A DIFFERENTIABLE OPTIMISATION FRAMEWORK FOR THE DESIGN OF INDIVIDUALISED DNN-BASED HEARING-AID STRATEGIES

Fotios Drakopoulos, Sarah Verhulst

Hearing Technology Lab, Department of Information Technology, Ghent University, Ghent, Belgium

ABSTRACT

Current hearing aids mostly provide sound amplification fittings based on individual hearing thresholds or perceived loudness, even though it is known that sensorineural hearing damage is functionally complex, and requires different treatment strategies. To meet this demand, we propose an optimisation framework for the design of individualised hearing-aid signal processing based on simulated (hearing-impaired) auditory-nerve responses. The framework is fully differentiable, thus the backpropagation algorithm can be used to train DNN-based hearing-aid models that optimally process sound to restore hearing in impaired cochleae. The auditory models within the framework can be tuned to the precise hearing-loss profile of a listener to yield trully individualised restoration strategies. Our simulations show that the trained hearing-aid models were able to enhance the auditory-nerve responses of hearing-impaired cochleae, and this provides a promising outlook for embedding our framework within future hearing aids and augmented-hearing applications.

***Index Terms*—** differentiable framework, hearing aid processing, individualisation, deep neural networks

1. INTRODUCTION

Although hearing aids can restore the inaudibility of faint sounds in many cases, they fail to provide a robust treatment for hearing impairment, especially if not configured correctly for the user. Hearing aids mostly apply amplification to compensate for outer-hair-cell (OHC) loss and elevated hearing thresholds [1, 2], leaving the remaining aspects of sensorineural hearing loss (SNHL) such as cochlear synaptopathy (CS; damage to the auditory-nerve (AN) synapses) untreated [3].

To address this challenge, we drew from the recent advances in deep neural networks (DNNs) to develop a differentiable framework in which a DNN-based hearing-aid (HA) model can be fully trained via backpropagation to optimally restore hearing in a hearing-impaired (HI) cochlea. Although differentiable approaches have been proposed before [4, 5], our framework is based on the restoration of simulated AN responses that account for different degrees of OHC loss and CS [6]. The biophysical detail of our HI models enables us to leverage precision diagnostics to precision treatment. We adopted a recent, differentiable description of human auditory processing [7, 4] to train three HA models by matching the HI cochlear responses to the reference response of a normal-hearing (NH) cochlear model. Different from traditional HA signal processing, we aimed to "reverse engineer" impaired cochlear processing using end-to-end DNNs, without posing prior constraints on the applied audio processing (e.g., frequency band analysis, fixed gain/compression function, signal resynthesis). In the next sections, we describe our framework in more detail and present the results for the three evaluated HI conditions.

This work was supported by the European Research Council (ERC) under the Horizon 2020 Research and Innovation Programme (grant agreement No 678120 RobSpear).

2. FRAMEWORK

The proposed framework for individualised DNN-based HA model training is shown in Figure 1 and consists of two pathways: one corresponding to the reference AN response of the NH cochlea r_f, and one corresponding to the response of a HI cochlea $\hat{r_f}$ which reflects the SNHL profile of an individual listener (frequency-dependent degrees of OHC loss and/or CS). In each iteration of this optimisation approach, the input x is given to the two pathways and the difference between the two generated responses is computed using a pre-defined loss function. Since all components in this framework are differentiable, the HA model can be trained via backpropagation to minimise the selected loss function. After training, the HA model is able to process an input auditory stimulus x and produce a processed version $\hat{x}$ that, when given as input to the specific HI cochlea, can generate an AN response $\hat{r_f}$ that optimally matches the reference NH response r_f.

The (input) training dataset we used consisted of 2310 randomly selected recordings from the TIMIT speech corpus [8], upsampled to 20 kHz and calibrated to 70 dB sound pressure level (SPL). The sentences were sliced into windows of 16384 samples for use as inputs to the auditory models [7, 4]. The entire framework was developed using the Tensorflow [9] and Keras [10] machine-learning libraries. For each evaluated HI condition, training was performed for 60 epochs using an Adam optimiser [11].

