Controllable joint noise reduction and hearing loss compensation using a differentiable auditory model

Philippe Gonzalez[1,†], *Torsten Dau*[2], *Tobias May*[2]

[1]Demant A/S, Denmark
[2]Department of Health Technology, Technical University of Denmark

phgo@demant.com, tdau@dtu.dk, tobmay@dtu.dk

Abstract

Deep learning-based hearing loss compensation (HLC) seeks to enhance speech intelligibility and quality for hearing impaired listeners using neural networks. One major challenge of HLC is the lack of a ground-truth target. Recent works have used neural networks to emulate non-differentiable auditory peripheral models in closed-loop frameworks, but this approach lacks flexibility. Alternatively, differentiable auditory models allow direct optimization, yet previous studies focused on individual listener profiles, or joint noise reduction (NR) and HLC without balancing each task. This work formulates NR and HLC as a multi-task learning problem, training a system to simultaneously predict denoised and compensated signals from noisy speech and audiograms using a differentiable auditory model. Results show the system achieves similar objective metric performance to systems trained for each task separately, while being able to adjust the balance between NR and HLC during inference.

Index Terms: differentiable auditory model, noise reduction, hearing loss compensation

1. Introduction

Hearing impairment affects millions of people worldwide, impacting social interactions, cognitive function, and overall quality of life. While hearing aids can mitigate these challenges, users often report difficulties understanding speech in complex acoustic environments. Currently, NR and HLC in hearing aids primarily rely on simplistic algorithms like beamforming and frequency band amplification. However, deep learning offers the potential to surpass these traditional methods, as it allows complex non-linear mappings between noisy speech, listener profiles, and optimal HLC strategies.

Unlike conventional tasks such as speech enhancement where defining a training objective using the ground-truth target is straightforward, HLC is more challenging due to the absence of such a target. Recent studies have tackled this by training auxiliary neural networks to emulate non-differentiable auditory peripheral models [1, 2, 3]. These emulators were then used in closed-loop optimization frameworks [4, 5, 6, 7]. While this approach facilitates HLC algorithm training, it lacks flexibility, since retraining is required whenever the auditory model is updated. Additionally, most studies have trained these systems for individual listener profiles, using auditory models describing detailed physiological processes, some of which may not be necessary for achieving perceptual benefits in hearing aids.

Another approach involves designing differentiable auditory models that can directly be integrated into the optimization of the HLC algorithm. This makes it easier to explore which auditory model stages are essential for delivering perceptual benefits. Studies adopting this approach [8, 9, 10] have shown that trained HLC algorithms can outperform traditional hearing aid prescriptions in both quiet [8] and noisy [9] conditions, as measured by objective metrics. However, these efforts often utilize simplistic systems with very few trainable parameters, such as fixed finite impulse response (FIR) filterbanks with learnable gains, and they continue to train systems for individual listener profiles.

Recent studies have trained listener-independent systems for NR and HLC [11, 12, 13]. However, the proposed algorithms are unable to perform NR or HLC in a controllable manner during inference. Studies have shown that there exist sub-populations of HI listeners who prefer strong NR over mild NR [14, 15]. Additionally, users may want to prioritize NR or HLC depending on the listening environment. For example, in very noisy environments, users may wish for strong NR, while in social gatherings, they may prefer HLC without NR. Therefore, the ability to adjust the balance between NR and HLC during inference can be a valuable feature in hearing aids.

In this work, we propose a multi-task learning framework for joint NR and HLC. A differentiable auditory model is used for training a speech processor to simultaneously predict a denoised and a compensated signal from a noisy input speech signal and an audiogram. Each task is assigned a distinct training objective, and these objectives are combined using an uncertainty-based weighting scheme [16]. During inference, the system can flexibly mix the two output signals to enable controllable joint NR and HLC. Code is available online at https://github.com/philgzl/cnrhlc.

