HEARING HEARING LOSS: INVERTING THE AUDITORY NERVE MODEL

Anes Macić
Department of Mechanical Engineering, University of Rochester
amacic@ur.rochester.edu

ABSTRACT

This paper details the development of an inverse auditory nerve model (iAN) using deep learning, specifically aimed at inverting neurograms to better understand the auditory experience of hearing loss. The approach utilizes a convolutional neural network to process neurograms generated from speech data, with a focus on the challenges posed by the complex nonlinearities in sound coding and auditory nerve responses. The model primarily decodes at a single loudness level and grapples with the intricacies of reconstructing both low and high-frequency components. The study acknowledges the limitations in its current methodology, particularly in fully replicating the auditory experience of hearing loss, and suggests potential areas for future refinement and research.

1. INTRODUCTION

The process of auditory mechanotransduction within the inner ear is a complex one, where the conversion of ear canal pressure into impulses along the auditory nerve is intricately governed by a series of nonlinear interactions. These interactions predominantly originate in the cochlear hair cells, which are instrumental in conducting and refining spectral analysis. The loss of these hair cells leads to sensorineural hearing loss, a condition marked by a significant reduction in the perceived loudness of incoming auditory stimuli. This form of hearing loss is inherently a non-linear system due to the pivotal role of hair cells in mechanotransduction.

Conventional approaches to mitigate hearing loss, such as amplifying auditory input to aid in hearing, have yielded limited success. Although such methods can induce a measurable increase in neural activity, they fail to substantially enhance intelligibility [2]. This inefficacy is primarily because many hearing aids are designed as linear, non-adaptive filters, which are inadequate in addressing the complex nonlinearities associated with hearing loss.

Understanding the auditory experience of individuals with hearing loss is a challenging endeavor. Unlike opticians, who can simulate visual impairment by defocusing images, audiologists do not have a direct analog for auditory impairment. This complexity is further compounded by the fact that many auditory periphery models either do not account for, or are not designed to simulate, the effects of hearing loss. This limitation obstructs our understanding of the auditory world experienced by those with impaired hearing. However, an exception exists in the model developed by Zilany et al. (2014) [1], which accurately simulates the auditory periphery and incorporates the capacity to parameterize hearing loss via an audiogram.

The primary objective of this research is to acoustically experience what hearing loss sounds like. To achieve this, we utilize deep learning techniques to reverse-engineer and enhance existing models of the auditory periphery. This approach aims to bridge the current research gap by not only audibly illustrating the experience of hearing loss but also by contributing to the development of hearing aids that more accurately mirror the complexities of the inner ear.

2. BACKGROUND

2.1 The Neurogram: An Inner Ear *Spectrogram*

The journey of sound through the human auditory system is a fascinating process of transformation and encoding. When sound waves reach the outer ear, they are first focused by the pinna into the ear canal. This funneling effect leads the sound waves toward the eardrum, where they induce vibrations. These vibrations are then transferred through the ossicles in the middle ear, which act as critical impedance matchers, ensuring efficient transmission of sound energy into the inner ear.

The inner ear houses the cochlea, a snail-shaped sensory organ that plays a pivotal role in sound perception. The cochlea operates on the principle of tonotopy, which means that different parts of the cochlea are sensitive to different frequencies of sound. High-frequency sound waves primarily stimulate the base of the cochlea, while lower frequencies elicit vibrations in the apical regions. This spatial distribution of frequency sensitivity effectively separates the sound into its spectral components, allowing for an intricate process of sound amplification and analysis.

Within the cochlea, the auditory nerve fibers (ANFs) are the critical elements that transduce these mechanical vibrations into electrical signals. Each segment of the cochlea is innervated by a specific set of ANFs, each finely tuned to respond to a unique frequency band. This arrangement ensures that the entire spectrum of audible frequencies is effectively captured and represented.

