Improved sound quality human-inspired DNN-based audio applications

Chuan Wen [1] Guy Torfs [2] Sarah Verhulst[1]
[1] Hearing Technology Lab, Department of Information Technology, Ghent University, Ghent, Belgium
[2] IDLab, Department of information Technology (INTEC), Ghent University–imec, Ghent, Belgium

Abstract

The human auditory system evolved into a structure that provides sharp frequency tuning while transforming sound into a neural code that is optimized for speech understanding in challenging acoustic environments. Employing hallmark features of human hearing in audio applications might thus leverage these systems beyond what is currently possible with purely data-driven approaches. A key requirement for such bio-inspired audio applications is a fully differentiable closed-loop system that includes a biophysically-realistic model of (hearing-impaired) auditory processing. However, existing state-of-the-art models introduce tonal artifacts within their processing that end up as detrimental audible artifacts in the resulting audio application. We propose a solution that improves the architecture of CNN-based auditory processing block to avoid the creation of spurious distortions, while we optimize computations to ensure that the audio applications have real-time capabilities (latency <10ms). We provide a proof-of-principle example for the case of closed-loop, CNN-based hearing-aid algorithms, and conclude that CNN-based auditory models embedded in closed-loop training systems hold great promise for the next generation of bio-inspired audio applications.

Keywords: human-inspired, closed loop, deep neural network, audio processing, artifacts

1 Introduction

The human auditory system, with its intricate and sophisticated mechanisms for sharp frequency selectivity, has long been a source of inspiration for audio processing techniques. For example, the Mel-Frequency Cepstral Coefficients (MFCCs) are a widely-used feature extraction method in Speech Recognition, as they provide a compact, noise-robust representation that mimics human auditory filtering [1]. Biophysically-inspired speech enhancement models furthermore demonstrate stronger robustness at negative signal-to-noise ratios (SNR) than standard models with spectrogram and waveform features [2, 3]. Lastly, [4] proposed a bio-inspired DNN-based closed-loop framework that is capable of restoring all hearing-loss cases, while including realistic audio processing models. The autoencoder-based model (CoNNear) that is used within the closed-loop system comprises a differentiable description of the cochlea, inner-haircell (IHC), and auditory-nerve fiber (ANF) processing stages, derived from a biophysically realistic mechanistic auditory model [5] to train DNN-based hearing aid (HA) algorithms (fig. 1a).

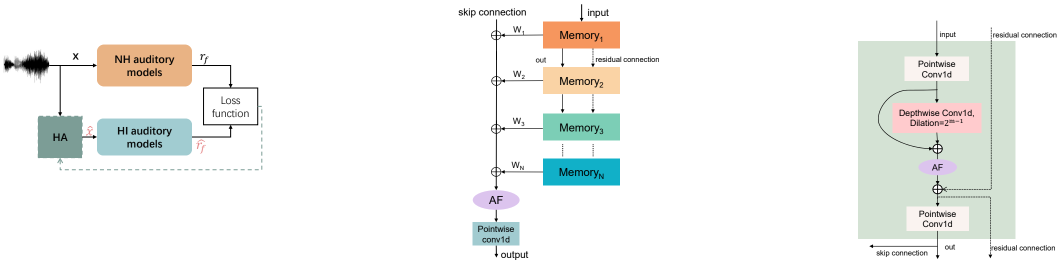
Even though the HA models trained by the framework in [4] are capable of compensating for different sensorineural hearing loss aspects, the network produces audible artifacts that compromise the sound quality of the HA model. In these autoencoder-based auditory models (CoNNear$_{\text{cochlear}}$, CoNNear$_{\text{IHC}}$ and CoNNear$_{\text{ANF}}$), the undesired artifacts originate from the transposed convolutions in the decoder [6]. When the different CoNNear elements are placed in cascade, the distortions will propagate throughout the closed-loop system, and become overamplified

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

and audible. The artifacts associated with CoNNear processing thus prevent the bio-inspired closed-loop system for audio applications.

Several alternative upsampling techniques for transpose convolution have been proposed to eliminate the artifacts. Interpolation + convolution can substitute the transposed convolution [7]. But it still shows obvious artifacts. The subpixel CNN was proposed based on reshape +convolution as an efficient upsampling layer [8, 9] but introduces tonal artifacts due to the periodic shuffle operator [8]. [10] proposed a WaveNet-based model to approximate the IHC receptor potentials that used dilated convolutions to avoided the upsampling operation. However, the WaveNet-based model still produces undesired tonal artifacts that can become audible.