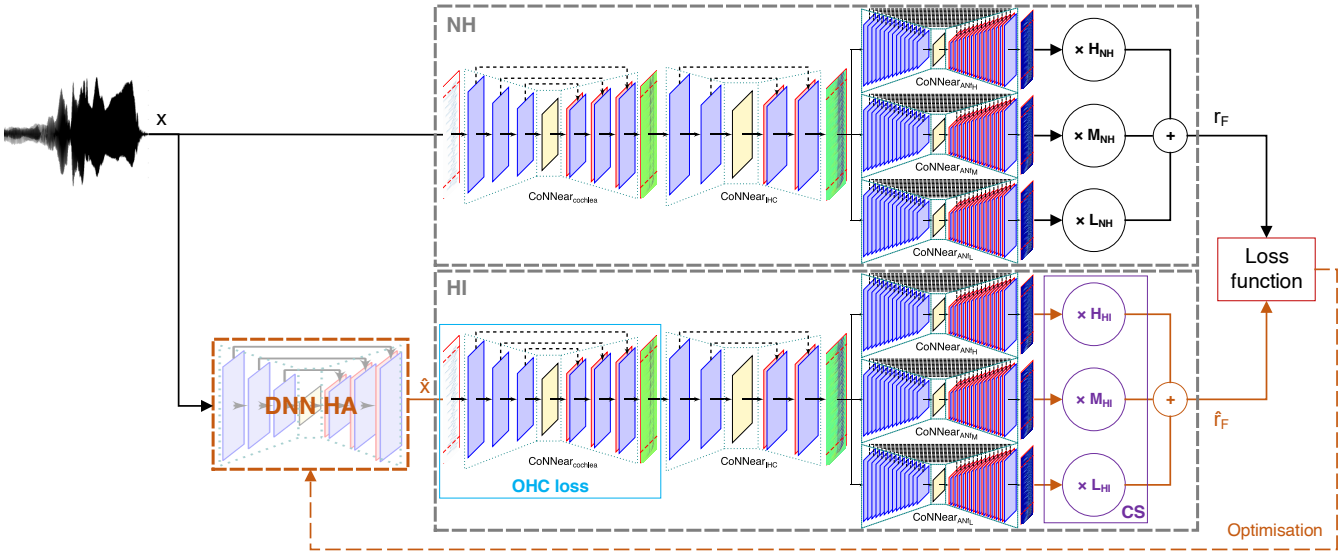

Fig. 1: Diagram of the proposed framework for the design of individualised DNN-based hearing-aid models.

2.1. DNN-based hearing aid model

An end-to-end, encoder-decoder convolutional neural-network (CNN) architecture [12] was used for the HA model and comprised 16 convolutional layers, i.e., 8 in the encoder and 8 in the decoder. The number of filters used in the encoder layers was 16, 32, 32, 64, 64, 128, 128 and 256, respectively, mirrored in reverse order in the decoder. Each convolutional layer had a filter length of 32 and was followed by a PReLU non-linearity except for the last layer. A stride of 2 was used in each layer to halve the temporal (time) dimension of the input, resulting in a shrunk dimension of $L/2^8$ samples. Skip connections were added and the decoder followed the opposite procedure to double the temporal dimension, resulting in a final output size that is identical to that of the input signal.

The HA models were trained to process input sub-frames of $L_p = 2048$ samples (102.4 ms for $f_s = 20$ kHz), but the CNN architectures can process inputs of any size after training.

2.2. Auditory cochlear models

Both NH and HI auditory cochleae were biophysically-inspired CNN-based models that accurately describe human cochlear, inner-hair-cell (IHC) and AN fiber (ANF) processing, named CoNNear$_{\text{cochlea}}$ [7] and CoNNear$_{\text{IHC-ANF}}$ [4]. By default, the CoNNear models simulate responses across 201 simulated tonotopic cochlear locations, corresponding to center frequencies (CFs) between 112 Hz and 12 kHz [13]. Here, $N_{CF} = 21$ frequency channels were selected out of the 201 to simplify and speed up the training procedure. Thus, when using inputs x of $L_c = 16384$ samples, the CoNNear cochleae generated AN responses r_f of $L = 8192$ samples over $N_{CF} = 21$ different frequency channels, after accounting for the context of the time dimension [4]. The AN responses correspond to the firing rates of all ANFs attached to an IHC, simulated for each CF across time. These responses were summed across the simulated CFs to yield the AN$_{\text{summed}}$ responses.

