A DNN-BASED HEARING-AID STRATEGY FOR REAL-TIME PROCESSING: ONE SIZE FITS ALL

Fotios Drakopoulos, Arthur Van Den Broucke, Sarah Verhulst

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

ABSTRACT

Although hearing aids (HAs) can compensate for elevated hearing thresholds using sound amplification, they often fail to restore auditory perception in adverse listening conditions. To achieve robust treatment outcomes for diverse HA users, we use a differentiable framework that can compensate for impaired auditory processing based on a biophysically realistic and personalisable auditory model. Here, we present a deep-neural-network (DNN) HA processing strategy that can provide individualised sound processing for the audiogram of a listener using a single model architecture. The DNN architecture was trained to compensate for different audiogram inputs and was able to enhance simulated responses and intelligibility even for audiograms that were not part of training. Our multi-purpose HA model can be used for different individuals and can process audio inputs of 3.2 ms in $<$0.5 ms, thus paving the way for precise DNN-based treatments of hearing loss that can be embedded in hearing devices.

***Index Terms*—** differentiable framework, hearing aids, real-time processing, deep neural networks, audiogram

1. INTRODUCTION

Although hearing aids (HAs) can restore the audibility of faint sounds in many cases, they fall short of improving speech understanding in everyday listening conditions such as in noisy environments [1]. To achieve precise treatment outcomes and restore auditory perception of HA users to normal levels, advanced computational methods are necessary that can leverage the highly non-linear character of the human auditory system. To this end, we recently presented a differentiable framework that can compensate for impaired auditory processing [2, 3]. Our framework is based on CoNNear [4, 5], a deep-neural-network (DNN) version of a biophysically realistic model of the human auditory system [6]. Due to its differentiable nature, CoNNear can be used to train a DNN-based HA (DNN-HA) model that optimally processes sound to restore hearing in hearing-impaired (HI) auditory systems. Although similar approaches have been proposed in other studies to improve HA processing [7, 8, 9], the biophysical detail of our differentiable HI models enables us to develop novel DNN-based HA-processing strategies that provide precise treatment of sensorineural hearing loss.

DNN-based HA strategies have the potential to transform hearing healthcare [10], but they can suffer from poor generalisability and high computational costs. The latter aspect can be a significant caveat for embedding DNNs in low-resource devices that require latencies below 10 ms [11], such as HAs or hearables. To address these shortcomings, we present a DNN-HA model that uses an audiogram input to provide individualised HA processing on the fly, i.e. without having to retrain the model for an individual user. Our DNN-HA model uniquely comprises a single DNN architecture that can process sound in real-time to accurately compensate for elevated hearing thresholds of different individuals, such that it can be implemented in future designated chips or hearing devices.

2. METHOD

2.1. Framework

The closed-loop framework of Fig. 1 was used to design the DNN-based HA model and consists of two pathways: One corresponding to the response r of a normal-hearing (NH) auditory system, and one corresponding to the (differentiable) response $\hat{r}$ of a HI auditory system. The optimisation approach of [2, 3] was followed to train the DNN-HA model via backpropagation by minimising a pre-defined loss function between the simulated NH and HI responses. Both NH and HI CoNNear auditory models are biophysically inspired convolutional-neural-network (CNN) based models that accurately describe human cochlear, inner-hair-cell and auditory-nerve (AN) processing across distinct tonotopic cochlear locations [4, 5]. Thus, the derived AN responses r and $\hat{r}$ correspond to biophysically realistic time-frequency representations of sound (neurograms) simulated across 201 cochlear channels with centre frequencies (CFs) between 112 Hz and 12 kHz. These neurogram responses can be summed across the simulated CFs to yield the AN population responses [6].

Here, we develop a single, generalisable DNN-HA model architecture that can achieve optimal compensation for the elevated hearing thresholds of different individuals. To this end, an audiogram input y was included in the framework (Fig. 1) and was used to define the cochlear parameters in the

This work was supported by European Research Council ERC-StG-678120 (RobSpear) and FWO grant G063821N Machine Hearing 2.0.