The neurogram emerges as a crucial concept in understanding this complex auditory processing. It represents the collective response of these ANFs to sound stimulation. By plotting the firing rates of various ANFs over time and across different cochlear regions, we can construct a

comprehensive map of neural impulses. This map, or neurogram, is akin to a spectrogram but for neural activity. It provides a detailed representation of how the brain interprets and processes sound at the most fundamental level of auditory coding. This unique visualization not only illustrates the frequency and intensity of sound as it travels through the auditory pathway but also offers insights into the temporal dynamics of auditory perception.

2.2 Important properties of the neurogram

While there are similarities between a neurogram and a spectrogram, it is crucial to recognize their three fundamental differences:

1. **Tuning Bandwidth** The spectrogram, based on an orthogonal Fourier basis, assigns each frequency bin a bandwidth proportional to its bin size. This design allows for the attainment of a very narrow tuning bandwidth (high tuning factor Q) through high-rate and extended sampling. In contrast, a neurogram exhibits considerably broader bandwidths around the center frequency of each auditory nerve fiber. This difference in bandwidth distribution is evident in (1).

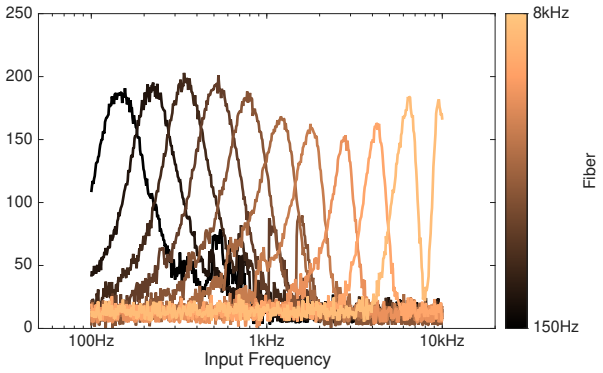

Figure 1. Tuning of auditory nerve fibers (high spontaneous rate) along the cochlear length. Y-axis is the mean firing rate. Input frequency was presented at 65 dB SPL.

2. **Linearity and the Dynamic Range** The spectrogram operates as a linear tool where the loudness is encoded in a linear manner. However, the neurogram functions in a highly nonlinear domain. Auditory nerve fibers differ in their dynamic ranges; high spontaneous rate (HSR) fibers have a narrow dynamic range, with bulk sensitivity in quiet sounds within 20-40 dB SPL, while low spontaneous rate (LSR) fibers have a broader dynamic range, responding to a wider range of 20-100 dB SPL, as depicted in figure (2).

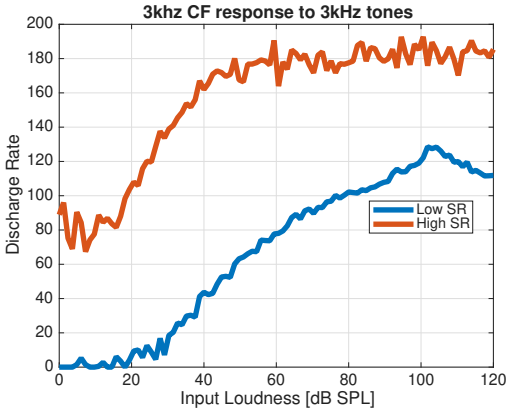

Figure 2. Input-output level curves for two families of fibers: low and high spontaneous rates (LSR, HSR). LSR are wide-dynamic range, HSR are narrow.

3. **Time/Phase Representation** Unlike the spectrogram which computes a complex variable representing both magnitude and timing for each frequency bin, the neurogram is limited to real values. In the neurogram, sound magnitude is depicted through the saturating firing rate of nerve fibers, and phase information is conveyed via amplitude modulations in the firing rate, particularly for low-frequency components. This encoding is consistent with human auditory perception, which can differentiate phase variations in low but not high frequencies. This phenomenon is illustrated in figure (3).