We propose a new CoNNear (dCoNNear) inspired by temporal convolutional networks (TCN) [11] and deep feedforward sequential memory networks (DFSMN) [12]. It comprises a sequence of stacked memory blocks. For each memory block, depthwise dilated 1-D convolutions are employed to model the long-term dependencies of audio processing (e.g. cochlear impulse response and neuronal adaptation) while avoiding the need for upsampling in original CNN-based systems. The dCoNNear architecture is applied to all auditory processing stages including cochlear, IHC, and ANF, as well as a full closed-loop system used to train a hearing-aid signal processing (HA-model) as shown in fig. 1a. We show that dCoNNear cannot only accurately simulate all processing stages of non-DNN-based SOTA biophysical auditory processing, but also does so without introducing spurious and audible artifacts in the resulting audio application. Our simulations and provided sound examples show that the HA-models trained from this new system generate no audible artifacts. The dCoNNear-based human-inspired closed-loop system has low latency (<10ms) capabilities and hence fosters the advancement of sophisticated audio processing algorithms.



(a) (b) (c)

Figure 1: The dCoNNear-based human-inspired closed-loop system (a) Diagram of the human-inspired closed-loop framework exemplified for designing the individualized hearing-aid models. (b) The block diagram of the dCoNNear. (c) The diagram of the memory block.

2 Artifacts analysis

The artifacts associated with the autoencoder-based CoNNear, herein referred to as "prior CoNNear", were examined at each distinct auditory processing stage: CoNNear$_{\text{cochlear}}$, CoNNear$_{\text{IHC}}$ and CoNNear$_{\text{ANF}}$. The reference for identifying and quantifying these artifacts comes from the deterministic auditory computational models [5], which were used as targets during prior CoNNear training. We used pre-trained prior CoNNear models from [13, 14] in the analysis and the evaluating stimuli were not part of the training materials for prior CoNNear.

Figure 2 (a-c) illustrate the frequency responses across different auditory stages given a 1kHz tonal input. Figure 2d shows the output of the HA model trained with the prior CoNNear framework to compensate the HI type "Slope35-7,0,0" in [4]. The prior CoNNear-based models exhibit spectral peaks in the magnitude spectrum of the cochlear, IHC, and ANF responses that are not present in the target model. The origin of these discrepancies can be attributed to kernel overlap during striding in the transposed convolution, leading to non-uniform output space coverage. This unevenness can produce unnatural oscillations or tones [8]. During training to minimize NH and HI AN responses, the HA models therefore incorporated tonal artifacts as shown in fig. 2d. Such artifacts, which are not inherently related to auditory processing, should be avoided in neural network-based auditory models. While there are different methods available to circumvent the problem of distortions created by the upsampling operation, none of the existing methods removes the tonal artifacts while maintaining the quality of cochlear processing.

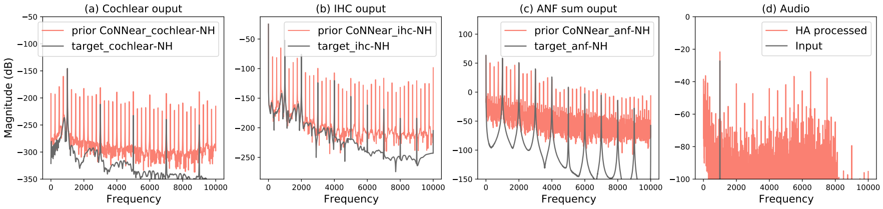


Figure 2: The 1kHz tone response at different auditory processing stages compared against the target model for normal-hearing (**a-c**). **d** The 1 kHz tonal input against the output of HA model trained from prior CoNNear-based framework.

3 Artifact-free dCoNNear-based closed-loop system

The artifact-free dCoNNear-based closed-loop framework for individualized DNN-based HA model training is shown in fig. 1a. It consists of two pathways: One corresponding to the auditory nerve (AN) response r_f of an NH auditory system, and one corresponding to the response $\hat{r_f}$ of a HI system. The HA model was trained by minimizing a pre-defined loss function between the simulated NH and HI responses [4]. All auditory elements and HA models are based on dCONNear topology.

Motivated by the temporal convolutional network (TCN) [11] and deep feedforward sequential memory networks (DFSMN) [12], we propose the dCoNNear as illustrated in fig. 1b. The dCoNNear comprises a sequence of stacked memory blocks with dilation factors exponentially increasing, which enables the network to capture long-term dependencies. M memory blocks with dilation factors 2^{m-1} are repeated R times, yielding a total of N = M * R blocks. Residual connections are employed between the memory blocks to facilitate the learning process by mitigating the vanishing gradient problem. The skip connections are added to aggregate the outputs of each memory block through a weighted summation. The resulting values are subsequently directed to a non-linear activation and convolutional layer to obtain the model outputs.