The trained DNN-HA model, together with a usage example, is available via `https://github.com/fotisdr/DNN-HA`.

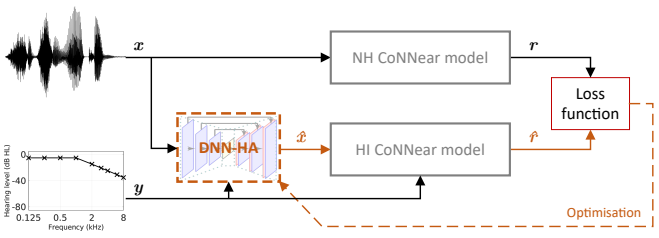

Fig. 1: Closed-loop diagram for the design of a multi-purpose DNN-HA model with adjustable sound processing.

HI CoNNear model in each training iteration so that it accurately simulates the outer-hair-cell damage pattern that gave rise to the individual audiogram. The input y was then used by the DNN-HA model to process the auditory stimulus x based on the given audiogram, generating a processed signal $\hat{x}$ that maximally matches the individual HI response to the reference NH response. Given a successful training using a broad audiogram dataset, the resulting DNN-HA model will be able to generalise to new audiograms to provide optimal HA processing for any user without retraining the model.

2.2. Training parameters

We used an audio training dataset of 2310 randomly selected recordings from the TIMIT speech corpus [12], calibrated to 70 dB sound pressure level (SPL) and upsampled to 20 kHz to match the CoNNear sampling frequency. The sentences were sliced into windows of 81920 samples, out of which 8192 samples provided context for the time dimension of the CoNNear models [4, 5]. N_{CF} = 21 frequency channels were selected out of the 201 with equal spacing to speed up the training procedure. Based on our previous findings [2, 3], we used a loss function that achieved the best benefits for speech in quiet and in noise. The loss function was defined as the mean-absolute error between the squared AN responses, squared AN population responses and power spectrograms of the AN population responses ($\mathcal{L}_{r^2,r_p^2,R_p^2}$ in [3]). The DNN-HA model was trained to minimise the loss function by processing the acoustic waveform for frequencies up to 8 kHz.

The audiogram training dataset was chosen to include audiometric thresholds for 21 hearing-loss (HL) profiles at 8 frequencies, i.e. [0.125, 0.25, 0.5, 1, 2, 3, 4, 6, 8] kHz. Before training, the NH CoNNear cochlear module was retrained for each HL profile using transfer learning, so that it simulated HI cochlear processing for each individual audiogram [13]. This allowed us to load the corresponding weights as non-trainable (frozen) parameters in each iteration of the training procedure based on the input audiogram. The selected HI profiles were:

- 7 Flat*XX* profiles; "Flat" HL of *XX* dB across frequency,
- 7 Slope*XX*_00 profiles; High-frequency "sloping" audiograms starting from 1 kHz with *XX* dB HL at 8 kHz,
- 7 Slope*XX*_05 profiles; "Flat" HL of 5 dB until 1 kHz and "sloping" HL from 1 kHz with *XX* dB HL at 8 kHz,

where *XX* = 5-35 dB HL with a step size of 5 dB.

2.3. DNN-based hearing-aid model

For the DNN-HA model, we used an end-to-end, encoder-decoder CNN architecture that was adopted in [14, 2, 3] and can be executed in real-time on low-resource platforms [11]. Here, the architecture had 5,088,385 trainable parameters and comprised 12 strided convolutional layers with a filter length of 32, i.e. 6 layers in the encoder and 6 in the decoder. The number of filters used in the encoder layers was [32, 32, 64, 64, 128, 128, 256], mirrored in reverse order in the decoder.