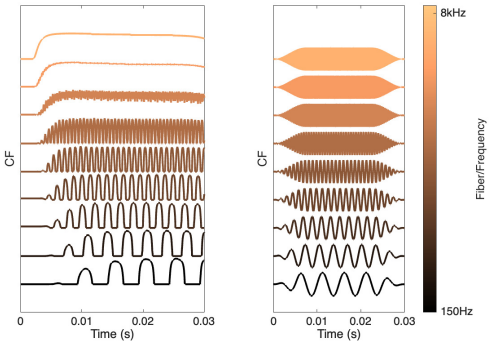

Figure 3. Temporal response of probability of firing (left) of different fibers when stimulated at their characteristic frequency with a sine tone (right).

These three distinctions allow us to critically compare the spectrogram and the neurogram.

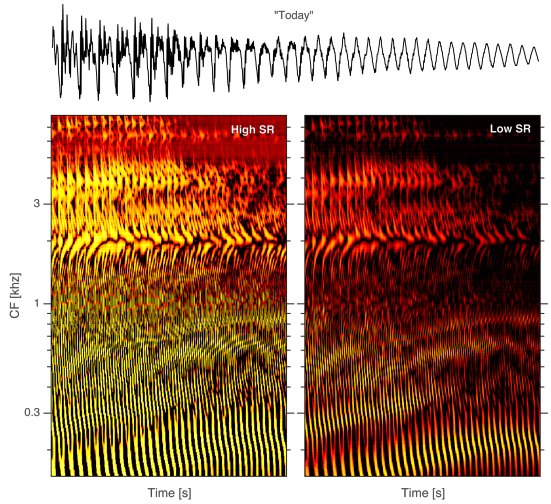

Figure 4. An example of a neurogram.

3. METHODS

The core of our study is the inverse auditory nerve model (iAN), a deep learning-based surrogate model designed to transform a neurogram corresponding to a sound at 65 dB SPL into its time-domain waveform representation.

3.1 Auditory Nerve Simulations

For the purpose of training the iAN, we utilized the model developed by Zilany et al. [1] to generate neurograms. Our dataset comprised 2 hours of clean speech from the LibriSpeech corpus, encompassing a diverse range of utterances from 40 different speakers. The speech data, sampled at 16kHz, provided a rich and varied acoustic landscape for our analysis.

In creating the neurograms, we focused on a dual-channel setup to represent two distinct types of auditory nerve fibers. Each channel consisted of 512 frequency bins, covering a range from 150 Hz to 8 kHz, capturing a broad spectrum of auditory information. To manage the computational load, these neurograms were initially simulated at a high sampling frequency of 100 kHz. However, due to substantial storage demands, we later resampled them at a more manageable rate of 2 kHz, storing each value in a 16-bit format.

This process of simulating neurograms was computationally intensive. Generating neurograms for just a few minutes of speech required approximately one hour of parallel processing in MATLAB, utilizing 90GB of RAM and 20 CPU cores.

A crucial aspect of our simulation process was ensuring that all speech in the dataset was scaled to a standardized loudness level of approximately 65 dB SPL. This level is representative of the typical loudness encountered in comfortable conversational speech, providing a realistic and relevant auditory context for our study.

3.2 Inverse Auditory Nerve Surrogate Model

The inverse auditory nerve model (iAN), central to our study, is built upon the architecture of a convolutional neural network. This network is designed to process neurograms and translate them into time-domain waveforms.

At the heart of iAN's processing capability is its input structure, which accepts neurograms in a four-dimensional format: $(B, C, nF, nT) = (B, 2, 512, 50)$. In this structure, 'B' represents the batch size,C denotes the number of channels – fixed at two to correspond with the two types of auditory nerve fibers, nF indicates the number of frequency bins, set at 512 to cover the range from 150 Hz to 8 kHz, and nT is the window length, comprising 50 timesteps that correspond to a 25 ms duration.