2. Typical framework

Figure 1 shows a typical framework for training a speech processor $\mathcal{F}_\theta$ for NR-only, HLC-only, or joint NR and HLC without control. The speech processor input is noisy speech x and its output is fed to a normal hearing (NH) or hearing impaired (HI) differentiable auditory model $\mathcal{A}_{\mathrm{NH}}$ or $\mathcal{A}_{\mathrm{HI}}$. If the task includes HLC, $\mathcal{F}_\theta$ and $\mathcal{A}_{\mathrm{HI}}$ also take an audiogram a as input. The target signal is the corresponding clean speech y or the same noisy speech x, and is fed to a NH auditory model $\mathcal{A}_{\mathrm{NH}}$. The speech processor is then optimized to minimize a loss function ℓ between the output of the two auditory models. Whether the auditory model at the output of the speech processor is NH or HI and the target signal is clean or noisy depends on the task:

- If the auditory model at the output of the speech processor is NH and the target signal is clean, then the speech processor is optimized for NR-only. The training objective is

$$\mathcal{L}_{\mathrm{NR}} = \ell\Big(\mathcal{A}_{\mathrm{NH}}\big(\mathcal{F}_\theta(x)\big), \mathcal{A}_{\mathrm{NH}}(y)\Big). \qquad (1)$$

This is similar to traditional speech enhancement, except that

---
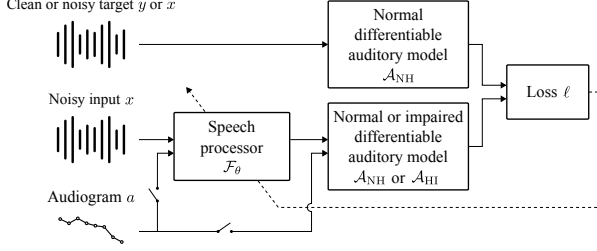
Figure 1: *Typical framework for NR-only, HLC-only, or joint NR and HLC without control.*

Figure 2: *Proposed framework for controllable joint NR and HLC.*

the training objective is based on an auditory model instead of e.g. signal-to-distortion ratio (SDR). For this task, no audiogram is fed to the speech processor nor the auditory model at its output. While denoising can improve the intelligibility and quality of speech, HI listeners require additional HLC.

- If the auditory model at the output of the speech processor is HI and the target signal is noisy, then the speech processor is optimized for HLC-only,

$$\mathcal{L}_{\mathrm{HLC}} = \ell\big(\mathcal{A}_{\mathrm{HI}}\big(\mathcal{F}_\theta(x, a), a\big), \mathcal{A}_{\mathrm{NH}}(x)\big). \qquad (2)$$

The speech processor is tasked with compensating for the hearing loss modeled in the auditory model at its output. However, background noise is not removed, which can be detrimental to the intelligibility and quality of speech.

- If the auditory model at the output of the speech processor is HI and the target signal is clean, then the speech processor is optimized for joint NR and HLC,

$$\mathcal{L}_{\mathrm{NR\text{-}HLC}} = \ell\big(\mathcal{A}_{\mathrm{HI}}\big(\mathcal{F}_\theta(x, a), a\big), \mathcal{A}_{\mathrm{NH}}(y)\big). \qquad (3)$$

However, since a single loss term is used, it is unclear if the speech processor prioritizes NR or HLC, and it is not possible to control for each task at inference time.

3. Proposed framework

Figure 2 shows the proposed framework for training the speech processor $\mathcal{F}_\theta$ for controllable joint NR and HLC. Given noisy speech x and an audiogram a, the speech processor outputs both a denoised signal $\hat{y}_{\mathrm{NR}}$ and a compensated signal $\hat{y}_{\mathrm{HLC}}$,

$$\mathcal{F}_\theta(x, a) = (\hat{y}_{\mathrm{NR}}, \hat{y}_{\mathrm{HLC}}). \qquad (4)$$

A loss term is defined for each task,

$$\mathcal{L}_{\mathrm{NR}} = \ell\big(\mathcal{A}_{\mathrm{NH}}(\hat{y}_{\mathrm{NR}}), \mathcal{A}_{\mathrm{NH}}(y)\big), \qquad (5)$$

$$\mathcal{L}_{\mathrm{HLC}} = \ell\big(\mathcal{A}_{\mathrm{HI}}(\hat{y}_{\mathrm{HLC}}, a), \mathcal{A}_{\mathrm{NH}}(x)\big). \qquad (6)$$

