Effect of cochlear compression on predicted speech in noise perception

Student Name：	Liu Yuyang
Student Number：	490525
Major：	EECS
Supervisor：	Ing. Václav Vencovský, Ph.D.

Declaration:

"I hereby declare that this bachelor's thesis is the product of my own independent work and that I have clearly stated all information sources used in the thesis according to Methodological Instruction No.1/2009 – "On maintaining ethical principles when working on a university final project, CTU in Prague. "

18/05/2024 **Signature**

BACHELOR‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Liu Yuyang** Personal ID number: **490525**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Electrical Power Engineering**

Study program: **Electrical Engineering and Computer Science**

II. Bachelor’s thesis details

Bachelor’s thesis title in English:

Effect of Cochlear Compression on Predicted Speech in Noise Perception

Bachelor’s thesis title in Czech:

Vliv nelineární odezvy v kochlee na vjem řeči v šumu

Guidelines:

Use an auditory model composed of a dual resonance non-linear (DRNL) filterbank to predict the similarity between the auditory model outputs for speech and for speech in bubble noise. The objective is to investigate how the overall signal intensity, specifically the intensity of speech in bubble noise, influences the predicted results. The model predictions will be compared with results obtained from a listening experiment conducted using the same speech signals. The preliminary experimental findings suggest that there is an improvement in perception with an intensity of up to approximately 70 dB sound pressure level. The thesis aims to shed light on whether cochlear compression plays a significant role in this perceptual improvement.

Bibliography / sources:

[1] Dubno, J.R., Ahlstrom, J.B., Wang, X., Horwitz, A.R. Level-Dependent Changes in Perception of Speech Envelope Cues. J. Assoc. Res. Otolaryngol. 13, 835–852 (2012). https://doi.org/10.1007/s10162-012-0343-2
[2] Lopez-Poveda, E., Meddis. R. A human nonlinear cochlear filterbank. J. Acoust. Soc. Am., 110 (6): 3107–3118. https://doi.org/10.1121/1.1416197

Name and workplace of bachelor’s thesis supervisor:

Ing. Václav Vencovský, Ph.D. Department of Radioelectronics FEE

Name and workplace of second bachelor’s thesis supervisor or consultant:

Date of bachelor’s thesis assignment: **19.09.2023** Deadline for bachelor thesis submission: **09.01.2024**

Assignment valid until: **16.02.2025**

Ing. Václav Vencovský, Ph.D.	doc. Ing. Zdeněk Müller, Ph.D.	prof. Mgr. Petr Páta, Ph.D.
Supervisor’s signature	Head of department’s signature	Dean’s signature

III. Assignment receipt

The student acknowledges that the bachelor’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor’s thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student’s signature

Acknowledgments:

I want to sincerely thank my supervisor, Václav Vencovský, for his invaluable assistance and exceptional patience while writing this thesis. His guidance and support have been instrumental in the completion of this work. I would also like to thank my university for providing precious learning opportunities and a high-quality study environment. Lastly, I must thank my family for their support during my studies and the writing of this thesis.

ABSTRACT

This thesis investigates the role of cochlear compression in speech perception, particularly in noisy environments, using an advanced auditory model. The primary focus is on the analysis of speech in multitalker babble noise by using an auditory model composed with a dual resonance non-linear (DRNL) filterbank. A key aspect of this study is examining how the overall signal intensity, meaning the intensity of speech and babble noise with constant signal-to-noise ratio (-5 dB), influences these predictions. The research employs a systematic approach, comparing the auditory model's predictions and showing the similarity between the model outputs to speech and speech plus babble noise with actual experimental listening test results from the literature. The currently used listening test results reveal a notable improvement in speech perception in noise at intensities up to about 70 dB sound pressure level (SPL). In contrast, the model predictions presented in this thesis do not fully show improvement with increasing sound levels. This discrepancy highlights the complexity of cochlear compression and suggests that additional factors beyond cochlear compression may contribute to speech perception in noisy environments. The findings are pivotal in understanding the extent to which cochlear compression contributes to this improvement.

Contents.

1 INTRODUCTION

1.1 Introduction and Motivation for the Study of Cochlear Compression in Speech Perception

The human auditory system, particularly the cochlea, plays a pivotal role in perceiving and understanding speech, especially in challenging listening environments. A key aspect of this system is cochlear compression, a dynamic process that significantly influences our ability to discern speech amidst background noise. This phenomenon becomes particularly evident at moderate to high sound levels, roughly ranging from about 30 to 60 dB sound pressure level (SPL). Recent studies in the field further substantiate this observation, highlighting that the non-linear processing characteristics of the cochlea are most noticeable within this specific range of sound intensities [1]. Such findings are crucial as they provide a deeper understanding of the cochlear mechanics under various auditory conditions.

This cochlear compression is a nonlinear process that takes place inside the cochlea, which is responsible for converting sound vibrations into signals that our brain can interpret. The cochlea contains several key structures, including the basilar membrane, Reissner's membrane, and the tectorial membrane, along with hair cells that are critical for converting sound vibrations into neural signals. These components allow for a response to frequencies and sound levels. When sounds are at moderate levels (30 – 60 dB SPL), the basilar membrane responds nonlinearly, a characteristic that facilitates cochlear compression [2]. This compression effectively reduces the range of intensities we perceive, making it easier for our auditory system to handle sounds with varying intensities.

When it comes to perceiving speech, the presence of cochlear compression in the healthy cochlea and its absence when cochlear amplification is compromised is notable. Although speech signals are complex with huge frequency and dynamic range, the role of cochlear compression is not yet fully understood. Studies suggest that it plays a significant role in enhancing speech perception, particularly in noisy environments. For instance, research by Dubno et al. highlights that cochlear compression affects the perception of speech envelope

cues, which may be crucial for understanding speech against background noise [3].

Dubno et al. investigated speech perception in noise as a function of stimulus intensity, specifically examining how speech and noise intensities impact understanding when the signal-to-noise ratio is held constant. Their findings indicated that comprehension improves as intensity increases up to about 60 dB SPL, beyond which the understanding of speech in noisy environments deteriorates [3]. These insights suggest that cochlear compression plays a pivotal role in speech perception amid background noise. Similarly, the work presented by Vencovský and Bureš [4] in a poster, which discusses 'Level-Dependent Responses to Speech in Noise derived from a Nonlinear Cochlear Model,' also yielded results that align with the observations by Dubno et al. [3]. This supports the notion that cochlear compression is crucial for enhancing speech intelligibility in noisy settings.