The output from iAN is equally structured to reflect the time-domain representation of the sound. It follows the format $(B, nT) = (B, 400)$, where B remains the batch size, and nT now represents the time dimension of the reconstructed waveform. This output corresponds to a 25 ms segment of the waveform, sampled at a rate of 16 kHz, effectively translating the frequency information from the neurogram into a temporal sound waveform.

To enhance the learning process, the batch size was set at 16, and the training data was shuffled in the time domain during each training step to ensure robustness and generalization of the model. The network processed the neurograms with a hop size of 5 ms, allowing for detailed and accurate waveform reconstruction.

A notable feature of iAN is its final nonlinearity, characterized by the *tanh* function. This design choice confines the output waveform within the codomain of $[-1, 1]$ ensuring that the reconstructed sound waveforms remain within a normalized and standardized range. Such a constraint is critical for maintaining the fidelity and consistency of the output waveforms, making them suitable for further auditory analysis and applications.

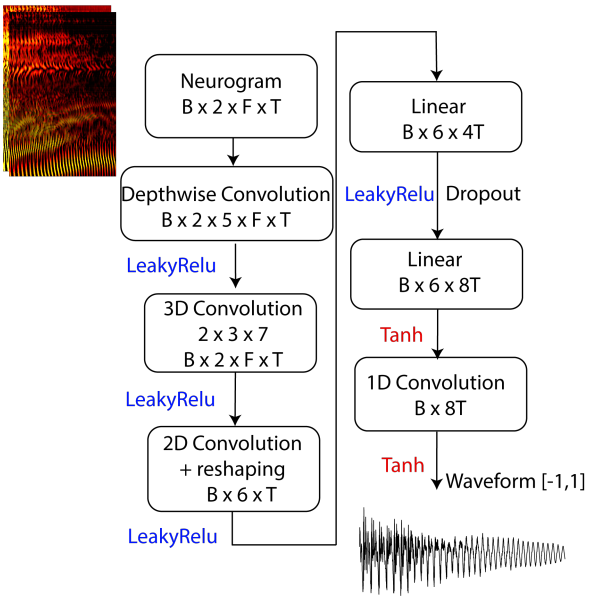

Figure 5. iAN Model Architecture

4. EXPERIMENTS / RESULTS

4.1 Inverse Model Results

4.1.1 Loss Function

Initially, the iAN was trained using a mean-squared error (MSE) loss function, which resulted in suboptimal performance. Notably, the model demonstrated proficiency in reconstructing frequencies below 3 kHz, but struggled with frequencies above this threshold, where virtually no energy was present. This issue can be attributed to the inherent properties of the neurogram and the limitations of MSE in encoding phase information for high-frequency components.

In an effort to address this, a loss function was devised that adapts to the frequency content of the signal. The approach involved matching the envelope and variance for high-frequency segments, where phase distinctions are less perceptually relevant, and focusing on time-domain accuracy for low-frequency components, sensitive to phase variations. The normalized spectral centroid over time, denoted as $S_c(t)$, was used to distinguish between high and low-frequency segments. The modified loss function was defined as follows:

$$L(y, \hat{y}) = (1 - S_c)\alpha L_{\text{MSE}}(y, \hat{y}) + S_c(\beta L_{\text{MSE}}(y_{\text{ENV}}, \hat{y}_{\text{ENV}}) + \gamma L_{\text{MSE}}(\sigma^2, \hat{\sigma}^2))$$

This approach yielded a slight improvement in the model's performance, particularly for listening samples, yet it did not fully resolve the issue. The complexity arises from the multiple ways of matching the envelope and variance, especially in signals where low-frequency energy coexists with significant high-frequency content.