The final training objective can be defined as a weighted sum of $\mathcal{L}_{\mathrm{NR}}$ and $\mathcal{L}_{\mathrm{HLC}}$. However, finding the optimal weights using a grid search is time-consuming. Moreover, the results can be very sensitive to the choice of the weights, and the optimal weights can vary during training. One option is to model the predictions as isotropic Gaussian distributions, and adjust the weights dynamically based on the uncertainty of the predictions [16]. In practice, the method consists in optimizing two additional parameters $u_{\mathrm{NR}} = \log \sigma_{\mathrm{NR}}^2$ and $u_{\mathrm{HLC}} = \log \sigma_{\mathrm{HLC}}^2$, where $\sigma_{\mathrm{NR}} > 0$ and $\sigma_{\mathrm{HLC}} > 0$ represent the homoscedastic uncertainty related to each task. The final training objective is

$$\mathcal{L}_{\mathrm{C\text{-}NR\text{-}HLC}} = \frac{\mathcal{L}_{\mathrm{NR}}}{e^{u_{\mathrm{NR}}}} + u_{\mathrm{NR}} + \frac{\mathcal{L}_{\mathrm{HLC}}}{e^{u_{\mathrm{HLC}}}} + u_{\mathrm{HLC}}. \qquad (7)$$
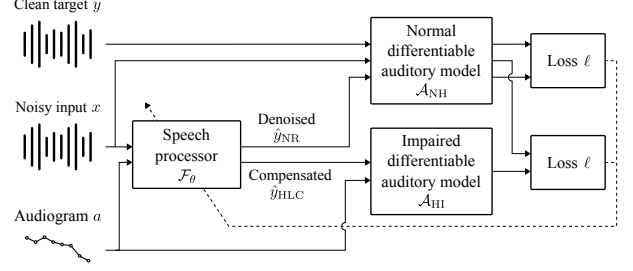
Intuitively, each loss term is weighted down if the related uncertainty is high. At the same time, since the log-uncertainties are added to the final training objective, they are encouraged to be as small as possible.

Similar to how unprocessed and beamformed signals are mixed in hearing devices [17], the denoised and compensated signals are mixed using a parameter $\alpha \in [0, 1]$,

$$\hat{y} = \alpha\hat{y}_{\mathrm{NR}} + (1 - \alpha)\hat{y}_{\mathrm{HLC}}. \qquad (8)$$

This allows the balance between NR and HLC to be adjusted without the need to retrain the speech processor.

4. Experimental setup

4.1. Speech processor

The speech processor is based on the band-split recurrent neural network (BSRNN) [18]. BSRNN achieves state-of-the-art results for speech enhancement [19, 20], and provides an excellent trade-off between computational complexity and performance [21]. Figure 3 shows an overview of the speech processor architecture. The speech processor takes as input the short-time Fourier transform (STFT) of the noisy speech $X \in \mathbb{C}^{F \times T}$, where F is the number of frequency bins and T is the number of frames. The band-split module transforms the real and imaginary part of K pre-defined frequency bands into a fixed number of channels N using band-specific fully connected layers. Dual-path modelling [22] across time frames and frequency bands is performed using residual long short-term memory (LSTM) blocks in L layers. Features are extracted from the audiogram using a fully connected layer followed by Tanh activation, and fed to each layer using FiLM conditioning [23]. The mask estimation module predicts both a mask and a residual spectrogram similarly to [20] for each output signal using band-specific multilayer perceptrons (MLPs). The mask and residual spectrogram for the denoised signal are denoted as $M_{\mathrm{NR}} \in \mathbb{C}^{F \times T}$ and $R_{\mathrm{NR}} \in \mathbb{C}^{F \times T}$, respectively. The mask and residual spectrogram for the compensated signal are denoted as $M_{\mathrm{HLC}} \in \mathbb{C}^{F \times T}$ and $R_{\mathrm{HLC}} \in \mathbb{C}^{F \times T}$, respectively. The STFT of the denoised and compensated signals $\hat{Y}_{\mathrm{NR}}$ and $\hat{Y}_{\mathrm{HLC}}$ are calculated as

$$\hat{Y}_{\mathrm{NR}} = M_{\mathrm{NR}} \odot X + R_{\mathrm{NR}}, \qquad (9)$$

$$\hat{Y}_{\mathrm{HLC}} = M_{\mathrm{HLC}} \odot X + R_{\mathrm{HLC}}. \qquad (10)$$