This thesis presents work employing a nonlinear filterbank cochlear model. Unlike the model used in Vencovský and Bureš [4], the model used in this thesis incorporates a nonlinear filterbank. Therefore, this model enables the study of cochlear compression on speech perception in noisy environments. In particular, the aim is to utilize this Dual Resonance Non-Linear (DRNL) model to predict the similarity between the model outputs in r response to speech combined with multitalker (babble) noise.

2 Methodology

In this section, we comprehensively describe the methodologies employed in this study. We begin by detailing the DRNL (Dual Resonance Non-Linear) filterbank model and the specific implementation used in our research. Following this, we describe the auditory model in which the DRNL filterbank is integrated, highlighting its essential components and functionalities. Finally, we outline the method used to calculate the correlation between the auditory model outputs in response to isolated speech and speech combined with multitalker noise.

2.1 OVERVIEW OF DUAL RESONANCE NON-LINEAR (DRNL) FILTERBANK

The Dual Resonance Non-Linear (DRNL) filterbank, as detailed in the work of Lopez-Poveda and Meddis [5], is a sophisticated auditory model designed to simulate the human cochlear response to sound. This model is particularly significant for replicating the complex, nonlinear processing of sounds in the cochlea, which is crucial for speech perception and auditory signal processing.

2.1.1 Simplified Description of the DRNL Filterbank:

The DRNL filterbank is a computational model designed to simulate the nonlinear characteristics of the human cochlea. It features a dual-pathway structure incorporating both linear and nonlinear responses to sound. The nonlinear path handles low-level sounds, employing dynamic compression to model the cochlea's complex response to these quieter sounds. As the intensity increases, the response becomes saturated, introducing compression. Conversely, the linear path becomes dominant at higher sound levels, maintaining a straightforward response with minimal distortion. Figure 1 shows the exact process, but note that in the case of the model used in this thesis, the nonlinearity is not a broken stick nonlinearity but a different nonlinearity described below. This compression is vital for representing how the cochlea handles a range of sound intensities, making the DRNL model particularly useful for understanding auditory processes such as speech perception. Integrating cascaded gammatone filters in both pathways allows the DRNL model to capture the frequency selectivity observed in psychoacoustic experiments, effectively simulating the auditory filtering performed by the basilar membrane [6].

This model, therefore, provides a robust and computationally effective framework for exploring how different sound levels affect cochlear behavior, making it a powerful tool for auditory research and applications like hearing aid design. Additionally, Ray Meddis aimed to integrate this model into a larger auditory periphery model to create a profile of a listener with hearing

loss.

2.1.2 Other Key Features of the DRNL Filterbank:

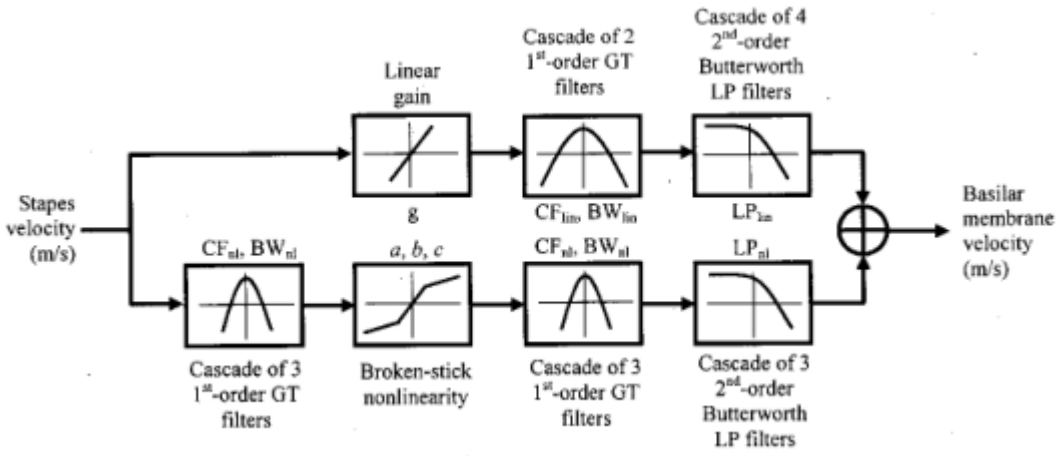

Fig.1 Diagram of the DRNL filter, taken from [5]. The parameters of each block are shown in the space between the linear (top) and the nonlinear (bottom) paths. The output signal from the DRNL filter is the sum of the signal coming out of each path. Note that in the case of the model implementation used in this thesis, the broken stick nonlinearity is adapted to the so-called "Nick's compression algorithm" described in subsection 2.1.3.

A vital component of the DRNL model is the Gammatone (GT) filters, which are bandpass filters.GT filters are used to simulate the response of the basilar membrane in the human auditory system. GT filters are bandpass filters capable of extracting signal components in a specific frequency range. The center frequency and bandwidth can be adjusted as required to simulate the response of the basilar membrane at different locations in the human cochlea. These filters have center frequencies based on computational rules designed to cover the entire hearing range (approximately 20 Hz to 16 kHz). This method of frequency spacing correlates with the frequency distribution found in natural sounds, ensuring that the model accurately mimics human auditory processing. In addition to the GT filters, the DRNL filter also includes cascades of second-order Butterworth lowpass filters, which further contribute to the accurate simulation of the cochlear response.

Applications and Significance:

The DRNL filterbank is a sophisticated model that simulates the nonlinear processing of sounds

in the human cochlea. It integrates linear and nonlinear pathways to handle varying sound levels, with the nonlinear path mimicking cochlear amplification for low-level sounds and the linear path ensuring minimal distortion at higher sound levels. This model is crucial for predicting the similarity between the cochlea's response to isolated speech and to speech combined with babble noise, which is essential for understanding complex auditory tasks such as speech perception.

2.1.3 Compression algorithm in DRNL Model—Nick's compression algorithm

In the Dual Resonance Non-Linear (DRNL) model, the nonlinear compression algorithm plays a crucial role in simulating how the cochlea processes the sound of varying intensities. The "Nick's Compression Algorithm," as denoted in the model implementation from the MATLAB auditory periphery [7], offers an efficient method to emulate this biological process by dynamically adjusting the nonlinear output to accommodate different sound pressure levels. The essence of this algorithm lies in its segmental handling of the input signal, comparing the absolute value of the signal to a compression threshold to determine if compression is needed.

