DOI: 10.61782/fa.2025.0677

FORUM ACUSTICUM
EURONOISE 2025

PERCEPTUAL EVALUATION OF DEEP LEARNING-BASED SPEECH ENHANCEMENT FOR HEARING AIDS

Martí Baig[1,2,3*] **Enric Gusó**[1,4] **Joanna Luberadzka**[1] **Umut Sayin**[1]
[1] Eurecat, Centre Tecnològic de Catalunya, Multimedia Technologies, Barcelona, Spain
[2] Universitat de Vic - Universitat Central de Catalunya, Faculty of Medicine, Barcelona, Spain
[3] Microson, Amplifon Group, Barcelona, Spain
[4] Universitat Pompeu Fabra, Music Technology Group, Barcelona, Spain

ABSTRACT

Hearing aids (HA) are widely used to compensate for hearing loss, although users often have difficulty understanding speech in complex environments, regardless of the several available signal processing algorithms. Recent advances in deep neural networks (DNNs) for speech processing suggest that these approaches could become a promising alternative to traditional HA technologies. In this study, we compare the performance of conventional speech enhancement algorithms used in commercially available HAs with that of DNN-based techniques. We generate speech in noise situations in an Ambisonics setup and record them with a dummy head. The signals are either enhanced using the HA default settings or using a DNN as a post-filter and presented to hearing-impaired and normal hearing individuals. We evaluate both causal and non-causal DNN variants, training the models to either fully remove or partially preserve reverberation using anechoic and pseudo-anechoic targets, which we refer to as strong and mild speech enhancement, respectively. We observe a significant preference for the mild models which are less prone to contain sound artifacts and distortion. Our results conclude that a mild speech enhancement DNN has the potential to improve HA performance in noisy environments.

*__Corresponding author__: marti.baig@amplifon.com.

Keywords: *hearing aids, listening test, speech enhancement, deep neural network.*

1. INTRODUCTION

HA users often find that noisy environments are the most unsatisfactory aspect of their experience [1], which may indicate that the performance of conventional HA speech enhancement algorithms is still limited in these challenging acoustic conditions.

This paper expands [2]: an objective evaluation of hearing aids and DNN-based binaural speech enhancement in complex acoustic scenes. In that study, three virtual Ambisonics sound environments (VSE) from the ARTE database [3] were simulated in the laboratory: a crowded office, a cocktail party, and a restaurant. A target speech signal from [4] was presented at a distance of one meter in front of the listener, with a signal-to-noise ratio (SNR) of +5 dB and an overall level of 70 dB SPL. Six conditions were created by combining three different acoustic environments with two target speaker positions: on-axis (0°) and off-axis (30°). This resulted in the following scenes: *office0*, *office30*, *party0*, *party30*, *restaurant0*, and *restaurant30*. The sound was reproduced using an Ambisonics loudspeaker setup surrounding a KU100 dummy head wearing HAs. Five receiver-in-canal HAs, configured for a flat 20 dB HL, were recorded in each scenario in two modes: *enabled*, with default signal enhancement features activated, and *bypass*, with all enhancement features disabled. The *bypass* recordings served as input to an offline DNN-based speech enhancement system. We evaluated two architectures: a non-causal model (SuDoRM-RF++GC) and a causal, HA-oriented model

(C-SuDoRM-RF++) [5]. The evaluation was based on four objective metrics: Hearing-Aid Speech Quality Index (HASQI) [6], Hearing-Aid Speech Perception Index (HASPI) [7], Modified Binaural Short-Time Objective Intelligibility (MBSTOI) [8] and Scale-invariant signal-to-distortion ratio (SISDR) as in [5]. We identified that conventional speech enhancement algorithms, such as digital noise reduction [9] and directional microphones or adaptive beamformers [10] face significant challenges under adverse acoustic conditions, sometimes even resulting in a reduced performance compared to the absence of any active algorithm. DNN-based approaches were observed to surpass traditional methods in terms of noise reduction and speech intelligibility, albeit with a trade-off in speech quality.

In this follow-up study, we perform a subjective evaluation of the different processing models with the addition of what we refer to as mild models. These are the same DNN models trained using a pseudo-anechoic target to remove the noise but preserve some reverberation, in contrast to the strong models trained on anechoic targets. The six types of processing models described in Tab. 1, which include the four DNN-based models and the two HA modes (*enabled* and *bypass*), tested across the six virtual acoustic scenes, serve as the material for the listening test.