The STFT uses a frame length of 32 ms, a hop size of 16 ms, and a Hann window. We use $K = 27$ frequency bands with a bandwidth of 200 Hz between 0 and 4 kHz, 500 Hz between 4 and 7 kHz, and 1 kHz between 7 and 8 kHz. The number of channels is $N = 64$, and the number of layers is $L = 6$. The number of trainable parameters varies between 4.1 and 4.7 M
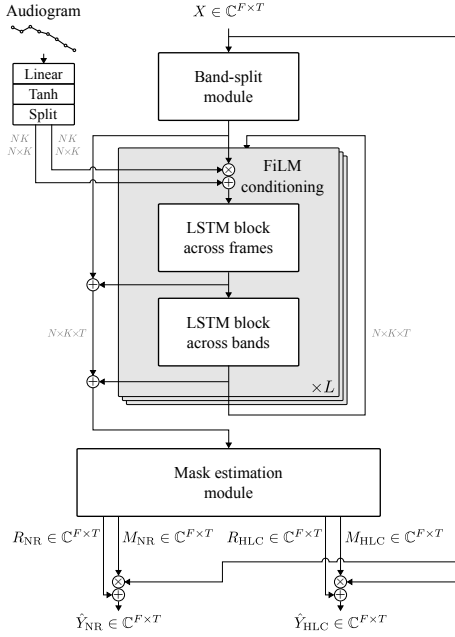
Figure 3: *Speech processor based on BSRNN [18]. Features extracted from the audiogram are fed to each layer using FiLM conditioning. The mask estimation module outputs a complex-valued mask and residual spectrogram for the denoised signal, the compensated signal, or both.*

depending on whether the speech processor is trained for NR, HLC, or both. Processing a 1-s-long input requires 4.6 giga multiply-accumulate (GMAC) operations.

4.2. Differentiable auditory model

The differentiable auditory model is based on the computational model of human auditory signal processing and perception (CASP) [24]. CASP can successfully predict the behavior of NH listeners in a wide range of psychoacoustic tasks, such as intensity discrimination, spectral and temporal masking, and modulation detection. It can also simulate hearing loss, and predict the average behavior of HI listeners [25]. The original model includes outer and middle ear filters, a dual-resonance non-linear (DRNL) filterbank [26], an inner hair cell transduction stage, adaptation loops [27], and a modulation filterbank [28].

A simplified and differentiable version of CASP is implemented. The outer ear filter is removed. The gammatone and low-pass filters in the DRNL filterbank are implemented as 32-ms-long FIR filters to reduce training runtimes. We use 31 filters with center frequencies spaced by one unit on the equivalent rectangular bandwidth (ERB)-rate scale between 80 and 7643 Hz. The inner hair cell transduction stage consists of half-wave rectification and low-pass filtering with a cutoff frequency of 1 kHz. The adaptation loops are replaced with an instantaneous log-compression stage achieving the same steady-state gain. The modulation filterbank is removed.

Following previous studies [29, 30, 31], outer and inner hair cell hearing losses HL_{OHC} and HL_{IHC} are defined for each frequency in dB as

$$\text{HL}_{\text{OHC}} = \min\left(\frac{2}{3}\text{HL}_{\text{tot}}, \text{HL}_{\text{OHC}}^{\max}\right), \quad (11)$$

$$\text{HL}_{\text{IHC}} = \text{HL}_{\text{tot}} - \text{HL}_{\text{OHC}}, \quad (12)$$

where HL_{tot} is the total hearing loss as indicated by the audiogram, and $\text{HL}_{\text{OHC}}^{\max}$ is the maximum outer hair cell loss that can be modeled by the DRNL filterbank. Outer and inner hair cell losses are applied as gain functions to the broken stick non-linearity of the DRNL filterbank and to the output of the inner hair cell transduction stage, respectively. Audiogram thresholds are extrapolated to the center frequencies of the DRNL filterbank using linear interpolation.