To ensure that the model can effectively process audio inputs of any size after training, the sentences were processed in randomly selected frames from 64 to 8192 samples (and also the whole audio signal) during training. The "conditional" HA processing based on the audiogram input was achieved by concatenating the audiogram with the output of the last encoder layer across the filter dimension. Based on the input size and the resulting latent size after the last encoder layer (bottleneck dimension), the audiogram input was repeated across the time dimension so that it could match the temporal size in the bottleneck. Training was performed for 40 epochs using an Adam optimiser [15] on Tensorflow [16] and Keras [17], with a learning rate of 0.0002 for the first 20 epochs and a learning rate of 0.0001 for the next 20.

3. EVALUATION

To evaluate the trained DNN-HA model on HI sound processing, the Flemish Matrix corpus [18] was used as our test dataset. Although the DNN-HA model was trained on a clean speech dataset, the generated speech-shaped noise of the test dataset was used to evaluate the capabilities of the DNN-HA model for speech in noise (SNR = 0 dB). The achieved HA processing was evaluated on the basis of: (a) Restoration of simulated AN responses using root-mean-square errors (RMSEs) and (b) Objective improvement of speech intelligibility using HASPI [19]. The evaluation was performed for conversational speech levels using sentences with root-mean-square (RMS) energy levels between 30 and 70 dB SPL in 10 dB steps. The NAL-R amplification strategy [20] was included as a baseline HA strategy using the implementation provided by the authors of HASPI. Four test audiograms were selected (Fig. 3): Two that were part of the training dataset ("seen") and two that were not included in training ("unseen"). The first two corresponded to the most severe cases of HL that the model was trained to compensate for (Flat35, Slope35_05), while the latter comprised two high-frequency sloping audiograms (Slope35_30, SlopeN1).

3.1. Simulated restoration

The simulated restoration was quantified using the normalised RMSE (NRMSE), computed between the simulated NH and HI AN population responses across the 260 sentences of the Flemish Matrix and normalised to the maximum of the NH

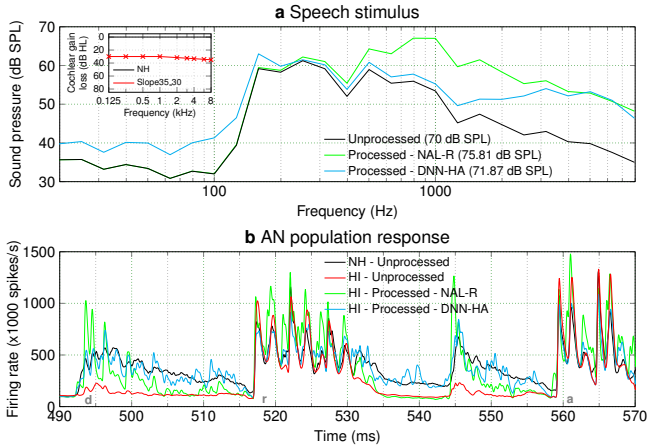

Fig. 2: Simulated restoration for a Flemish Matrix sentence, processed by the NAL-R strategy [20] and the trained DNN-HA model to compensate for an "unseen" HI audiogram (inset in **a**). Panel **a** shows the magnitude of the speech stimulus across frequency before and after processing. Panel **b** shows the simulated AN population responses of the NH and the respective HI CoNNear model for a segment of the sentence.

response for each sentence:

$$\text{NRMSE} = \frac{1}{\max\left(r_p\right)}\sqrt{\frac{\sum_{n=1}^{L}(r_p(n)-\hat{r}_p(n))^2}{L}}, \qquad (1)$$

$$r_p(n) = \sum_{w=1}^{N_{CF}} r(n,w), \qquad \hat{r}_p(n) = \sum_{j=1}^{N_{CF}} \hat{r}(n,w), \qquad (2)$$

where n corresponds to each time sample of the AN population responses r_p and $\hat{r}_p$, and L to the total number of samples. The NH and HI AN population responses r_p and $\hat{r}_p$ were computed by summing the simulated AN responses r and $\hat{r}$ across the frequency dimension w (Eq. 2). Even though we used $N_{\text{CF}} = 21$ frequency channels during training, the full 201-channel CoNNear models were used for the evaluation.