2.2.1. *Hearing impairment*

Although the CoNNear models describe the processing of a NH cochlea, they can easily be adjusted to simulate different degrees of OHC loss and CS [14]. Using transfer learning, OHC loss can be introduced by retraining the NH cochlear model so that it corresponds to a specific gain loss profile or an individual audiogram [15].

To simulate CS, the contribution of the three AN modules (high-, medium- and low-spontaneous-rate ANFs) can be adjusted to reflect a selective loss of AN fibers. In the NH cochlea, the simulated responses of the three ANF types are multiplied by $H_{NH} = 13$, $M_{NH} = 13$ and $L_{NH} = 3$ [16], respectively, to derive the NH AN response r_f. Accordingly, we decreased the values of H_{HI}, M_{HI} and L_{HI} in the HI cochlea to simulate a specific degree of CS.

2.3. Loss function

We considered a combined loss function that included different representations of the difference between the NH response r_f and HI response $\hat{r_f}$, as well as between the unprocessed x and processed input $\hat{x}$. The final loss function L was the concatenation of all individual losses $[20 \cdot L_{ch}, 1 \cdot L_{sum}, 5 \cdot L_{stft}, 1000 \cdot L_{FFTx}]$, each described below:

$$L_{ch} = \text{MAE}\{(r_{f(i,j)})^2, (\hat{r_f}_{(i,j)})^2\} \quad (1)$$

$$L_{sum} = \text{MAE}\{\sum_{j=1}^{N_{CF}} (r_{f(i,j)})^2, \sum_{j=1}^{N_{CF}} (\hat{r_f}_{(i,j)})^2\} \quad (2)$$

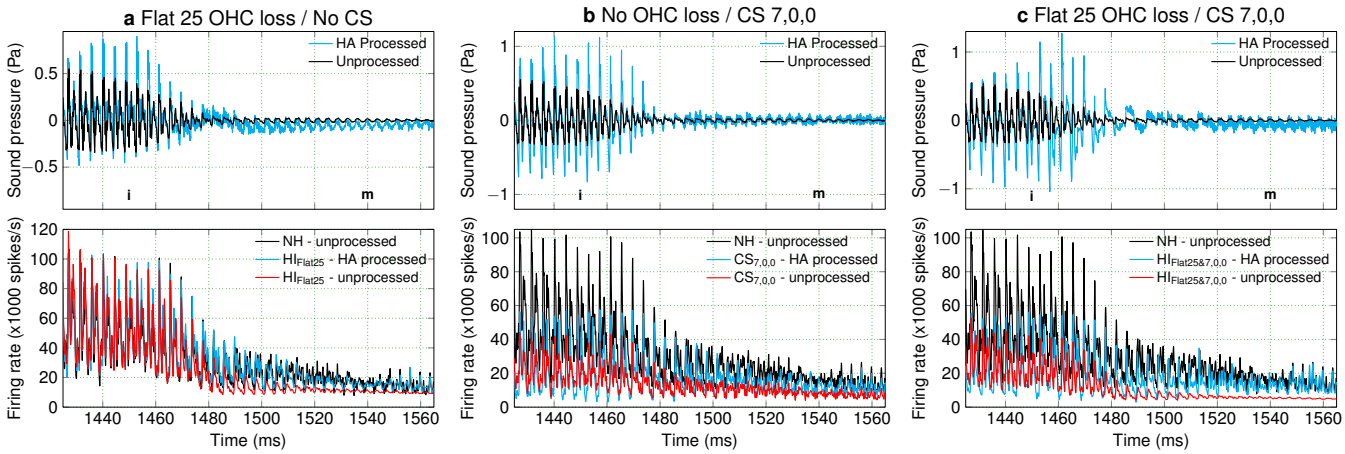

Fig. 2: Simulated $\text{AN}_{\text{summed}}$ responses for the different SNHL combinations and the respective HA models. A speech segment is shown before and after processing (top), and the enhancement that the processed stimulus yielded to the HI response (bottom). An ideal HA model would match the simulated HI response after processing (cyan) to the NH response (black).