This paper is structured as follows. We begin with a description of the subjective experimental setup, followed by statistical analysis of the results and participant metadata obtained via post-test questionnaires. We then present an objective evaluation using the HASPI and HASQI metrics and explore the relationship between the participant's preferences and the objectives scores. In the discussion section, we analyze group-specific differences, noting that both participant groups preferred DNN models with mild speech enhancement. We conclude by highlighting the relevance of these findings in supporting the potential of DNN-based enhancement models to overcome current limitations in HA performance under complex acoustic conditions.

2. MATERIALS AND METHODS

2.1 Participants

The study involved two groups of participants: individuals with hearing impairments (HI) who use HAs, and individuals with normal hearing (NH) who do not use HAs. Adult HA users were recruited based on the following inclusion criteria: moderate to severe sensorineural or mixed symmetrical hearing loss, daily use of the HAs, and binaurally fitted for more than six months. Twenty participants (7 female, 13 male; age range: 58–80 years, $M = 67.8 \pm 7.6$) completed the listening test, each wearing one of the HA models listed in [2]. In total, four high-end HA models with similar speech enhancement strategies were tested. The average binaural hearing loss of the participant, estimated using the American Medical Association (AMA) formula [11], was $M = 29.7 \pm 8.3$ %. None of the participants had previous experience with listening tests or sound quality rating procedures beyond standard clinical hearing assessments. The median of the audiograms demonstrated a sloping high-frequency hearing loss, characteristic of age-related sensorineural hearing loss (presbycusis).

The second group consisted of 21 NH individuals, out of which one participant was excluded due to self-reported hearing loss. The remaining participants either reported no hearing loss or were uncertain about it. The final group comprised 10 males and 10 females, with ages ranging from 19 to 69 years ($M = 38.8 \pm 11.6$).

Table 1. Overview of the different processing models.

Model	Type	Causal	Target
mildCausal	C-SuDoRM-RF++	✓	*pseudo*
strongCausal			*anech*
mildNon	SuDoRM-RF++GC	✗	*pseudo*
strongNon			*anech*
enabled	HA enh. ON	✓	NA
bypass	HA enh. OFF	✓	NA

2.2 Listening test setup

To assess user preference for different speech enhancement models, a Multi Stimulus with Hidden Reference and Hidden Anchors (MUSHRA)-like test was conducted [12]. This method allows participants to simultaneously rate multiple stimuli on screen, compare them easily, and replay them as needed, minimizing time and fatigue compared to traditional paired comparisons [13]. The reference signal was the *bypass* recording, with no speech enhancement features activated. Participants were informed of the reference as the baseline condition and they were asked to rate the six unlabeled models on a scale of 0 to 100, with 100 being the most preferred sound

to hear with their HA. No anchors were used. The test was double-blind using the web-based MUSHRA framework [14]. Before data collection, participants underwent a brief training session in which they familiarized themselves with the user interface and stimuli, and adjusted the volume to a comfortable level.

For the HI group, the listening test was conducted in a hearing center. The audio files were played from a computer and streamed directly to their HAs, with the enhancement algorithms disabled, allowing only hearing loss compensation. The session lasted approximately 45 minutes per participant and was conducted with the assistance of a researcher. In contrast, the NH group participated remotely, using closed headphones in a quiet setting, and completed the test in an average of 25 minutes without direct supervision. For this group, the audio files were based on recordings from a single HA. All participants completed a questionnaire after the test to collect statistics on the use of HA and their experience with the listening task.

2.3 Statistical Analysis

Tab. 2 provides an overview of the test variables, including the factors and their corresponding levels used in the statistical analysis. To evaluate the effects of experimental conditions on participant responses, we performed independent Friedman tests for the within-subject factors (*model* and *scene*, each with six levels) and a Kruskal-Wallis test for the between-subject factor (*group*: HI vs. NH). The Friedman test was selected as a non-parametric alternative to the repeated-measures ANOVA, as it does not assume normality or homogeneity of variances and is suitable for analyzing ranked or continuous but non-normally distributed data. Given the ordinal nature of some responses and the non-Gaussian distribution observed in preliminary data checks, the Friedman test provided a robust method for detecting differences across conditions. Similarly, we employed the Kruskal-Wallis test as a non-parametric alternative to one-way ANOVA. All statistical tests were conducted at a significance level of $\alpha = 0.05$, consistent with the standard threshold commonly adopted in behavioral and perceptual research [13].