3.2. Intelligibility assessment

Speech intelligibility was assessed using the HASPI model v2 (implementation provided by the authors). For each tested audiogram, the average HASPI scores were computed across the 260 Flemish Matrix sentences before and after processing.

3.3. Real-time processing

The DNN-HA model was trained with different input frames so that it can effectively process input sizes as low as 64 samples (3.2 ms) after training. Although the simulated restoration and intelligibility assessment was performed for the HA processing applied to the whole sentences, we also evaluated the real-time capabilities of the trained model by processing the sentences in frames. Here, input sizes from 64 to 512 samples were tested to define the latency that yielded the best trade-off between execution time and HA performance.

Table 1: The trained DNN-HA model was evaluated for real-time processing using input windows from 64 to 512 samples (3.2-25.6 ms). For each input size, the average time needed to process a Flemish Matrix sentence (average length of 2.7 s) was computed on a CPU (AMD EPYC 7413) and a GPU (NVIDIA A30). The NRMSE percentages and HASPI scores correspond to the results of the Slope35_30 audiogram (Fig. 3**c**) for speech in quiet presented at 50 dB SPL.

	CPU (ms)	GPU (ms)	NRMSE (%)	HASPI
Unprocessed	-	-	11.5160	0.7020
25.6 ms latency (512 samples)				
NAL-R	0.0057	-	9.4334	**0.9952**
DNN-HA	1.9630	0.5667	5.0840	0.9881
DNN-HA (50%)	1.6696	0.3455	**4.8101**	0.9753
12.8 ms latency (256 samples)				
NAL-R	0.0029	-	9.4846	0.9737
DNN-HA	1.1586	0.3327	5.2216	0.9472
DNN-HA (50%)	1.0012	0.2223	4.9089	0.9895
6.4 ms latency (128 samples)				
NAL-R	0.0019	-	9.5645	0.9871
DNN-HA	0.7805	0.2187	5.7451	0.9756
DNN-HA (50%)	0.6992	0.1732	5.3778	0.9911
3.2 ms latency (64 samples)				
NAL-R	**0.0014**	-	9.7153	0.9742
DNN-HA	0.5017	0.1826	6.8077	0.9662
DNN-HA (50%)	0.4612	**0.1578**	6.3940	0.9922

4. RESULTS AND DISCUSSION

Figure 2 shows the simulated restoration of the trained DNN-HA model for an example Flemish Matrix sentence ('David draagt drie gele boten'). Even though the tested audiogram (Slope35_30) was not seen during training, the DNN-HA model achieved a much more precise restoration of the simulated AN population responses than the NAL-R strategy (Fig. 2**b**) which mostly enhanced the onset peaks. The amplification that the DNN-HA provided was less than 2 dB, while the NAL-R strategy applied almost 6 dB gain to compensate for this HL profile. This suggests that our strategy might be able to compensate for more severe HL profiles without relying on excess amplification of the audio signal.

The overall NRMSE and HASPI results for the four evaluated audiograms are shown in Fig. 3. The NRMSE showed similar trends for speech in quiet and in noise, with the DNN-HA model providing an average improvement of 4.8% for noisy speech and outperforming the NAL-R strategy ($\sim$1.2% improvement). For low stimulus levels, both strategies restored the HASPI scores back to $\sim$1 for speech in quiet and in noise. However, when compared to our trained DNN-HA model, NAL-R achieved higher scores for speech in noise at high stimulus levels. The DNN-HA model restored the simulated responses for clean and noisy speech equally well, but it was not able to improve the HASPI scores for 60- and 70-