The operational principle of the algorithm is as follows: it first calculates the absolute value and the sign of the nonlinear output. When the absolute value of the nonlinear output is below the compression threshold, the algorithm processes this portion of the signal linearly, multiplying it by a predetermined gain coefficient. For parts exceeding the threshold, the algorithm employs an exponential function for compression, which not only considers the magnitude of the signal but also preserves its original sign, ensuring that the processed signal remains consistent with compressively growing basilar membrane response at moderate levels.

Figure 2 depicts the transfer function of Nick's Compression Algorithm within the DRNL model. The x-axis represents the input signal amplitude, while the y-axis represents the algorithm's output. The graph illustrates which branch of the DRNL filter dominates. At the lowest intensities, the nonlinear branch yields a larger response and, therefore, dominates the model output. In contrast, the linear branch starts to dominate as the stimulus amplitude grows.

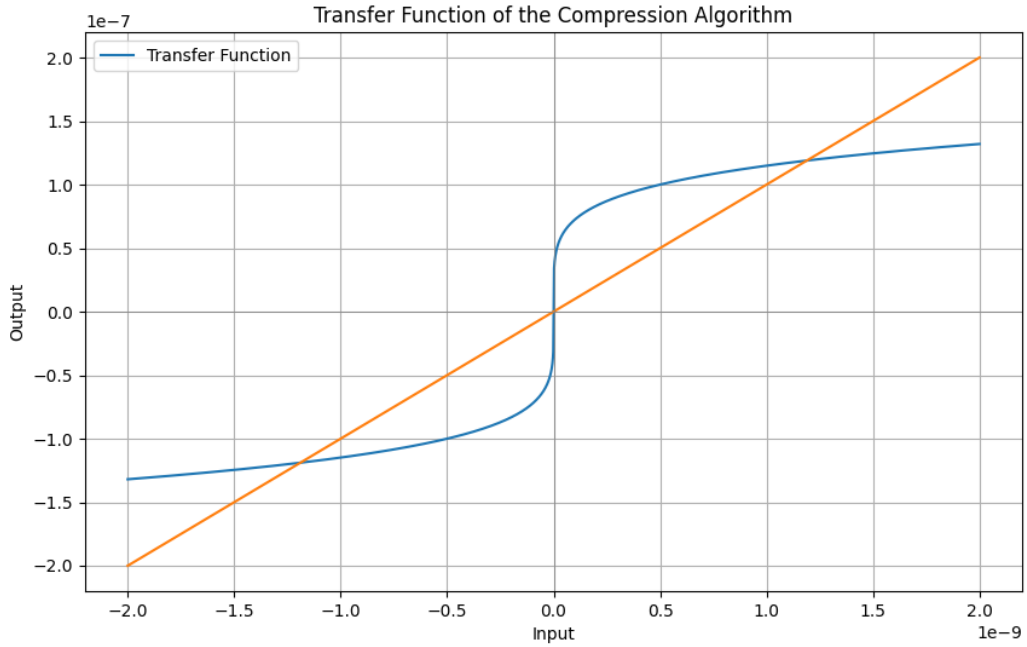

Fig.2 The transfer functions for the algorithm's linear (orange line) and nonlinear (blue line) branches. The nonlinear branch produces larger amplitudes at low intensities, demonstrating its role in handling low-level sounds with dynamic compression. The linear branch begins to dominate as the input signal intensity increases and exceeds the cochlear compression threshold, providing a more straightforward response with minimal distortion.

Figure 3 expands the input range, providing a broader view of the algorithm's response to input signal intensity. The x-axis spans a broader range of input values, and the y-axis displays the corresponding output. The graph showcases three distinct lines: the blue line, the same as in Figure 2, represents the transfer function of the exponential compression, indicating the model's nonlinear response. The orange line shows the model's transfer function at the lowest intensities, where it behaves linearly. The green curve, which corresponds to the linear branch's transfer function shown as the orange line in Fig. 2, appears very small here due to the much smaller scale on the x-axis than in Fig. 2. This demonstrates the compression effect more extensively: inputs below the threshold experience linear amplification (orange line), while those above it undergo exponential compression (blue line). The algorithm's ability to maintain linearity at lower intensities and apply compression at higher intensities is crucial for realistic auditory modeling.

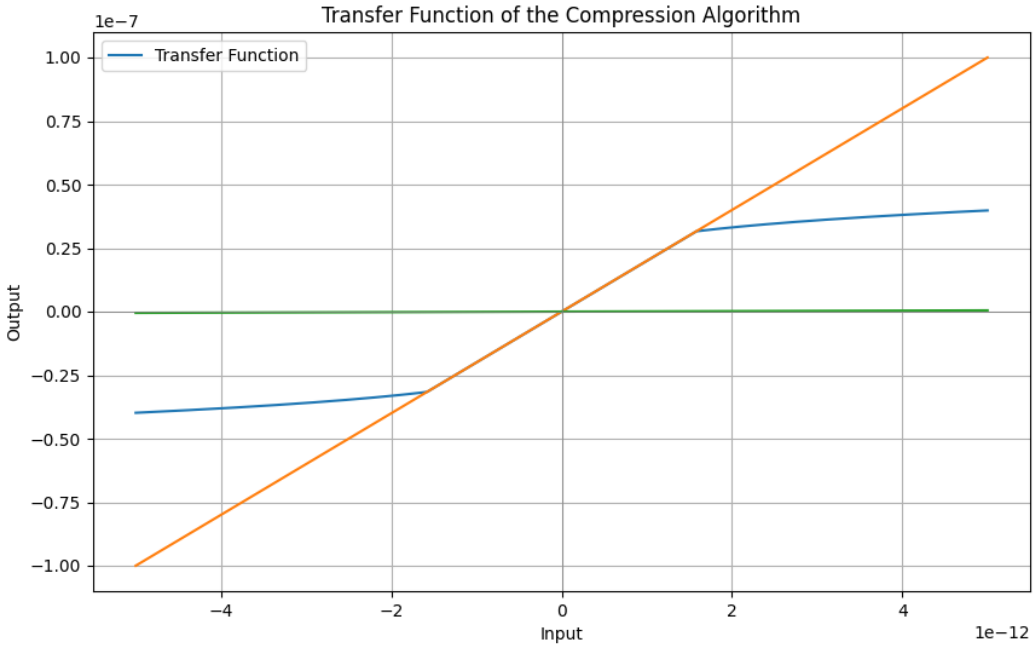

Fig.3 Extended view of the transfer function for Nick's compression algorithm. The orange line demonstrates the transition from linear gain to logarithmic compression across a broader input range, illustrating the algorithm's dual response characteristic.