The architecture of the memory block (Figure 1c) shows how the inputs from the preceding block undergo processing through pointwise convolution, followed by depthwise convolution with dilation factors of 2^{m-1}. The combined outputs from the pointwise and depthwise convolutions are then summed up and passed through an activation function. The final outputs of the memory block, obtained after the pointwise operation, serve as the inputs to the subsequent block and the skip connection paths. The cochlear and IHC models are architecturally congruent with the dCoNNear topology. The branched ANF model is employed to predict three ANF types respectively [4], with the first M blocks serving as shared layers. Each branch, equipped with M blocks, predicts different ANF types independently. The AN summed response r_f is derived by summing up the three ANF responses. The details of each model are shown in table 1.

To illustrate an example of audio processing, it is important to use our individualized HI audio processing model inside the closed-loop system (fig. 1a). For our example, we adjusted the NH auditory modules to simulate different degrees of out-hair-cell (OHC) loss and cochlear synaptopathy (CS). To simulate OHC loss, the NH cochlear model undergoes retraining through transfer learning, based on a specific gain-loss profile. The NH AN response r_f is derived by summing up the three ANF responses which are multiplied by H_r = 13, M_r = 3, L_r=3. To simulate CS, modifications are made by adjusting the values of H_r, M_r, and L_r.

Models	M	R	K	H	Activation	parameters	L	L_l	L_r
dCoNNear$_{\text{cochlear}}$	4	3	64	512	tanh	6.7M	2560	256	256
dCoNNear$_{\text{IHC}}$	4	2	32	128	tanh	0.3M	2560	256	256
dCoNNear$_{\text{ANF}}$	8	2	16	32	ReLU	0.1M	16384	7936	256
HA-model	4	3	32	256	tanh	1.5M	16384	7936	256

Table 1: Hyperparamters of the dCoNNears and HA-model. K denotes the kernel length of the 1-D dilated convolutional layer in each memory block; H indicates the channel number of hidden layers; L, Ll, and Lr represent the input length, left context, and right context, respectively

4 Experiments

To train the auditory (cochlear, IHC and ANF) and HA models, we used 2310 randomly selected samples from the TIMIT speech corpus [15] which provides an adequate representation of the acoustic diversity of speech while being phonetically balanced. The training targets for cochlear, IHC and ANF were the outputs of a biophysical transmission line (TL) model [5], an analytical Hodgkin–Huxley-type IHC model [16], and a three-store diffusion model of the ANF synapse [17], respectively. The input signals to the nonlinear TL model were generated by upsampling the original 16 kHz audio materials to 100 kHz and adjusting their root mean square (RMS) energy to 70 dB SPL. The input signal and TL-model outputs were then re-sampled at 20 kHz to train dCoNNear. The training strategies of [14] were followed to train the dCoNNear and the loss function optimized with the mean absolute error (MAE) between the analytical and predicted dCoNNear outputs.

The trained dCoNNears simulate cochlear responses corresponding to 201 center frequencies between 100Hz and 12 kHz. When training the HA model, NCF = 21 equally spaced frequency channels were selected out of the 201 to speed up the training procedure. The HA-model was optimized with the mean squared error (MSE) between dCoNNear-simulated NH and HI AN responses, while dCoNNears elements were kept frozen. A learning rate of 0.0001 was used with an Adam optimizer [18] and the entire framework was developed in PyTorch.

5 Results

We comprehensively evaluated the dCoNNear elements using stimuli absent from the training material and adopted in auditory neuroscience to quantify key properties of the auditory system. $\mathrm{Q}_{\mathrm{ERB}}$ describes level-dependent cochlear filter characteristics. The half-wave-rectified receptor potential demonstrates the compression feature of IHC mechanical-to-electrical transduction. The ANF synchrony level describes the non-monotonic relation between ANF response and the stimulus level. Figure 3a-c shows a good match between analytical target and dCoNNear predictions, which indicates that the dCoNNears accurately capture the biophysical properties of the human auditory system.