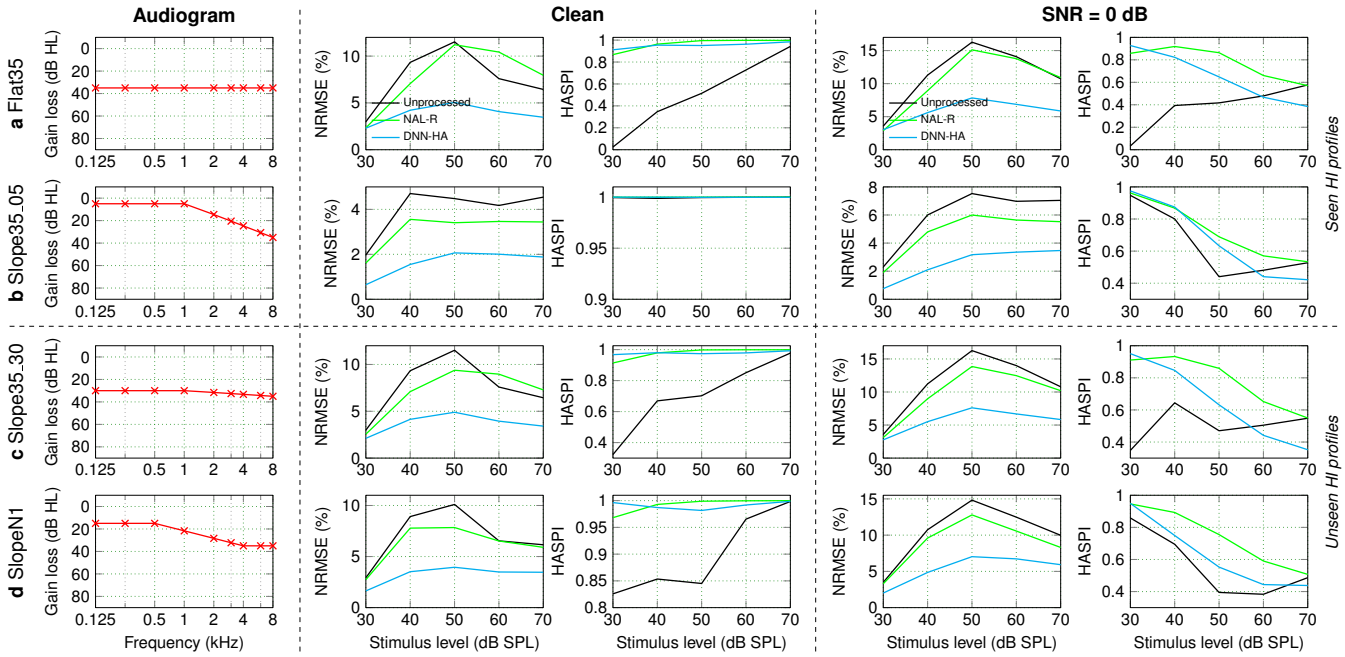

Fig. 3: Simulated restoration for two "seen" (**a**,**b**) and two "unseen" (**c**,**d**) HI profiles. For each corresponding audiogram, the NRMSE and HASPI scores were computed for the 260 Flemish Matrix sentences presented at levels between 30 and 70 dB SPL in quiet and in noise (0 dB SNR). Each plot compares the average NRMSE and HASPI scores before and after processing with the NAL-R amplification strategy and the trained DNN-HA model.

dB-SPL sentences in noise. Although it is possible that the HASPI model cannot generalise to predict performance in noise for the new non-linear processing that our DNN-HA model provides, the performance of the DNN-HA model in noise might still be improved by including noise in the training dataset.

At the same time, similar performance was achieved for seen and unseen audiograms (Fig. 3), which suggests that our trained DNN-HA model can provide optimal HA processing for any audiogram input. In this work, we were limited to audiograms with thresholds up to 35 dB HL for the training and evaluation of the proposed DNN-HA model, since the analytical auditory model allows for the simulation of up to 35 dB HL in the cochlear module [6]. An extension of the analytical model could allow CoNNear to simulate more severe HL profiles in the future which will further improve the application range of the DNN-HA model.