Introducing this compression algorithm significantly enhances the practicality and accuracy of the DRNL model, enabling it to mimic the dynamic compression effects of the human cochlea more precisely during actual auditory processing. This way, the model better reflects the cochlear response to different sound intensities and demonstrates the characteristics of nonlinear compression in handling high-intensity sounds, which is crucial for improving the design of hearing aids and optimizing auditory signal processing algorithms.

2.2 DRNL Model Implementation and Characteristics:

The DRNL (Dual Resonance Non-Linear) filterbank has been digitally implemented in MATLAB using the framework provided by the MATLAB Auditory Periphery (MAP) project, which is freely available on GitHub [7]. This section first describes the implementation specifics of the model, followed by an exposition of its characteristics.

Implementation Details:

The DRNL filterbank is implemented in the time domain and includes various components like GT filters and linear lowpass (LP) filters as part of its architecture. This implementation is available in MATLAB, utilizing a sampling rate of 44.1 kHz. This sampling rate is chosen because it is high enough to accurately capture the frequency range of human hearing, which extends up to approximately 20 kHz. The sampling frequency determines the highest frequency that can be accurately represented, which is half of the sampling rate. Therefore, a sampling rate of 44.1 kHz ensures that frequencies up to 22.05 kHz can be accurately processed, encompassing the entire range of audible sounds.

GT filters:

This filter has an impulse response to the form. [8]

$$h(t) = kt^{(n-1)}\exp(-2\pi Bt)\cos(2\pi f_c t + \phi) \quad (t \geq 0), \tag{1}$$

$$h(t) = 0 \quad (t < 0),$$

where n is the order of the filter, B is the bandwidth, fc is the center frequency, φ is the phase and *k* are gain.

The DRNL filter uses several cascades of first-order (n = 1) GT filters only. They were implemented digitally as an infinite impulse response filter as follows [9]:

$$y[i] = a_0 \cdot x[i] + a_1 \cdot x[i-1] - b_1 \cdot y[i-1] - b_2 \cdot y[i-2], \tag{2}$$

where *[i]* refers to the ith sample of the digital signal, *x,* and *y* are the input and output signals to/from the filter, respectively, and the coefficients are calculated as follows:

$$a_0 = \left| \frac{1 + b\_1\cos(\theta) - jb\_1\sin(\theta) + b\_2\,co1 + b\_1\cos(\theta) - jb\_1\sin(\theta) + b\_2\cos(2\theta) - jb\_2\sin(2\theta)}{1 + \alpha\cos(\theta) - j\alpha\sin(\theta)} \right|, \tag{3}$$

where $a1=\alpha \cdot a0$, $b1=2\alpha$, *and* $b2=\exp(-2\phi)$. Here, $\theta=2\pi fc \cdot dt$, $\phi=2\pi B \cdot dt$, $\alpha=-\exp(-\phi)\cos(\theta)$, and $j=-1$. The variable dt represents the sampling period of the digital signal.

The low-pass filters:

The low-pass filters in the DRNL model are implemented as second-order Butterworth lowpass filters. The digital implementation formula is:

$$y(i) = C \cdot x[i] - 2 \cdot C \cdot x[i-1] + C \cdot x[i-2] - D \cdot y[i-1] - E \cdot y[i-2], \tag{4}$$

Where the coefficients that define the behavior of the filter are:

$$C = \frac{1}{1 + \sqrt{2} \cdot \cot(\pi f_c dt) + \cot^2(\pi f_c dt)}, \quad (5)$$

$$D = 2 * C * (1 - cot\text{^}2(pi * fc * dt)), \quad (6)$$

$$E = C * (1 - \sqrt{2}\, cot(pi * fc * dt) + cot\text{^}2(pi * fc * dt)), \quad (7)$$

where *fc* is the 3-dB-down cutoff frequency of the low-pass filter, *dt* is the sampling period of the digital signal.

The linear gain

The linear gain in the DRNL model is implemented in the digital implementation formula:

$$y[i] = g * x[i], \quad (8)$$

where *[i]* refers to the ith sample of the digital signal, and *x* and *y* are the input and output signals to/from the linear gain stage, respectively.

The nonlinearity

The formula for the nonlinearity component in the DRNL model is given as:

$$y(i) = sign(x([i]) * min(a * abs(x[i]), b * abs(x([i])\text{^}c), \quad (9)$$

Where:

y[i]: Output signal of the nonlinearity at the i-th sample.

x[i]: Input signal to the nonlinearity at the i-th sample.

a, b, c: Parameters of the nonlinearity.