4.3. Datasets

The speech processors are trained with noisy and reverberant speech generated using simulated room impulse responses (RIRs). The clean speech utterances are selected from DNS5 [32], LibriSpeech [33], MLS [34], VCTK [35], and EARS [36]. The noise segments are selected from DNS5 [32], WHAM! [37], FSD50K [38], and FMA [39]. The total amount of available speech and noise is 1713 and 541 h, respectively. The RIRs are simulated as in [40]. Each scene is simulated by placing one speech source and up to three noise sources in the same room. The room size is randomly selected between 3×3×2.5 and 10×10×4 m³. The reverberation time T_{60} is randomly selected between 0.1 and 0.7 s. The signal-to-noise ratio (SNR) between the reverberant speech and each reverberant noise source is randomly selected between −10 and 20 dB. Early reflections are included in the clean signal y using a reflection boundary of 50 ms [41]. Scenes are generated on-the-fly during training. For testing, 1000 scenes are generated using speech utterances from Clarity [42], noise segments from TUT [43], and RIRs from DNS5 [32]. All the considered datasets are publicly available. The sampling frequency is 16 kHz.

4.4. Training

The speech processors are trained with 2 000 000 4-s-long scenes. We use a batch size of 32 and the Adam optimizer [44] with an initial learning rate of $1e^{-3}$. The learning rate is reduced by a factor of 0.99 every 10 000 scenes. Gradients are clipped with a maximum L_2 norm of 5. Training takes approximately 36 h on a single A100 40 GB graphics processing unit (GPU).

4.5. Audiograms

We consider the 10 standard audiograms from [45] as well as the NH audiogram. For each training scene, an audiogram a is randomly selected, and a random jitter in [-10, 10] dB is applied to each threshold to increase diversity. Thresholds are finally clipped to [0, 105] dB. During testing, each scene is processed for each of the 11 profiles. Audiogram frequencies are fixed to 250, 375, 500, 750, 1000, 1500, 2000, 3000, 4000 and 6000 Hz.

4.6. Configurations

The following speech processor configurations are compared:

- *BSRNN-SDR*: system trained for NR-only using SDR.
- *BSRNN-NR*: system trained for NR-only using Eq. (1).
- *BSRNN-HLC*: system trained for HLC-only using Eq. (2).
- *BSRNN-NR-HLC*: system trained for uncontrollable joint NR and HLC using Eq. (3).
- *BSRNN-C-NR-HLC*: system trained for controllable joint NR and HLC using Eq. (7).

Additionally, the systems using auditory model-based objectives are trained with either the mean squared error (MSE) or the mean absolute error (MAE) as the loss function ℓ.