2.4 Objective evaluation

HASPI and HASQI were used to predict speech intelligibility and quality, respectively. Although both are intrusive metrics, they differ from alternatives such as MB-STOI or SI-SDR by incorporating an auditory periphery model that accounts for individual hearing thresholds and simulates key physiological processes, including those of the middle ear, cochlea, basilar membrane, and hair cells. Scores range from 0 to 1, with 1 indicating maximum intelligibility or quality. We used the best ear scores and normalized the signals following the procedure in [15]. Unlike our previous study, we included individualized hearing thresholds in the metric computations: measured audiograms for the HI group and age- and sex-based estimations for the NH group [16]. Finally, we evaluated the relationship between the objective scores and the participant ratings.

Table 2. Listening test variables and conditions.

Factor	Type	Variable	Levels
Score	Continuous	Dependent	0-100
Group	Categorical	Independent	*HI*
			NH
Model	Categorical	Independent	*enabled*
			mildCausal
			mildNon
			strongCausal
			strongNon
			bypass
Scene	Categorical	Independent	*office0*
			office30
			party0
			party30
			restaurant0
			restaurant30

3. RESULTS

3.1 Post-test questionnaires

All participants completed a questionnaire at the end of the experiment. Fig. 1 presents the responses for each group. Among participants with hearing impairment, 60% had used HAs for 1-2 years, 15% for 3-4 years, and 15% for more than 4 years, with 10% not used them for more than a year. Most reported regular use, with 50% almost always, 35% always, and 15% sometimes. Satisfaction with their HAs was high, with 70% satisfied and 20% very satisfied, although all reported difficulties understanding speech in noisy environments. Regarding the hearing test,

65% of the HI group found it quite easy, while 35% of the NH group reported it as quite easy or very easy. However, 10% of the NH group found the test quite difficult, probably due to the lack of on-site assistance. In terms of the realism of sound environments, 60% of the HI group and 65% of the NH group found them quite realistic. However, 10% of the HI participants considered the scenes somewhat unrealistic, suggesting that HA users may be more critical or sensitive to simulated complex speech-in-noise scenarios.

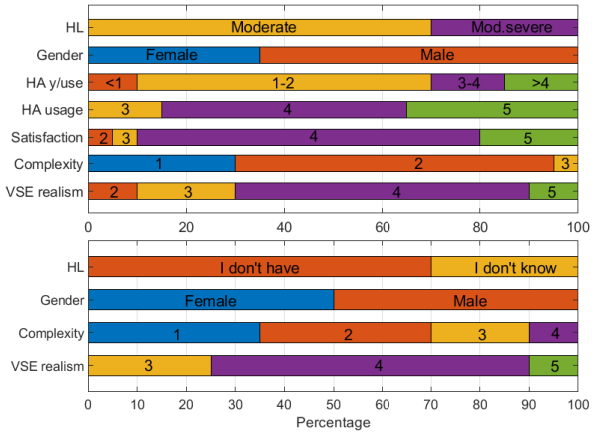

Figure 1. Questionnaire responses from the HI (top) and NH (bottom) groups. HL: degree of hearing loss; HA usage: average hours/day. Responses based on 5-point Likert scales: HA satisfaction, test complexity, and VSE realism, ranging from 1 (very dissatisfied/easy/unrealistic) to 5 (very satisfied/difficult/realistic).

3.2 Preference scores

We depict in Fig. 2 the preference scores for each speech enhancement model rated by the participants in both groups. The left and right halves of the violin plot represent the HI and NH groups, respectively. The *mildNon* model emerged as the most effective for both groups, achieving the highest median scores of 85.5 for the HI group and 77 for the NH group, alongside relatively stable responses (IQR of 29.5 and 25.5, respectively). The *mildCausal* model was the second highest rated, with median scores of 72.5 and 70 for the HI and NH groups, respectively, and IQR of 33 and 28. The *enabled* model demonstrated commendable performance for both groups, with median scores of 65 for the HI group and 69 for the NH group, although it showed considerable variability in responses (IQR of 38.5 and 36.5, respectively). In contrast, the *strongCausal* and *strongNon* models underperformed in both groups, with *strongCausal* recording the lowest median scores (35.5 for HI and 12.5 for NH), indicating a failure to meet the participants' expectations. Similarly, *strongNon* recorded low median scores (53 for HI and 28 for NH) and exhibited substantial variability in ratings. The *bypass* model demonstrated moderate performance for both groups, with a median score of 56 for the HI group and 60 for the NH group, and moderate variability (IQR of 34 and 30.5, respectively). In general, HI participants exhibited greater variability in their scores (larger IQR) compared to NH participants, particularly for models such as *strongNon* and *strongCausal*.