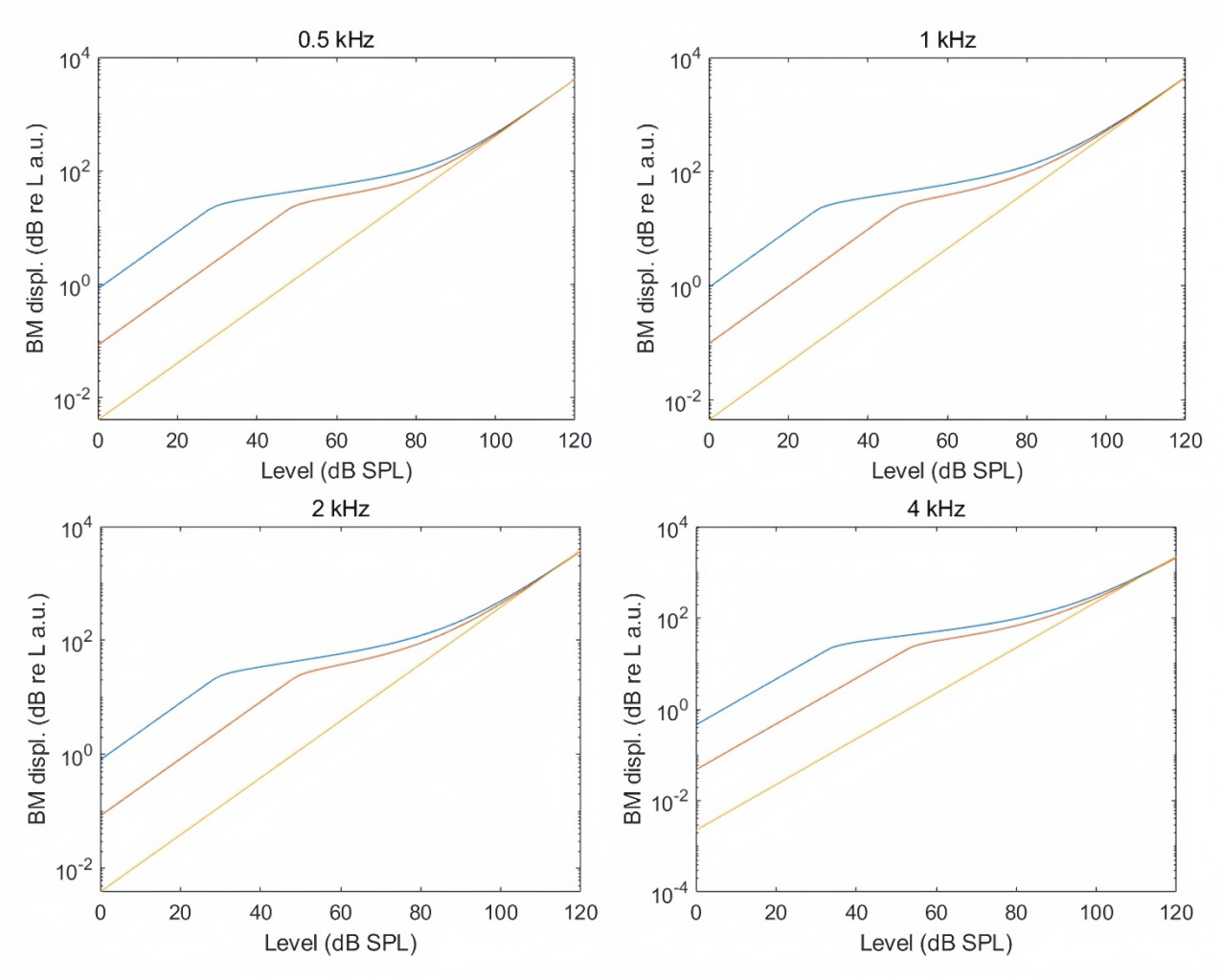

Fig.4 Input/output functions of the cochlear model showing the dependence of basilar membrane displacement (BM displ.) as a function of stimulus intensity for three different settings of the parameter *a*, which affects gain at characteristic frequencies (CFs). The output values were obtained using a pure tone at 0 dB SPL. In this figure, the blue line represents the model with a = 2e4, the red line represents the model with a = 2e3, and the yellow line represents the extrapolated response at the highest intensities, simulating the model with no outer hair cells (OHCs), corresponding to a = 0, indicating no nonlinear path. Note that the actual I/O function might slightly shift in amplitude due to the effect of gain on CF (characteristic frequency).

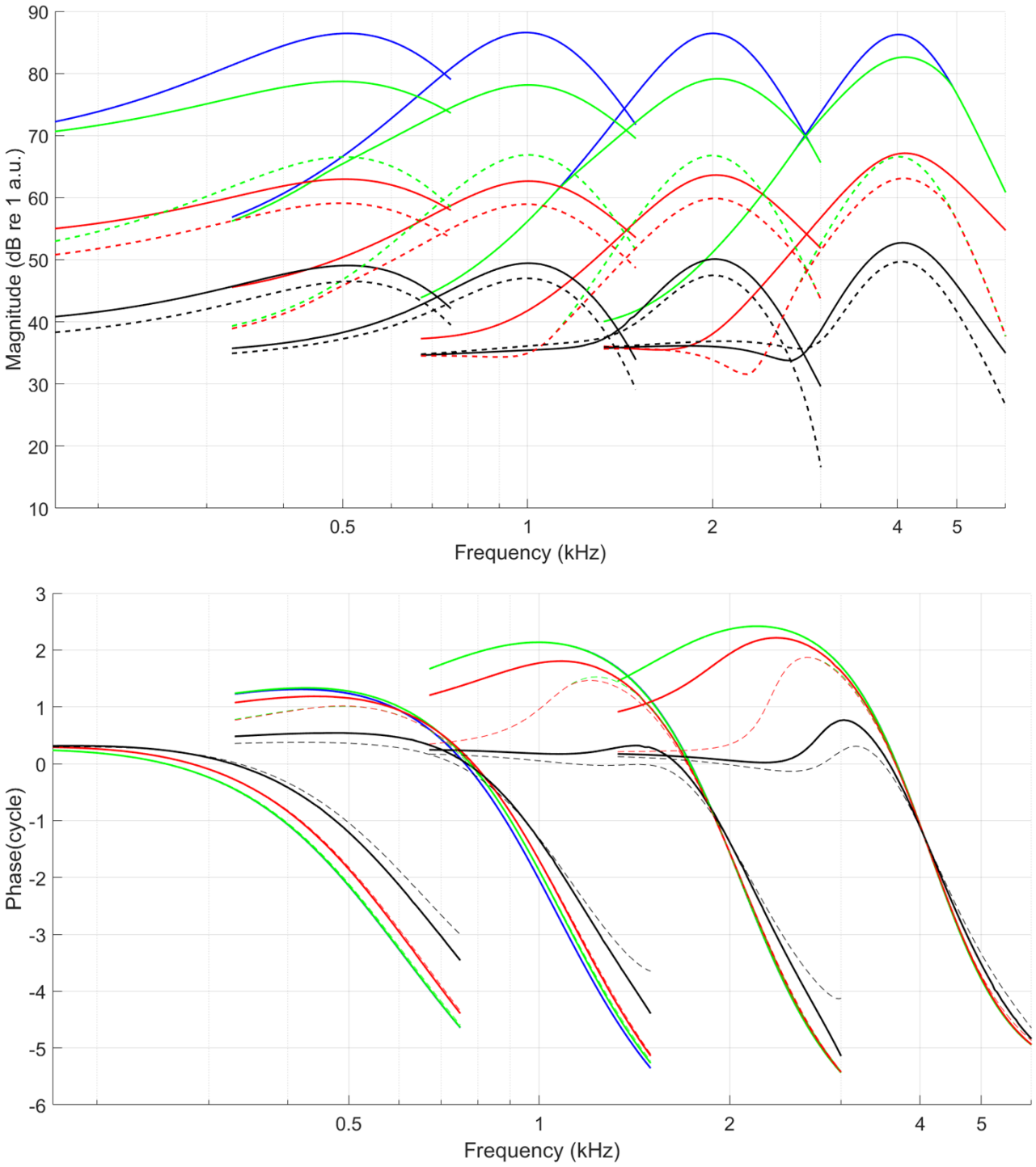

Fig.5 Characteristic transfer functions derived for selected segments of the cochlear model with characteristic frequencies (CF) at 0.5, 1, 2, and 4 kHz. Transfer functions derived from the model with reduced gain (indicated by dashed lines) exhibit wider cochlear filter bandwidths at lower sound intensities. In contrast, at a sound pressure level of 80 dB SPL, the cochlear filter bandwidths tend to converge for models of varying gain. In this figure, the color-coded