To further refine the model, the final loss function was crafted as follows:

$$L(y, \hat{y}) = \alpha L_{\text{MSE}}(y_{LP}, \hat{y}_{LP}) + \beta L_{\text{MSE}}(\text{STFT}(y_{HP}), \text{STFT}(\hat{y}_{HP}))$$

In this formulation, the first term computes the MSE for low-pass filtered signals (cutoff at 1 kHz with a first-order filter) in the time domain. The second term computes the MSE of the magnitude of spectrograms for high-pass filtered signals, focusing on accurately representing the magnitude of high-frequency content.

4.2 Sensitivity Evaluation

4.2.1 Time domain sensitivity

When reconstructing a speech segment using the iAN, one must define what the hop is and whether any windowing will be used in reconstruction upon combining the segments. To understand what the best approach to this, I calculated average value of the loss function at each of the 400 time steps (25 ms) that the model outputs on 10 minutes of unseen speech.

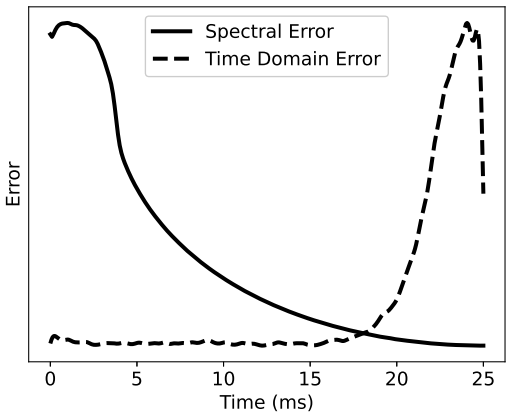

Figure 6. Errors over the predicted waveform window (25 ms at 16kHz: 400 predicted datapoints). Spectral error is the MSE error of high-passed spectrogram. Time domain error is the MSE error of low-passed time domain waveform

In figure (6) we see that at the edges, either the high or the low frequency errors are larger than at the center, indicating that the model outputs better fidelity waveforms in the middle range. The error difference may come from the fact that different delays are associated with different frequencies in the analytical auditory model. This analysis is informative as the full speech reconstruction now may be done with hop size of 1/2 of a window length (12.5ms) and weighted with a Hamming window.

4.3 Reconstruction Quality

When listening to the samples, an astounding quality of reconstructing the color of someone's voice may be observed. While the speech is arguably degraded, the color of the speaker's voice is extremely well matched. Since we primarily perceive color from the timbre - where both phase and magnitude of vowel harmonics matter, it is no suprise that vowels are predicted very accurately.

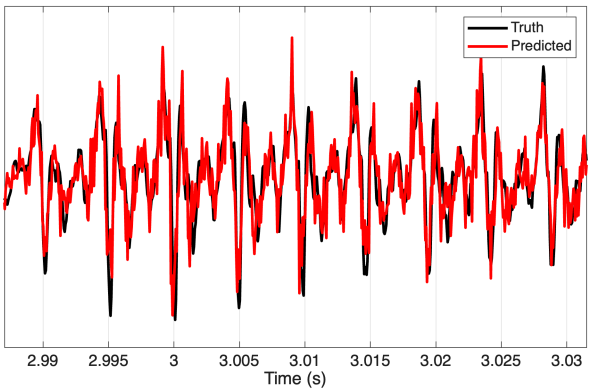

Figure 7. Model response to vowels showing almost perfect reconstruction.

Despite implementing significant modifications in various iterations of the model to achieve improved recon-

struction across all frequency ranges, the reconstruction of high-frequency regions continues to pose challenges. A plausible explanation for this difficulty lies in the unique response characteristics of high-frequency auditory nerve fibers (ANFs). These fibers are known to respond to low-frequency tones and can phase-lock to the envelope of sound waves, meaning their firing patterns are influenced by the entire waveform, not just the high-frequency components. This phenomenon adds a layer of complexity to the task of accurately reconstructing high-frequency regions.