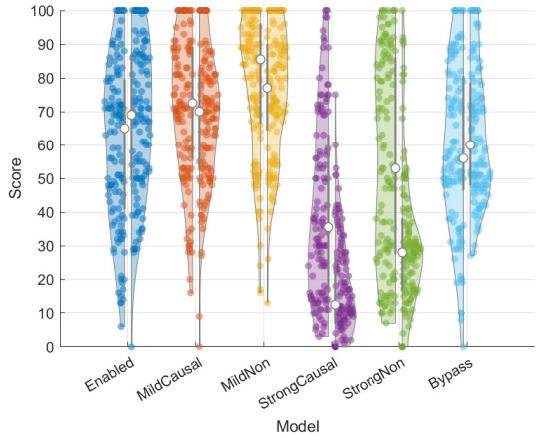

Figure 2. Preference scores for each processing model (HI: left, NH: right of each violin). White dots indicate the median and the IQR is represented by a gray vertical line.

Tab. 3 presents the ranking and percentages of participants in each group who selected each model as their top preference across the scenes. The findings indicate that the HI group exhibited a broader distribution of preferences, with *mildNon* being the most popular model (44%), followed by *mildCausal* (18%) and *strongNon* (16%). In contrast, the NH group preferred predominantly *mildNon* (36%), followed by *enabled* (31%) and *mildCausal* (20%). Both groups demonstrated lower preferences for *bypass* (HI: 5%, NH: 8%) and *strongCausal* (HI: 4%, NH: 0%).

Table 3. Ranking and percentage of top preferences for processing models for HI and NH groups.

Rank	HI		NH	
	Model	**Pref.(%)**	**Model**	**Pref.(%)**
1	*mildNon*	44	*mildNon*	36
2	*mildCausal*	18	*enabled*	31
3	*strongNon*	16	*mildCausal*	20
4	*enabled*	13	*bypass*	8
5	*bypass*	5	*strongNon*	6
6	*strongCausal*	4	*strongCausal*	0

The results of the statistical tests revealed significant main effects for all factors examined. For the within-subject factor *model*, the Friedman test yielded a statistically significant result $\chi^2(5) = 132.76$, $p < .001$, indicating that listener responses varied between different speech enhancement strategies. Similarly, a significant effect was observed for *scene* ($\chi^2(5) = 395.02$, $p < .001$), suggesting that the complexity or realism of the scene had a measurable impact on the ratings. The Kruskal-Wallis test revealed a significant main effect of the *group* ($\chi^2(1) = 21.53$, $p < .001$), indicating that the response patterns differed significantly between the participants in each group. To further investigate these effects, we performed post hoc analyzes using Tukey's Honest Significant Difference (HSD) criterion for pairwise comparisons.

Fig. 3 and Fig. 4 present the results of the pairwise comparisons for the HI and NH groups, respectively. Within the *HI* group, significant differences were identified between *mildCausal* and *strongCausal*, as well as between *mildNon* and *strongCausal*. In the *NH* group, several pairwise differences were observed: *enabled* differed significantly from both *strongCausal* and *strongNon*; *mildCausal* differed significantly from both *strongCausal* and *strongNon*; *mildNon* differed significantly from both *strongCausal* and *strongNon*; and *bypass* differed significantly from both *strongCausal* and *strongNon*. These results suggest that the *strongCausal* and *strongNon* models consistently elicited distinct response patterns compared to the other conditions, particularly within the NH group, where a wider range of contrasts achieved significance. Although the Friedman test indicated a significant main effect for the *scene* factor, none of the pairwise comparisons reached statistical significance after correcting for multiple comparisons in any group of participants. This suggests that the differences between individual *scene* conditions were not substantial enough to be detected when controlling for the family-wise error rate.