Table 1: *Objective metrics for the different speech processor configurations.*

	ℓ	α	a	SDR	PESQ	ESTOI	HASPI	HASQI
Noisy	-	-	NH	−0.15	1.31	0.58	0.85	0.33
BSRNN-SDR	-	-	NH	**13.21**	**2.19**	0.85	**0.96**	0.47
BSRNN-NR	MSE	-	NH	11.78	1.80	0.82	0.93	0.42
BSRNN-NR	MAE	-	NH	12.80	1.96	**0.85**	0.95	0.46
BSRNN-HLC	MSE	-	NH	−0.25	1.31	0.58	0.85	0.33
BSRNN-HLC	MAE	-	NH	−0.12	1.31	0.58	0.85	0.33
BSRNN-NR-HLC	MSE	-	NH	9.45	1.67	0.77	0.91	0.38
BSRNN-NR-HLC	MAE	-	NH	7.16	1.16	0.84	0.89	0.37
BSRNN-C-NR-HLC	MSE	0.0	NH	−0.20	1.31	0.58	0.85	0.33
BSRNN-C-NR-HLC	MAE	0.0	NH	−0.17	1.31	0.58	0.86	0.33
BSRNN-C-NR-HLC	MSE	0.8	NH	9.10	1.64	0.68	0.87	0.37
BSRNN-C-NR-HLC	MAE	0.8	NH	9.91	1.82	0.70	0.92	0.44
BSRNN-C-NR-HLC	MSE	1.0	NH	10.88	1.76	0.80	0.93	0.41
BSRNN-C-NR-HLC	MAE	1.0	NH	13.00	2.14	0.85	0.95	**0.48**
Noisy	-	-	HI	−0.15	1.31	0.58	0.39	0.23
BSRNN-SDR	-	-	HI	**13.21**	**2.19**	0.85	0.43	0.25
BSRNN-NR	MSE	-	HI	11.78	1.80	0.82	0.41	0.25
BSRNN-NR	MAE	-	HI	12.80	1.96	**0.85**	0.43	0.26
BSRNN-HLC	MSE	-	HI	−38.38	1.05	0.45	0.64	0.27
BSRNN-HLC	MAE	-	HI	−39.82	1.05	0.47	0.68	0.27
BSRNN-NR-HLC	MSE	-	HI	−34.29	1.07	0.59	0.74	0.32
BSRNN-NR-HLC	MAE	-	HI	−35.17	1.11	0.65	**0.80**	0.33
BSRNN-C-NR-HLC	MSE	0.0	HI	−38.24	1.06	0.47	0.64	0.28
BSRNN-C-NR-HLC	MAE	0.0	HI	−39.94	1.06	0.48	0.67	0.28
BSRNN-C-NR-HLC	MSE	0.8	HI	−24.60	1.08	0.51	0.65	0.36
BSRNN-C-NR-HLC	MAE	0.8	HI	−26.29	1.09	0.51	0.68	**0.36**
BSRNN-C-NR-HLC	MSE	1.0	HI	10.97	1.76	0.80	0.41	0.24
BSRNN-C-NR-HLC	MAE	1.0	HI	13.04	2.14	0.85	0.43	0.26

Figure 4: *HASPI and HASQI as a function of the mixing parameter α, when using HI audiograms and MAE loss.*

5. Results

The different speech processor configurations are evaluated using SDR, perceptual evaluation of speech quality (PESQ) [46], extended short-term objective intelligibility (ESTOI) [47], hearing aid speech perception index (HASPI) [48], and hearing aid speech quality index (HASQI) [49]. SDR, PESQ, and ESTOI reflect NR performance, while HASPI and HASQI reflect joint NR and HLC performance. The results are averaged separately for scenes processed with NH and HI audiograms a, and reported in Tab. 1. Systems trained for joint NR and HLC can be prompted with a NH audiogram to compare them with systems trained for NR-only. When prompted with HI audiograms, systems trained for NR-only achieve poor HASPI and HASQI results, since HLC is required. Conversely, systems trained for HLC achieve poor SDR, PESQ, and ESTOI results, since the provided amplification causes the output to deviate from the clean signal. For the controllable system BSRNN-C-NR-HLC trained for joint NR and HLC, setting the mixing parameter $\alpha = 0$ enables HLC-only, while $\alpha = 1$ enables NR-only. Key observations include:

- When prompted with $\alpha = 1$, the controllable system BSRNN-C-NR-HLC trained with MAE outperforms its counterpart BSRNN-NR trained for NR-only in terms of SDR and PESQ. This is observed despite BSRNN-C-NR-HLC being designed to predict both a denoised and a compensated signal for a wide range of listener profiles using a similar number of trainable parameters. This aligns with the assumption that learning multiple related tasks simultaneously can produce better internal representations and stronger generalization compared to learning each task separately [50].
- Similarly, BSRNN-C-NR-HLC achieves superior HASQI results to the uncontrollable system BSRNN-NR-HLC for HI listeners and $\alpha = 0.8$, and superior SDR and PESQ results for NH listeners and $\alpha = 1$, even though BSRNN-C-NR-HLC predicts both a denoised and a compensated signal using a similar number of trainable parameters.
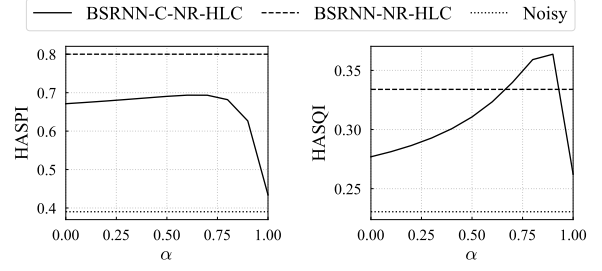
- While the highest SDR and PESQ results are achieved by BSRNN-SDR, systems trained using the proposed differentiable auditory model achieve comparable NR performance. This suggests that the output of the proposed differentiable auditory model is a valid training target for NR.