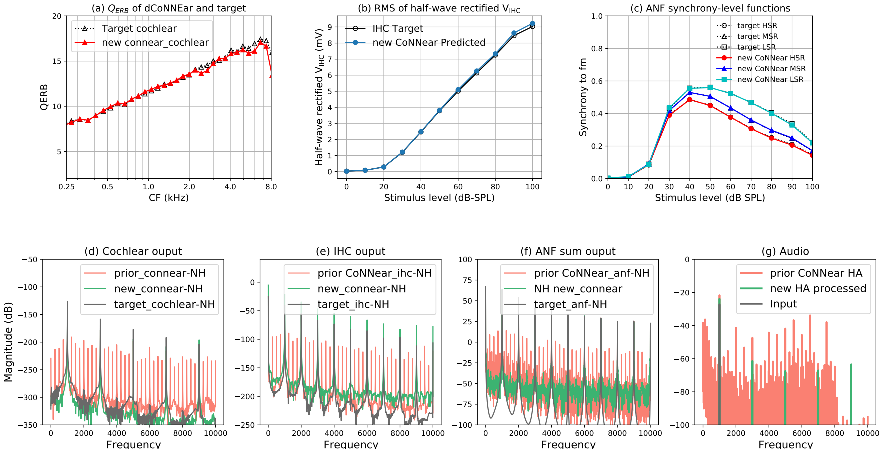


Figure 3: Evaluations of the dCoNNears based on key properties . (a) The $\mathrm{dCoNNear}_{\mathrm{Cochlear}}$ prediction of filter tuning (Q_{ERB}) compared to TL model predictions, for a stimulation level of 40 dB SPL. (b) The simulated $\mathrm{dCoNNear}_{\mathrm{IHC}}$ half-wave rectified receptor potential as a function of sound level. (c) The simulated $\mathrm{dCoNNear}_{\mathrm{ANF}}$ synchrony-level functions for three fiber types. (d-g) The 1kHz tone response at different auditory processing stages and the outputs of HA models based on different CoNNear models.

Subsequently, we assessed whether the artifacts associated with the prior CoNNEar architecture were removed in the dCoNNear architecture. Figure 3 (d-g) illustrates the 1kHz tone response at different auditory processing stages, as well as the HA models trained with diverse frameworks. The artifacts associated with the outputs of the dCoNNear models were markedly reduced compared to prior CoNNears, as observed by red tonal lines in the magnitude spectrum. Furthermore, the output from the HA model, trained employing the dCoNNear-based framework, exhibited a notable reduction in artifacts. These suggest that the dCoNNear-based framework can overcome artifacts generated by the prior CoNNears. The sound quality of the resulting HA algorithm thus markedly improved as can be appreciated from the samples provided at the link.

This work was supported by FWO Machine Hearing 2.0 (216318G) and EIC-Transition EarDiTech (101058278).

References

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, “Deep learning for audio signal processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[2] D. Baby and S. Verhulst, “Biophysically-inspired features improve the generalizability of neural network-based speech enhancement systems,” in *19th Annual Conference of the International-Speech-Communication-Association (INTERSPEECH 2018)*. ISCA, 2018, pp. 3264–3268.

[3] C. Wen and S. Verhulst, “Biophysically-inspired single-channel speech enhancement in the time domain.”

[4] F. Drakopoulos and S. Verhulst, “A neural-network framework for the design of individualised hearing-loss compensation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[5] S. Verhulst, A. Altoe, and V. Vasilkov, “Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss,” *Hearing research*, vol. 360, pp. 55–75, 2018.

[6] J. Pons, S. Pascual, G. Cengarle, and J. Serrà, “Upsampling artifacts in neural audio synthesis,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3005–3009.

[7] A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and checkerboard artifacts,” *Distill*, vol. 1, no. 10, p. e3, 2016.

[8] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, “Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize,” *arXiv preprint arXiv:1707.02937*, 2017.

[9] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[10] A. Nagathil and I. C. Bruce, “Wavenet-based approximation of a cochlear filtering and hair cell transduction model,” *The Journal of the Acoustical Society of America*, vol. 154, no. 1, pp. 191–202, 2023.

[11] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks: A unified approach to action segmentation,” in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 47–54.

[12] S. Zhang, M. Lei, Z. Yan, and L. Dai, “Deep-fsmn for large vocabulary continuous speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5869–5873.

[13] 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,” *Nature machine intelligence*, vol. 3, no. 2, pp. 134–143, 2021.

[14] F. Drakopoulos, D. Baby, and S. Verhulst, “A convolutional neural-network framework for modelling auditory sensory cells and synapses,” *Communications Biology*, vol. 4, no. 1, p. 827, 2021.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1,” *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[16] A. Altoè, V. Pulkki, and S. Verhulst, “Model-based estimation of the frequency tuning of the inner-hair-cell stereocilia from neural tuning curves,” *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4438–4451, 2017.

[17] A. Altoe, V. Pulkki, and S. Verhulst, "The effects of the activation of the inner-hair-cell basolateral k+ channels on auditory nerve responses," *Hearing research*, vol. 364, pp. 68–80, 2018.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.