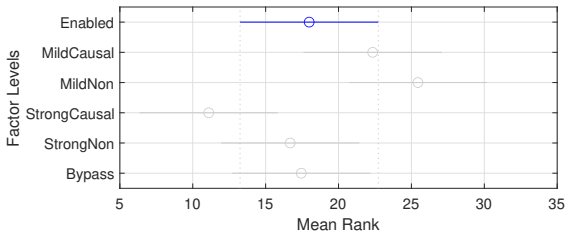

Figure 3. Pairwise comparisons of mean ranks among processing models in the HI group.

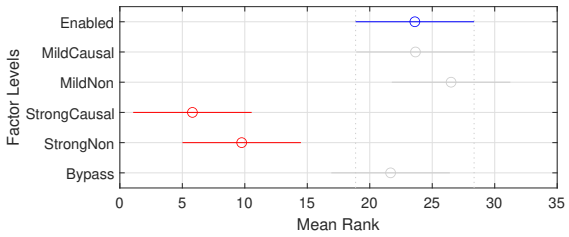

Figure 4. Pairwise comparisons of mean ranks among processing models in the NH group.

No significant interactions were observed between the different models and scenes.

The effect size, calculated using Hedge's g-test, averaged 0.65 for the HI group, indicating a moderate effect, and 1.49 for the NH group, indicating a large effect.

3.3 Speech intelligibility and quality

Fig. 5 and Fig. 6 show the violin plots of the HASPI and HASQI scores, respectively, for each model and participant group. For HASPI in the HI group, the *mildNon* model had the highest median score (0.56), followed by *mildCausal* and *strongNon* (0.54), *bypass* (0.51), *strongCausal* (0.52), and *enabled* (0.44). IQR values ranged from 0.08 for *strongNon* to 0.35 for *enabled*. In the NH group, the *enabled* model had the highest median score (0.72), followed by *mildNon* (0.64), *mildCausal* (0.61), *strongNon* (0.60), *strongCausal* (0.55), and *bypass* (0.51). IQR values ranged from 0.13 for *strongNon* to 0.44 for *bypass*. For HASQI in the HI group, the *mildNon* model had

the highest median score (0.22), followed by *mildCausal* (0.21), *strongNon* (0.20), *bypass* and *strongCausal* (0.19), and *enabled* (0.17). The IQR ranged from 0.06 for *bypass* to 0.09 for *mildNon*. In the NH group, both *mildNon* and *enabled* had a median score of 0.12, followed by *strongCausal*, *strongNon*, and *mildCausal* with a median of 0.11 each. *bypass* had the lowest median score (0.09). IQR values ranged from 0.02 for *strongNon* to 0.05 for *mildCausal* and *strongCausal*. In general, in the HI group, *mildNon* performed the best for both HASPI and HASQI in terms of median scores. In the NH group, *enabled* performed the best for HASPI, while *mildNon* and *enabled* showed similar performance for HASQI. IQR values indicate that the HI group generally exhibited more variability between models compared to the NH group. This is likely due to the fact that the NH group's recordings came from the same HA model, the only difference being the input audiogram used for the metrics calculation.

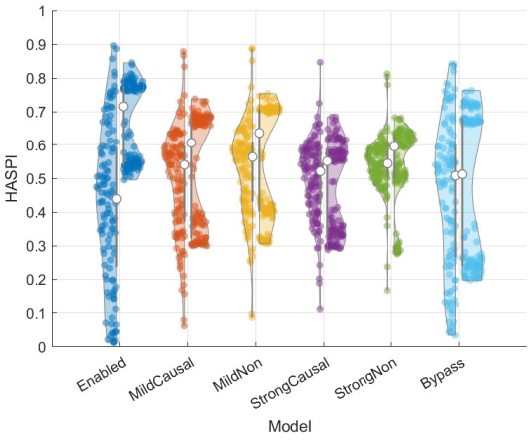

Figure 5. HASPI scores by processing model (HI: left, NH: right of each violin).

3.4 Correlation of metrics

The association between the metrics HASQI and HASPI and the ratings of the participants was assessed using Spearman's rank order correlation coefficient (ρ). *enabled* was the only speech enhancement model showing a significant correlation with both metrics. For HASPI, the correlation was positive with Spearman's ρ of 0.35 ($p < 0.001$), while for HASQI, the correlation was 0.28 at the same significance level.