Figure 4 shows the HASPI and HASQI results as a function of the mixing parameter α for the controllable system BSRNN-C-NR-HLC, using HI audiograms and MAE loss. The results for the uncontrollable system BSRNN-NR-HLC and the noisy input are plotted as horizontal lines. BSRNN-C-NR-HLC achieves optimal HASPI and HASQI results for $\alpha = 0.7$ and $\alpha = 0.9$, respectively. This suggests that a balanced combination of NR and HLC provides optimal speech intelligibility and perceptual quality for hearing aid users. While BSRNN-C-NR-HLC does not outperform BSRNN-NR-HLC in terms of HASPI for any α value, it does so in terms of HASQI for $\alpha \in [0.7, 0.9]$.

6. Conclusion

We proposed a novel framework for training a system capable of controllable joint NR and HLC. This approach eliminates the need for training auditory model emulators, training for individual listener profiles, or training for a fixed balance between NR and HLC. The system demonstrates comparable performance to specialized systems designed for either NR-only or HLC-only, as measured by objective metrics. Additionally, the ability to adjust the balance between NR and HLC enables optimal HASPI and HASQI results, underscoring its relevance for hearing aid users. Future work will evaluate the system with listening tests, investigate the influence of each differentiable auditory model stage on performance, and investigate the benefit of adjusting α dynamically based on short-time acoustic features.

7. References

[1] D. Baby, A. van den Broucke, and S. Verhulst, "A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications," *Nat. Mach. Intell.*, 2021.

[2] F. Drakopoulos, D. Baby, and S. Verhulst, "A convolutional neural-network framework for modelling auditory sensory cells and synapses," *Commun. Biol.*, 2021.

[3] P. Leer, J. Jensen, Z.-H. Tan, J. Østergaard, and L. Bramsløw, "How to train your ears: Auditory-model emulation for large-dynamic-range inputs and mild-to-severe hearing losses," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.

[4] F. Drakopoulos and S. Verhulst, "A differentiable optimisation framework for the design of individualised DNN-based hearing-aid strategies," in *Proc. ICASSP*, 2022.

[5] ——, "A neural-network framework for the design of individ-

ualised hearing-loss compensation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2023.

[6] F. Drakopoulos, A. van den Broucke, and S. Verhulst, "A DNN-based hearing-aid strategy for real-time processing: One size fits all," in *Proc. ICASSP*, 2023.

[7] P. Leer, J. Jensen, L. H. Carney, Z.-H. Tan, J. Østergaard, and L. Bramsløw, "Hearing-loss compensation using deep neural networks: A framework and results from a listening test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2025.

[8] Z. Tu, N. Ma, and J. Barker, "DHASP: Differentiable hearing aid speech processing," in *Proc. ICASSP*, 2021.

[9] ——, "Optimising hearing aid fittings for speech in noise with a differentiable hearing loss model," in *Proc. INTERSPEECH*, 2021.

[10] Z. Tu, J. Zhang, N. Ma, and J. Barker, "A two-stage end-to-end system for speech-in-noise hearing aid processing," in *Proc. Clarity*, 2021.

[11] K. Žmolíková and J. H. Černocký, "BUT system for the first Clarity Enhancement Challenge," in *Proc. Clarity*, 2021.

[12] J. Cheng, R. Liang, L. Zhao, C. Huang, and B. W. Schuller, "Speech denoising and compensation for hearing aids using an FTCRN-based metric GAN," *IEEE Signal Process. Lett.*, 2023.