A critical factor contributing to this challenge is the architectural design of the current model. Specifically, the convolutional kernels in our neural network are not configured to span a broad range of frequencies. This limitation restricts the model's ability to learn and replicate more complex patterns that are characteristic of high-frequency auditory responses. The absence of wide-ranging frequency convolutional kernels means that the model may not fully capture the intricate interactions between different frequency components of the sound, particularly in the high-frequency domain. Addressing this architectural limitation could be key to enhancing the model's ability to reconstruct high-frequency regions with greater accuracy and fidelity.

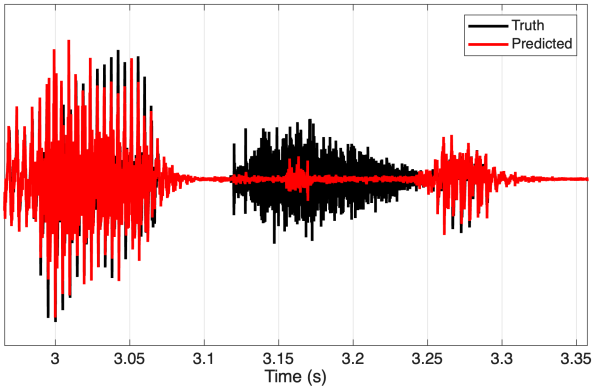

Figure 8. Model response to a vowel followed by a fricative. Fricative is not predicted well, indicating poor performance for sounds with a high spectral centroid.

4.4 Hearing Hearing Loss

The ultimate aspiration of inverting the Auditory Nerve (AN) model lies in its potential to simulate the auditory experience of hearing loss. By analytically generating neurograms that reflect the characteristics of impaired hearing, and then processing these through the inverse auditory nerve model (iAN) designed for healthy hearing, we can theoretically reconstruct the auditory perception of individuals with hearing loss.

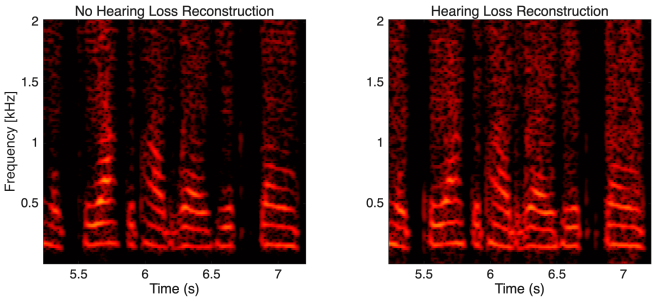

Figure 9. Spectrograms of two predicted waveforms. Left is the prediction made on a healthy neurogram. Right is the prediction made on a hearing loss neurogram (compensated for overall gain reduction).

Figure (9) shows the low frequency regions of the STFT of two reconstructed waveforms - one from a neurogram generated by a hearing loss cochlea, and one with a normal hearing. To match the energy loss, the hearing loss neurogram was generated at 90 dB SPL allowing the spectral energy scales in the reconstructed spectrograms to be similar. Additionally, figure (10) shows a frequency response of a vowel extracted from the same spectrograms shown in figure (9).

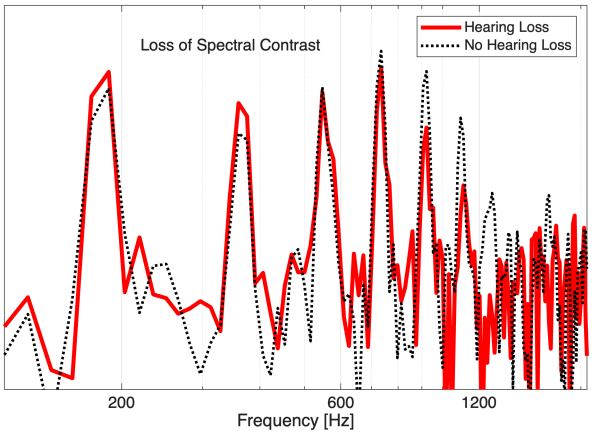

Figure 10. Frequency response of the inverted waveforms for a vowel.