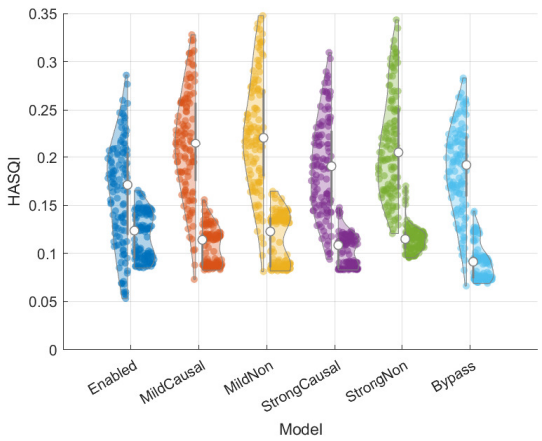

Figure 6. HASQI scores by processing model (HI: left, NH: right of each violin).

4. DISCUSSION

In this study, we extended the work presented in [2] by conducting a listening test to examine the hypothesis that DNN-based speech enhancement techniques can outperform traditional HA enhancement algorithms in challenging acoustic environments. In addition, we performed an objective evaluation using HASPI and HASQI metrics where, unlike [2], we included the participant's audiograms (HI group) or an age- and sex-based estimation (NH group). Finally, we assessed the association between objective metrics and subjective preference scores.

The *mildNon* model emerged as the most preferred model for both groups. It achieved the highest median scores with relatively stable responses, indicating consistent satisfaction among the participants, and was selected as the first choice in both groups. Its success can be attributed to its ability to strike a balance between enhancing speech intelligibility and preserving natural sound characteristics and spatial perception. In second place is the *mildCausal* model, which stands out as a promising candidate for HA use due to its real-time suitability, unlike the non-causal version. It was the second most preferred model for the HI group and the third for the NH group. In contrast, the strong models performed significantly worse in both groups, *strongCausal* being the least preferred. These models were criticized for overly aggressive noise suppression, resulting in lower speech quality and loss of spatial perception, which most participants perceived as unnatural. The *enabled* model performed moderately well,

although with high variability, and was the fifth preferred model in the HI group, surprisingly being more preferred in the NH group. The variability in the ratings suggests that while the *enabled* model worked well for some users, its performance was not consistently appreciated in complex environments, particularly for the HI group. Finally, the *bypass* model was moderately rated by both groups, with a small preference in general, but it still outperformed the strong models. The higher variability within the HI group suggests that hearing-impaired listeners have more diverse needs and may benefit from more personalized HA settings, while the NH group exhibited a clearer preference for models that preserve natural sound quality. Statistical analysis revealed significant main effects for the speech enhancement model and the participant group. Post hoc pairwise comparisons indicated that the strong models generated statistically significant differences compared to the scores of the mild models in both groups. Although the scene factor initially showed a significant effect, it did not influence the model preferences after controlling for multiple comparisons. The results of the HI group indicated a moderate effect size (0.65), while the NH group had a larger effect size (1.49), suggesting clearer distinctions between models.

The objective evaluation revealed that the *mildNon* model performed the best in both HASPI and HASQI evaluations within the HI group, followed closely by the *mildCausal* and *strongNon* models. In the NH group, the *enabled* model achieved the highest HASPI, while both *mildNon* and *enabled* exhibited the highest HASQI. However, the *enabled* model performed the worst in the HI group, even trailing behind the *bypass* model. This suggests that the performance of traditional speech enhancement algorithms can be detrimental in complex and noisy environments. The *strongCausal* and *strongNon* models performed moderately well in HASPI, surpassing *enabled* in the HI group and *bypass* in both groups, but performed worse in HASQI due to the aggressive suppression of noise and reverberation, which, while effective, compromises natural sound quality and introduces sound artifacts. In particular, the *enabled* model was the only one to show a statistically significant correlation with the participant ratings and HASPI and HASQI. Although the correlation was positive, it remained relatively modest, indicating a mild association between objective evaluation and subjective ratings. HASPI and HASQI, are designed to correlate with standardized hearing tests [17], further reinforcing their relevance in assessing HA performance.