Finally, Table 1 shows the results of the DNN-HA model for low-latency sound processing. For frame sizes from 3.2 to 25.6 ms, the DNN-HA model required 0.5-2 ms on a CPU and 0.2-0.6 ms on a GPU to process a Flemish Matrix sentence. The DNN-HA model was not able to match the execution time of the NAL-R strategy (implemented as a simple 140-order filter), but yielded much better NRMSE percentages and comparable HASPI scores. Performance did not drop significantly when processing the input signal in frames (4.9% NRMSE and 0.97 HASPI when processing the whole stimulus), and only decreased for the NRMSE for window sizes lower than 128 samples. Furthermore, the introduction of 50% overlap between the input frames (DNN-HA (50%)) improved performance, and even surpassed the processing of the whole signal in some cases (NRMSE of $\sim$4.8% for 512 samples). It should be noted that the minimum input the DNN-HA architecture can process is constrained by the number of strided layers in the encoder, i.e. $2^6 = 64$ samples for 6 encoder layers. Smaller input sizes can be achieved with zero-padding or with fewer encoder layers in the DNN-HA architecture.

5. CONCLUSION

In this work, we presented a single-architecture DNN-HA model that can provide adjustable HA processing based on the individual audiogram of a user without retraining. Different from traditional HA signal processing, the end-to-end DNN model was trained to compensate for impaired auditory processing without relying on prior assumptions of the applied audio processing. The proposed DNN-HA model achieved better simulated restoration and comparable intelligibility estimations than the NAL-R strategy by applying less amplification to the audio signal ($<$2 dB at 70 dB SPL for $\sim$35 dB HL). At the same time, our model was able to process inputs as short as 3.2 ms in $<$0.5 ms on a CPU and in $<$0.2 ms on a GPU without showing significant drops in restoration performance. Thus, the proposed DNN-HA model can pave the way for precise DNN-based HA strategies that can easily be embedded in future low-resource hearing devices.

6. REFERENCES

[1] Abby McCormack and Heather Fortnum, "Why do people fitted with hearing aids not wear them?," *International Journal of Audiology*, vol. 52, no. 5, pp. 360–368, 2013.

[2] Fotios Drakopoulos and Sarah Verhulst, "A differentiable optimisation framework for the design of individualised DNN-based hearing-aid strategies," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 351–355.

[3] Fotios Drakopoulos and Sarah Verhulst, "A neural-network framework for the design of individualised hearing-loss compensation," *arXiv preprint arXiv:2207.07091*, 2022.

[4] 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.

[5] 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.

[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] Jeff Bondy, Sue Becker, Ian Bruce, Laurel Trainor, and Simon Haykin, "A novel signal-processing strategy for hearing-aid design: Neurocompensation," *Signal Processing*, vol. 84, no. 7, pp. 1239–1253, 2004.

[8] 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.

[9] Anil Nagathil, Florian Göbel, Alexandru Nelus, and Ian C Bruce, "Computationally efficient DNN-based approximation of an auditory model for applications in speech processing," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 301–305.

[10] Nicholas A Lesica, Nishchay Mehta, Joseph G Manjaly, Li Deng, Blake S Wilson, and Fan-Gang Zeng, "Harnessing the power of artificial intelligence to transform hearing healthcare and research," *Nature Machine Intelligence*, vol. 3, no. 10, pp. 840–849, 2021.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] François Chollet et al., "Keras: The python deep learning library," *Astrophysics Source Code Library*, pp. ascl–1806, 2018.

[18] Heleen Luts, Sofie Jansen, Wouter Dreschler, and Jan Wouters, "Development and normative data for the Flemish/Dutch Matrix test," 2014.

[19] James M Kates and Kathryn H Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[20] 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.