responses correspond to different sound pressure levels: blue for 20 dB SPL, green for 40 dB SPL, red for 60 dB SPL, and black for 80 dB SPL, each tracing the model's performance at these respective intensities.

2.3 Analysis of DRNL Model Parameter Effects on Response

In this thesis, we use two variants of the cochlear model, which differ in the amount of cochlear compression, controlled by the parameter *a*. Higher values of *a* result in more pronounced dynamic range compression, enhancing the cochlea's sensitivity to weak signals while compressing strong signals. In contrast, $a = 0$ demonstrates a linear response with no dynamic range compression. A model employing two distinct values for the parameter a, specifically $a = 2e3$ and $a = 2e4$, was used to analyze the basilar membrane (BM) input/output functions.

Figure 4 shows the input/output functions of the cochlear model at characteristic frequencies of 0.5, 1, 2, and 4 kHz for three different gain settings. The output values were obtained using a pure tone at 0 dB SPL. In this figure, the blue line represents the model with $a=2e4$, the red line with $a=2e3$, and the yellow line represents the model without outer hair cells (OHCs), corresponding to $a=0$, indicating no nonlinear path. These input/output functions illustrate how the model's output varies with different gain settings. At low sound levels, the non-linear gain (blue and red lines) produces larger amplitudes, showing the compressive behavior of the model. As the input level increases, the higher gain model (blue line) shows greater compression, indicating a greater ability to handle high-intensity sounds. Conversely, the model without OHCs (yellow line) behaves linearly, showing no compression.

Figure 5 shows the characteristic transfer functions derived for cochlear model segments with CFs at 0.5, 1, 2, and 4 kHz. The transfer functions are shown for different sound pressure levels: blue for 20 dB SPL, green for 40 dB SPL, red for 60 dB SPL, and black for 70 dB SPL. Models with reduced gain (dashed lines) have wider cochlear filter bandwidths at lower sound levels, reflecting lower frequency selectivity. At 80 dB SPL, the cochlear filter bandwidths of models with different gain settings tend to converge, indicating similar frequency selectivity. This

shows how the parameter a affects the model: at lower gains (dashed lines), the filters are wider at low intensities, indicating less precise tuning. As gain increases, the model better represents the non-linear characteristics of the human cochlea, compressing high-intensity inputs and maintaining narrower bandwidths at moderate to high sound levels, thereby improving frequency selectivity.

2.4 DRNL Model Implementation

The Dual Resonance Non-Linear (DRNL) model achieves frequency selectivity through bandpass filters whose bandwidth varies with sound intensity. In this model, the bandwidth of the nonlinear pathway filters is a crucial factor, particularly due to its level-dependent nature. Specifically, the linear branch has wider filters than the nonlinear branch, resulting in a level-dependent bandwidth. This section describes the implementation of the DRNL model, including how the parameter *a* affects the bandwidth and the overall model behavior.

The model was implemented in MATLAB, utilizing a sampling rate of 44.1 kHz. The frequency range covered by the model spans from approximately 20 Hz to 18.734 kHz, providing a detailed frequency resolution. This resolution is consistent with the model used in Vencovský and Bureš, ensuring comparable results and analyses.

Controlling Bandwidths of GT filters:

The bandwidth of the DRNL model determines its frequency selectivity, which is an important characteristic of the auditory periphery. The filter bandwidth is narrowest at the lowest intensities, influencing the model's ability to selectively respond to different sound frequencies across all intensity levels.

By adjusting the order of the nonlinear GT filters, we can manipulate the model bandwidth. A lower filter order results in wider bandwidths, altering the model's response characteristics. In the current model implementation, the bandwidths of each GT filter in both branches of the DRNL filter are shown in Fig. 6.