5. CONCLUSIONS

We conclude that our findings support the potential of DNN-based methods to outperform traditional HA algorithms in ecologically valid complex scenarios, particularly when trained using pseudo-anechoic targets. Although statistical significance was not achieved between conventional processing techniques and DNN-based models, the latter emerged as the preferred choice among participants, particularly the mild speech enhancement models that preserved environmental sound details and natural sound qualities. The results demonstrate that DNN models can enhance speech intelligibility and quality, although their effectiveness depends on the training strategy, with better outcomes achieved through less aggressive noise suppression and dereverberation. The different virtual sound environments simulated did not significantly influence the preference for the speech enhancement strategy. Overall, the findings highlight the strong potential of DNN-based enhancement models to advance future HA technologies, albeit their practical integration remains challenging, especially in terms of power efficiency, real-time processing capabilities, device miniaturization, and consistent preservation of speech quality. Building on the feedback from this experiment, future work will focus on exploring the potential of spatial audio and other technologies such as virtual or augmented reality to enhance the realism and ecological validity of simulated speech-in-noise scenarios. These technologies remain relatively underexplored in hearing science, particularly with regard to their application in individual hearing evaluation and HA fitting.

6. ETHICS STATEMENT

This study was part of a Master's thesis at the University of Salford, with ethical approval granted by the University Ethics Panel (application ID: 8231). Participants were fully informed about the specifics of the experiment and their rights, received an information sheet, and provided their written consent. The data collected was completely anonymous and participants were informed of their right to withdraw from the study at any time.

7. CONFLICT OF INTEREST

The authors declare that the research was conducted without commercial or financial affiliations that could be perceived as potential conflicts of interest.

8. ACKNOWLEDGMENTS

We would like to thank Trevor Cox, James Kates, and the colleagues at the GAES hearing centers and headquarters for their support and assistance.

9. REFERENCES

[1] Anovum, "Results eurotrak españa 2023," tech. rep., EHIMA, 2023.

[2] E. Gusó, J. Luberadzka, M. Baig, U. Sayin, and X. Serra, "An objective evaluation of hearing aids and dnn-based binaural speech enhancement in complex acoustic scenes," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.

[3] A. Weisser, J. M. Buchholz, C. Oreinos, J. Badajoz-Davila, J. Galloway, T. Beechey, and G. Keidser, "The ambisonic recordings of typical environments (arte) database," *Acta Acustica United With Acustica*, vol. 105, no. 4, pp. 695–713, 2019.

[4] V. Aubanel, M. L. G. Lecumberri, and M. Cooke, "The sharvard corpus: A phonemically-balanced spanish sentence resource for audiology," *International journal of audiology*, vol. 53, no. 9, pp. 633–638, 2014.

[5] E. Tzinis, Z. Wang, X. Jiang, and P. Smaragdis, "Compute and memory efficient universal sound source separation," *Journal of Signal Processing Systems*, vol. 94, no. 2, pp. 245–259, 2022.

[6] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (hasqi) version 2," *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[7] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[8] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[9] T. Ricketts and S. Dhar, "Comparison of performance across three directional hearing aids," *Journal of the American Academy of Audiology*, vol. 10, no. 04, pp. 180–189, 1999.

[10] H. Gode and S. Doclo, "Adaptive dereverberation, noise and interferer reduction using sparse weighted linearly constrained minimum power beamforming," in *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 95–99, IEEE, 2022.

[11] R. A. Dobie, "The ama method of estimation of hearing loss," *Ear and Hearing*, vol. 32, no. 6, pp. 680–682, 2011.

[12] ITU-R, *Method for the subjective assessment of intermediate quality level of audio systems*, 2015. In RECOMMENDATION BS.1534-3 (10/2015) (Vol. BS.1534-3).

[13] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*. Chichester, UK: John Wiley & Sons, 2006.

[14] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra — a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, p. e187, 2018.

[15] M. A. Akeroyd, W. Bailey, J. Barker, T. J. Cox, J. F. Culling, S. Graetzer, G. Naylor, Z. Podwińska, and Z. Tu, "The 2nd clarity enhancement challenge for hearing aid speech intelligibility enhancement: Overview and outcomes," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[16] International Organization for Standardization, "Acoustics — statistical distribution of hearing thresholds related to age," Standard ISO 7029:2017, ISO, 2017.

[17] J. M. Kates and K. H. Arehart, "An overview of the haspi and hasqi metrics for predicting speech intelligibility and speech quality for normal hearing, hearing loss, and hearing aids," *Hearing Research*, vol. 426, p. 108608, 2022.