$$L_{stft} = \text{MAE}\{|\text{stft}[\sum_{j=1}^{N_{CF}} r_{f(i,j)}]|, |\text{stft}[\sum_{j=1}^{N_{CF}} \hat{r_f}_{(i,j)}]|\} \quad (3)$$

$$L_{FFTx} = \text{MAE}\{X_i, \hat{X}_i\}, \qquad f(i) > 8 \text{ kHz} \quad (4)$$

where MAE denotes the mean absolute error (L1-loss) between two components, stft the short-time Fourier transform (STFT), X and $\hat{X}$ the FFT magnitude spectra of the auditory stimuli x and $\hat{x}$, respectively. The L_{ch} loss was computed directly from the AN responses r_f and $\hat{r_f}$, while the L_{sum} from the $\text{AN}_{\text{summed}}$ responses (after summing the AN responses across the N_{CF} frequency channels). The AN responses were squared to enhance the differences and to exaggerate the peaks of the responses. The loss L_{stft} corresponded to the difference between the STFT magnitudes of the two $\text{AN}_{\text{summed}}$ responses. Lastly, the loss L_{FFTx} was added to ensure that no processing was applied to the signal above 8 kHz (Nyquist frequency of the speech corpus [8]).

3. EVALUATION

Three HA models were trained to compensate for three combinations of SNHL representing different degrees of OHC loss and CS deficits:

- $\text{HI}_{\text{Flat25}}$: 25 dB cochlear gain loss across all frequencies, corresponding to a flat audiogram shape
- $\text{CS}_{7,0,0}$: loss of LSR and MSR, 46% loss of HSR ANFs
- $\text{HI}_{\text{Flat25\&7,0,0}}$: combined Flat25 HL and 7,0,0 CS

The trained HA models were evaluated using speech material from the TIMIT test dataset that were not part of the training data. An example sentence (”should she wake him”) was used to visually examine the effect of the HA processing on a vowel and a consonant, and focusses on the last part of the word ”him” (Fig. 2).

4. RESULTS AND DISCUSSION

Figure 2 shows the effect of the trained HA models on the selected speech segment. For each HI condition (panels a-c), the stimulus is shown before and after processing with the respective HA model on top. Then, the simulated $\text{AN}_{\text{summed}}$ responses to the unprocessed stimulus are shown for the NH and HI cochleae, as well as the HI response using the processed stimulus as input. Ideal restoration would lead to perfectly-matching NH and HI responses (after processing). In the case of OHC loss (Fig. 2a), the trained HA model was able to accurately restore the $\text{AN}_{\text{summed}}$ responses. When CS was introduced (Fig. 2b,c), the trained HA models enhanced the HI response but were not able to fully restore it back to the reference response everywhere (e.g., 1425-1480 ms).

When inspecting the type of signal processing that the trained HA models applied, we can see that the *Flat25* HA model relied on optimally-selected gain (Fig. 2a; top) to compensate for the response difference across time (Fig. 2a; bottom). On the other hand, the other two models (Fig. 2b,c) predominantly amplified the peaks of the vowel \i\, yielding a compensated HI response with increased dynamic range (higher amplitude peaks but even lower dips). This is better demonstrated in Fig. 3, where the AN responses are plotted for the 21 frequency channels (before summing), either before or after processing with the $\text{HI}_{\text{Flat25\&7,0,0}}$ HA model. Overall, the processing did not enhance the most-excited (on-CF) regions (Fig. 3d) but instead added energy to the off-CF frequency channels that were not initially excited (white dashed boxes). In the case of the vowel (\i\), energy was added to the unexcited channels below the fundamental frequency (CFs < 300 Hz), while energy was removed in the regions between modulation peaks (1425-1460 ms for CFs 10-20) to enhance the resting periods between consecutive peaks and yield an $\text{AN}_{\text{summed}}$ response with higher peak amplitudes (Fig. 2c). In