There are two primary differences we can note in the reconstructed waveforms. First is that the SNR is degraded, which may be seen by increased background activity in the high frequency region (since this is all relative, it only means SNR is reduced) in figure (9). The second one may be seen in figure (10) where we note that spectral contrast is reduced. Spectral contrast, particularly in the context of human speech, plays a vital role in the process of formant identification, which is essential for recognizing distinct voices and vowel sounds. This identification relies on discerning the spectral peaks and valleys that constitute formant frequencies. In a typical auditory scenario (represented by the dashed black line in the figure), there is a significant amplitude difference between a spectral peak (e.g., at 700 Hz, representing F2) and a neighboring spectral valley (e.g., at 350 Hz). However, in the case of hearing loss,

this contrast diminishes: the spectral valleys become less pronounced (increase in amplitude), and the peaks reduce in amplitude. This effect not only attenuates the overall sound but also 'blurs' the frequency axis, thereby diminishing the distinctiveness of formant frequencies. Such a reduction in spectral contrast effectively leads to further degradation of the auditory signal, complicating the process of speech perception and recognition in individuals with hearing loss

5. DISCUSSION

5.1 Limitations and Improvements

When training the inverse auditory nerve model (iAN), several challenges emerge. Firstly, the process of generating data is notably time-consuming. This is largely attributed to the inherent nonlinearities in sound coding, which restrict the feasibility of data augmentation. Additionally, scaling different components of the loss function to accurately predict both low and high-frequency components of the signal with high fidelity presents a considerable challenge.

Another aspect not addressed in this study is the nonlinearity in auditory nerve fiber (ANF) responses related to temporal processes, such as adaptation. These processes involve the history of stimuli affecting the ANF firing rate. While not included in the current architecture, future iterations of the model could incorporate elements like Long Short-Term Memory (LSTM) layers to capture this temporal aspect.

A significant limitation of the current model is its restriction to decoding neurograms at a single loudness level. Expanding the model to encompass the full dynamic range, which involves outputting waveforms that vary dramatically in scale (from [-1,1] to [-100000,100000]), is a complex task. This extreme range represents a substantial nonlinearity that poses difficulties for consistent and accurate waveform reconstruction.

Regarding the interpretation of results, caution is paramount. A critical question arises: How can we distinguish whether the degradation observed in reconstructed audio is a result of changes in the neurogram due to hearing loss, or a consequence of the model's limited extrapolation capabilities? To address this in future research, it would be prudent to quantify the extent of degradation in specific auditory cues predicted by the model and compare these findings with results from psychoacoustic experiments.

Additionally, an important consideration in the context of healthy hearing is the role of efferent feedback from the brain, which has been shown to enhance spectral contrast [4]. This feedback mechanism, active in healthy ears, is instrumental in optimizing auditory perception. However, it is absent or altered in hearing-impaired conditions. Therefore, even with our model simulating hearing loss, there remains a gap – our brains, equipped with healthy auditory processing, are adept at enhancing what we hear, including increasing spectral contrast. This natural optimization means that our simulation, while insightful, may not fully replicate the true experience of hearing loss. This discrepancy highlights the complexity of auditory perception and the challenges in creating authentic simulations of impaired hearing.

6. REFERENCES

[1] Bruce, I. C., Erfani, Y., Zilany, M. S. A. (2018). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. Hearing research, 360, 40–54.

[2] Souza P. E. (2002). Effects of compression on speech acoustics, intelligibility, and sound quality. Trends in amplification, 6(4), 131–165.

[3] Zilany, M. S., Bruce, I. C., Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. The Journal of the Acoustical Society of America, 135(1), 283–286.

[4] Farhadi, A., Jennings, S. G., Strickland, E. A., Carney, L. H. (2023). Subcortical auditory model including efferent dynamic gain control with inputs from cochlear nucleus and inferior colliculus. The Journal of the Acoustical Society of America, 154(6), 3644–3659.