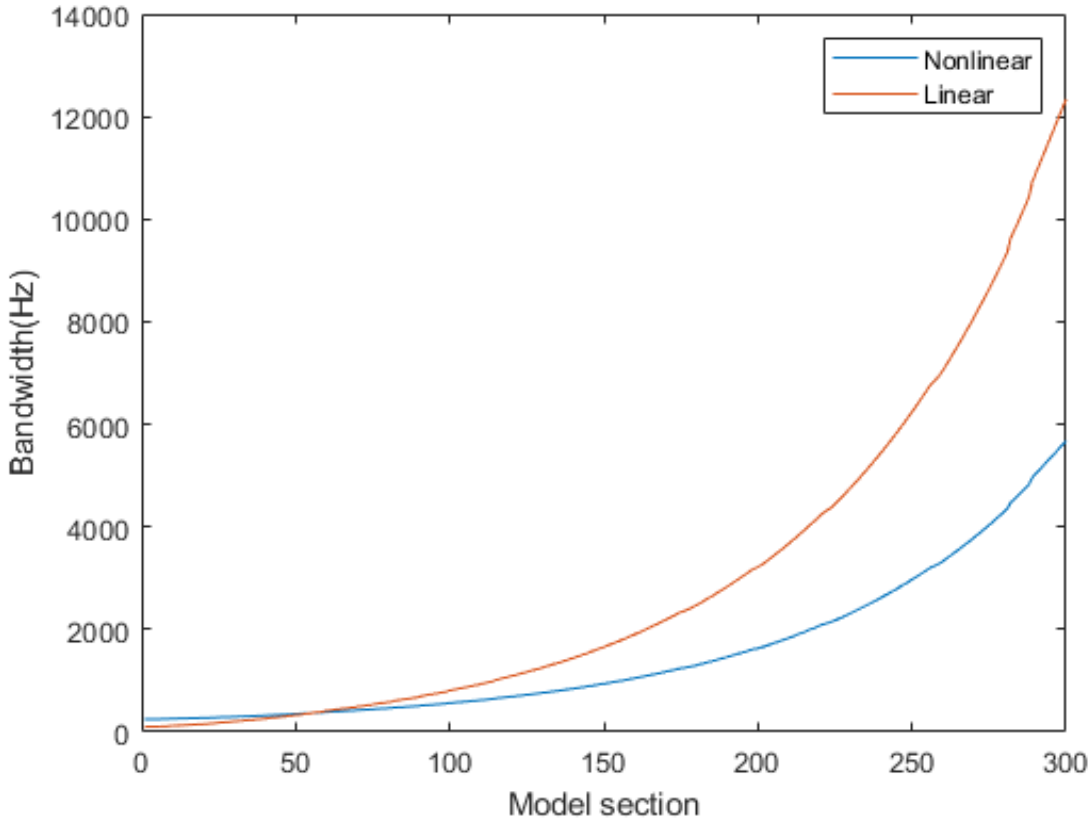

Fig.6 Comparison of nonlinear and linear filter bandwidth across 300 sections of the cochlear model

Figure 6 illustrates the bandwidth distribution of nonlinear and linear filters across 300 sections of the cochlear model, from the base to the apex. The x-axis represents the 300 sections of the cochlear model, each corresponding to a specific CF range. The y-axis represents the filter bandwidth in Hertz (Hz).

The bandwidth of the GT filters in the nonlinear part (blue line): The bandwidth of auditory filters increases with the center frequency, as larger numbers of the model sections are at higher frequencies where the bandwidth is largest. This is consistent with the critical band theory, which indicates that the bandwidth of auditory filters, or critical bands, increases as a function of frequency. Additionally, research has shown that as the intensity of sound grows, the bandwidth also increases, which explains why the bandwidth in the linear part of the DRNL filters is larger than in the nonlinear part [12].

Bandwidth of the GT filters in the linear part (red line): The linear filter bandwidth also changes with the CF.

2.5 Auditory Periphery Model

Within the framework of this thesis, the DRNL filterbank was incorporated into a model of the auditory periphery. This model consists of three main components:

1. Outer and Middle Ear model:

The input acoustic signal undergoes transformation, simulating the outer and middle ear transfer functions. This involves filtering by two parallel 1st-order Butterworth bandpass filters to model the resonances of the outer ear canal. The first filter has a gain of 10 dB with cutoff frequencies at 2.5 kHz and 4 kHz, while the second filter has a gain of 25 dB with cutoff frequencies at 2.5 kHz and 7 kHz. These filtered signals are combined and then processed by a middle ear model composed of a high-pass filter (cutoff frequency 50 Hz) and a low-pass filter (cutoff frequency 1 kHz), transforming the signal into stapes displacement. This model is sourced from the MATLAB Auditory Periphery (MAP) toolbox [7].

2. DRNL Filterbank:

The core component of the model is the DRNL filterbank, which simulates the nonlinear processing of the cochlea. The DRNL filterbank is described in detail above in section 2.1. It includes both linear and nonlinear pathways to replicate the cochlea's response to varying sound levels. The nonlinear pathway in the DRNL model is a hybrid: it responds linearly at the lowest intensities and transitions to compressive nonlinearity as the sound intensity increases. This behavior allows the model to simulate the basilar membrane's response to varying sound levels. At low intensities, the nonlinear path dominates, providing fine-tuned frequency selectivity. As the intensity increases, the linear path becomes more prominent, handling higher sound levels and ensuring the model's overall response remains accurate and robust. This dual-path structure allows the model to accurately represent the cochlear response across a wide range of sound intensities. The implementation details of the DRNL filterbank, including its components, such as GT filters and compression mechanisms, are described in section 2.1.

3. Inner Hair Cell and Auditory Nerve Synapse Model:

Post the DRNL filterbank, the processed signal in each channel of the DRNL filterbank is fed into a model simulating the inner hair cell (IHC) and auditory nerve synapse. This stage includes a half-wave rectifier followed by a low-pass filter (1st-order Butterworth filter with a cutoff frequency of 1 kHz). The signal is then thresholded and converted to a logarithmic scale, representing the internal signal representation at the IHC output [10].

2.6 Method for Assessing Noise Impact on Speech Perception

The primary aim of the thesis is to utilize this auditory periphery model to predict the similarity between model outputs in response to speech alone and speech combined with babble noise. We do not have a cognitive section in the auditory model that would perform speech recognition. Therefore, we assume that cross-correlation between internal representations (i.e., outputs of the peripheral ear model) derived from the model in response to speech and in response to speech + babble noise correlate with speech recognition. We adopt here an approach known from so-called intrusive objective audio quality assessment tools in which the signal under test and a reference are available to the system.

The similarity between the model outputs is quantified using cross-correlation methods adapted from the PEMO-Q approach for objective audio quality assessment [11]. The internal representations of the entire sentences are derived for both clean speech and speech with babble noise, and cross-correlation coefficients are computed as follows:

$$r = \frac{\sum_{t,f=1}^{N,M} (x_{tf} - \bar{x})(y_{tf-\bar{y}})}{\sqrt{\sum_{t,f=1}^{N,M} (x_{tf} - \bar{x})^2 \sum_{t,f=1}^{N,M} (y_{tf} - \bar{y})^2}}, \quad (11)$$

where x_{tf} and y_{tf} represent the time-frequency samples of the internal representations for speech with noise and speech only, respectively, the cross-correlation is calculated over non-overlapping 20-ms time frames, yielding the cross-correlation, which is the function of time providing a temporal similarity function. We use two different approaches to calculate the overall cross-correlation for the entire sentence. The first approach is to compute the 0.05

quantile of the cross-correlation values across all time frames, capturing the lower bound of similarity. The second approach calculates the mean value of the cross-correlation across time, providing an average measure of similarity for the entire sentence. The same approach was used in Vencovský and Bures [4].

3 RESULTS

Figure 7 below illustrates the recognition performance of sentences presented in babble noise as a function of sound pressure level (SPL). The data used to create this figure were taken from Vencovský and Bureš [4]. Listeners were presented with a sentence (Czech meaningful sentences) in babble noise with a signal-to-babble-noise-ratio of -5 dB and asked to repeat the sentence. If the sentence is repeated correctly, the listener is awarded a rating of 100%. If the listener repeats at least half of the words in the sentence correctly, they are awarded a rating of 50%. The overall level of speech and babble noise is variable. At each level, ten sentences are presented. The results are expressed as the mean values across the ratings of ten sentences presented at each level.