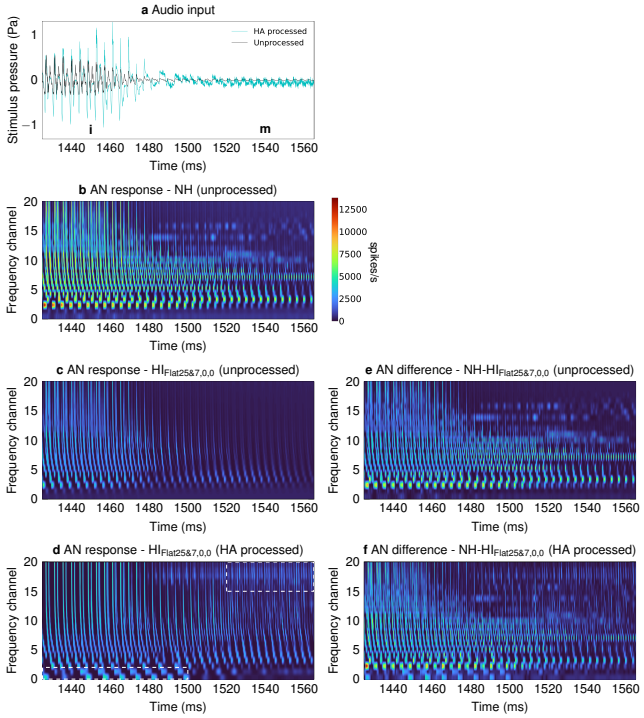

Fig. 3: Simulated AN responses across frequency for the NH cochlea (**b**) and the Flat25&7,0,0 HI cochlea, before (**c**) and after (**d**) applying the respective HA processing to the input. Panels **e** and **f** show the absolute differences between the NH response and the HI responses to the unprocessed and HA processed stimuli, respectively.

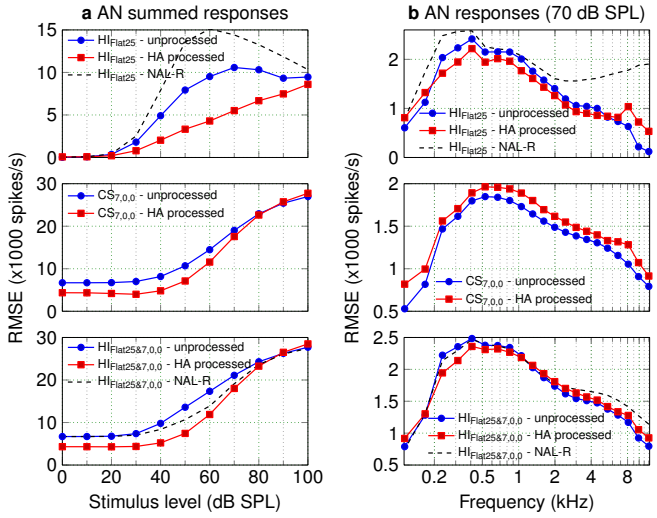

Fig. 4: **a** Average RMSEs for each HI condition as a function of input level, each time computed between the HI AN$_{summed}$ responses (before and after processing with the HA models) and the reference (NH) AN$_{summed}$ response. **b** Average RMSEs across the 21 frequency channels, computed between the HI AN responses (before and after processing) and the NH AN response at each CF in response to 70-dB-SPL stimuli.

the case of the consonant (\m\), the processing added energy in all high-frequency channels and especially to CFs > 5 kHz to enhance the degraded AN$_{summed}$ response.

To quantify the benefit of each HA model on HI sound processing, the test dataset was used to estimate the average root-mean-square error (RMSE) between the simulated NH and HI AN responses, before and after processing with each HA model. Figure 4a shows the AN$_{summed}$ RMSEs for different input levels, while Fig. 4b shows RMSEs of the AN responses at the 21 considered cochlear CFs, computed over the 70-dB-SPL dataset. Overall, the HA models decreased the RMSE across level and frequency, showing more benefit than the standard NAL-R strategy [1] for HI$_{Flat25}$ and HI$_{Flat25\&7,0,0}$.

5. CONCLUSION

We proposed a differentiable optimisation framework that can be used to design individualised audio-based treatments for HI cochleae. The framework was evaluated for three HI cases by training DNN-based HA models to restore the simulated HI AN responses to the reference NH responses. The trained HA models enhanced the AN responses in all three HI combinations. In the case of OHC loss alone, the HA processing was able to precisely restore the AN$_{summed}$ responses by decreasing the RMSE by a factor of 1.6 on average. When CS was introduced in the HI cochlea, the corresponding HA models were able to enhance the AN$_{summed}$ responses by increasing their dynamic range, but did not fully restore the HI responses to the NH response. The compensation of such severe hearing deficits (loss of more than half of the ANFs) resulted in HA models that applied more elaborate sound processing (e.g., sharpening of peaks, off-CF amplification) compared to traditional HA signal processing (e.g., dynamic compression or amplification of the signal).

