-331. DOI: 10.1109/TASLP.2019.2955253.

[22] Lesimple C, Kuehnel V, Siedenburg K (2024) Hearing aid evaluation for music: Accounting for acoustical variability of music stimuli. *JASA Express Lett* 4(9): 093201. DOI: 10.1121/10.0028397.

[23] Carney L (2018) Supra-threshold hearing and fluctuation profiles: Implications for sensorineural and hidden hearing loss. *J Assoc Res Otolaryngol* 19: 331-352. DOI: 10.1007/s10162-018-0669-5.