[13] S. Drgas, L. Bramsløw, A. Politis, G. Naithani, and T. Virtanen, "Dynamic processing neural network architecture for hearing loss compensation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.

[14] T. Neher, "Relating hearing loss and executive functions to hearing aid users' preference for, and speech recognition with, different combinations of binaural noise reduction and microphone directionality," *Front. Neurosci.*, 2014.

[15] T. Neher, K. C. Wagener, and R.-L. Fischer, "Directional processing and noise reduction in hearing aids: Individual and situational influences on preferred setting," *J. Am. Acad. Audiol.*, 2016.

[16] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. CVPR*, 2018.

[17] N. Gößling, D. Marquardt, and S. Doclo, "Performance analysis of the extended binaural MVDR beamformer with partial noise estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2020.

[18] Y. Luo and J. Yu, "Music source separation with band-split RNN," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2023.

[19] J. Yu and Y. Luo, "Efficient monaural speech enhancement with universal sample rate band-split RNN," in *Proc. ICASSP*, 2023.

[20] J. Yu, Y. Luo, H. Chen, R. Gu, and C. Weng, "High fidelity speech enhancement with band-split RNN," in *Proc. INTERSPEECH*, 2023.

[21] W. Zhang, K. Saijo, J.-W. Jung, C. Li, S. Watanabe, and Y. Qian, "Beyond performance plateaus: A comprehensive study on scalability in speech enhancement," in *Proc. INTERSPEECH*, 2024.

[22] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020.

[23] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018.

[24] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *J. Acoust. Soc. Am.*, 2008.

[25] M. L. Jepsen and T. Dau, "Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss," *J. Acoust. Soc. Am.*, 2011.

[26] E. A. Lopez-Poveda and R. Meddis, "A human nonlinear cochlear filterbank," *J. Acoust. Soc. Am.*, 2001.

[27] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, 1996.

[28] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.*, 1997.

[29] E. A. Lopez-Poveda and P. T. Johannesen, "Behavioral estimates of the contribution of inner and outer hair cell dysfunction to individualized audiometric loss," *J. Assoc. Res. Otolaryngol.*, 2012.

[30] J. Zaar and L. H. Carney, "Predicting speech intelligibility in hearing-impaired listeners using a physiologically inspired auditory model," *Hear. Res.*, 2022.

[31] H. Relaño-Iborra, J. Zaar, and T. Dau, "Evaluating an auditory model as predictor of speech understanding in hearing-impaired listeners," in *Proc. Forum Acust.*, 2023.

[32] H. Dubey *et al.*, "ICASSP 2023 Deep Noise Suppression Challenge," *IEEE Open J. Signal Process.*, 2024.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[34] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. INTERSPEECH*, 2020.

[35] C. Veaux, J. Yamagishi, and S. King, "The Voice Bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. O-COCOSDA/CASLRE*, 2013.

[36] J. Richter *et al.*, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Proc. INTERSPEECH*, 2024.

[37] G. Wichern *et al.*, "WHAM!: Extending speech separation to noisy environments," in *Proc. INTERSPEECH*, 2019.

[38] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2022.

[39] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proc. ISMIR*, 2017.

[40] Y. Luo and R. Gu, "Fast random approximation of multi-channel room impulse response," in *Proc. ICASSP*, 2024.

[41] N. Roman and J. Woodruff, "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *J. Acoust. Soc. Am.*, 2013.

[42] S. Graetzer *et al.*, "Dataset of British English speech recordings for psychoacoustics and speech processing research: The Clarity speech corpus," *Data Br.*, 2022.

[43] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. EUSIPCO*, 2016.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[45] N. Bisgaard, M. S. Vlaming, and M. Dahlquist, "Standard audiograms for the IEC 60118-15 measurement procedure," *Trends Amplif.*, 2010.

[46] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)–A new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001.

[47] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2016.

[48] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Commun.*, 2014.

[49] ——, "The hearing-aid speech quality index (HASQI)," *J. Acoust. Soc. Am.*, 2010.

[50] J. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.*, 2000.