Our baseline approach was based on biophysically accurate auditory (SNHL) models and can generate a novel type of end-to-end HA processing without applying any prior constrains to the signal processing. The components of our framework can be tuned to correspond to the individual HI cochlea of a listener [17] and yield trully individualised hearing restoration, benefitting people with OHC damage and/or CS. After training, the individualised HA model can be used alone (e.g., in a hearing device) to process sound such that hearing is optimally restored in the specific listener.

In future work, the loss function of our framework can be further adjusted to yield more specific hearing restoration, e.g., by including objective metrics such as PESQ or STOI to focus on speech quality or intelligibility improvement. After a successful evaluation of the speech restoration capabilities of the trained HA models, the architectures can be optimised for real-time applications [18] and integration within hearing instruments and cochlear implants.

6. REFERENCES

[1] Denis Byrne and Harvey Dillon, "The national acoustic laboratories'(nal) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and hearing*, vol. 7, no. 4, pp. 257–265, 1986.

[2] Gitte Keidser, Harvey Dillon, Matthew Flax, Teresa Ching, and Scott Brewer, "The nal-nl2 prescription procedure," *Audiology research*, vol. 1, no. 1, pp. 88–90, 2011.

[3] Sharon G Kujawa and M Charles Liberman, "Adding insult to injury: cochlear nerve degeneration after "temporary" noise-induced hearing loss," *Journal of Neuroscience*, vol. 29, no. 45, pp. 14077–14085, 2009.

[4] Fotios Drakopoulos, Deepak Baby, and Sarah Verhulst, "A convolutional neural-network framework for modelling auditory sensory cells and synapses," *Communications Biology*, vol. 4, no. 1, pp. 827, Jul 2021.

[5] Zehai Tu, Ning Ma, and Jon Barker, "Dhasp: Differentiable hearing aid speech processing," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 296–300.

[6] Sarah Verhulst, Alessandro Altoe, and Viacheslav Vasilkov, "Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss," *Hearing research*, vol. 360, pp. 55–75, 2018.

[7] Deepak Baby, Arthur Van Den Broucke, and Sarah 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.

[8] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David 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, pp. 27403, 1993.

[9] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[10] François Chollet et al., "Keras: Deep learning library for theano and tensorflow," *URL: https://keras. io/k*, vol. 7, no. 8, pp. T1, 2015.

[11] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[12] Deepak Baby and Sarah Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 106–110.

[13] Donald D Greenwood, "A cochlear frequency-position function for several species—29 years later," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2592–2605, 1990.

[14] Sarineh Keshishzadeh, Markus Garrett, Viacheslav Vasilkov, and Sarah Verhulst, "The derived-band envelope following response and its sensitivity to sensorineural hearing deficits," *Hearing research*, vol. 392, pp. 107979, 2020.

[15] Arthur Van Den Broucke, Deepak Baby, and Sarah Verhulst, "Hearing-impaired bio-inspired cochlear models for real-time auditory applications," in *21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020)*. International Speech Communication Association (ISCA), 2020, pp. 2842–2846.

[16] M Charles Liberman, "Auditory-nerve response from cats raised in a low-noise chamber," *The Journal of the Acoustical Society of America*, vol. 63, no. 2, pp. 442–455, 1978.

[17] Sarineh Keshishzadeh, Markus Garrett, and Sarah Verhulst, "Towards personalized auditory models: Predicting individual sensorineural hearing-loss profiles from recorded human auditory physiology," *Trends in Hearing*, vol. 25, pp. 2331216520988406, 2021.

[18] Fotios Drakopoulos, Deepak Baby, and Sarah Verhulst, "Real-time audio processing on a Raspberry Pi using deep neural networks," in *23rd International Congress on Acoustics (ICA 2019)*. Deutsche Gesellschaft für Akustik, 2019, pp. 2827–2834.