Figure 7 shows averaged values across 21 young and normally hearing adult listeners (8 females) with age below 30. It also illustrates the improvement in recognition performance as a function of increasing SPL up to a certain level, after which it may plateau at higher levels.

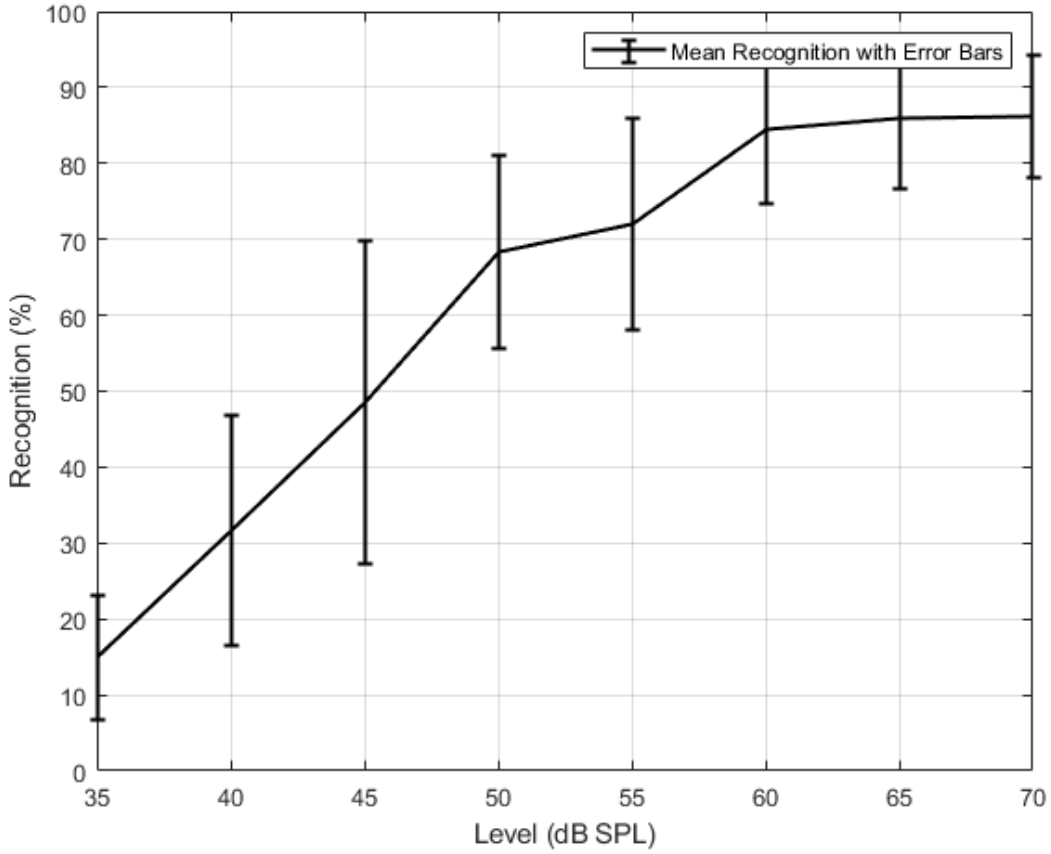

Fig.7 Recognition of sentences in babble noise as a function of stimulus level for fixed speech to babble noise ratio of -5 dB.

Figures 8, 9, 10, and 11 illustrate the cross-correlation coefficients between the model outputs in response to speech and speech + babble noise. Each panel depicts the cross-correlation coefficients for ten different sentences. The cross-correlation coefficients depicted in Figs. 8 and 9 were obtained with the model with a higher gain (parameter a = 2e4), while those illustrated in Figs. 10 and 11 were obtained with the model with a lower gain (parameter $a =$ 2e3). Figures 8 and 10 show the 5% quantile of r as a function of time, and Figs. 9 and 11 show the mean values.

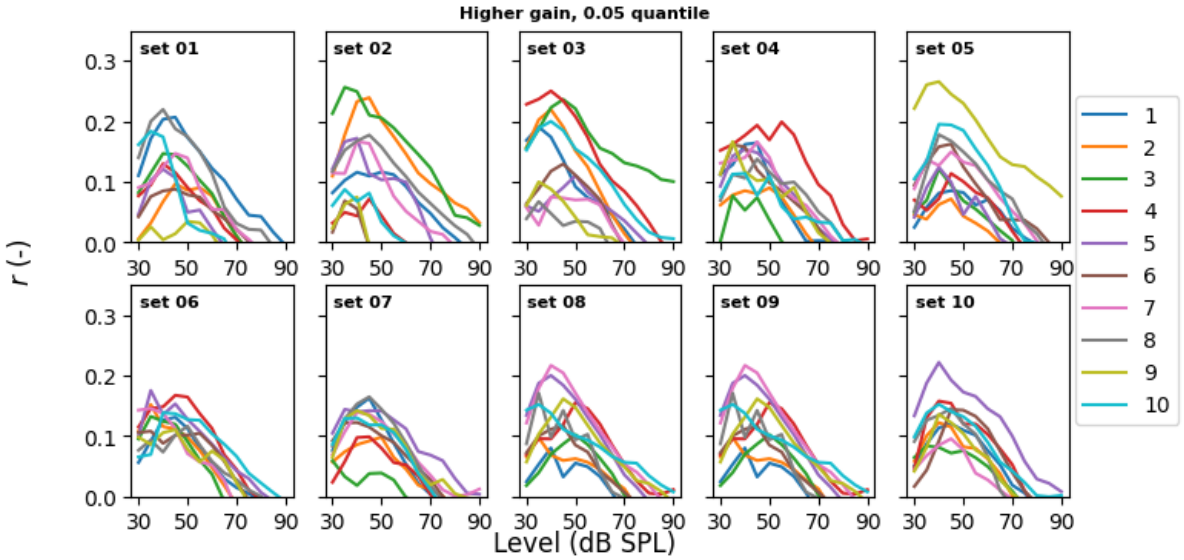

Fig.8 Cross-correlations between model responses for speech + babble noise and speech only. 0.05 quantile is taken across the cross-correlations for adjacent 20-ms time frames. The cochlear model has a higher gain (better hearing threshold).

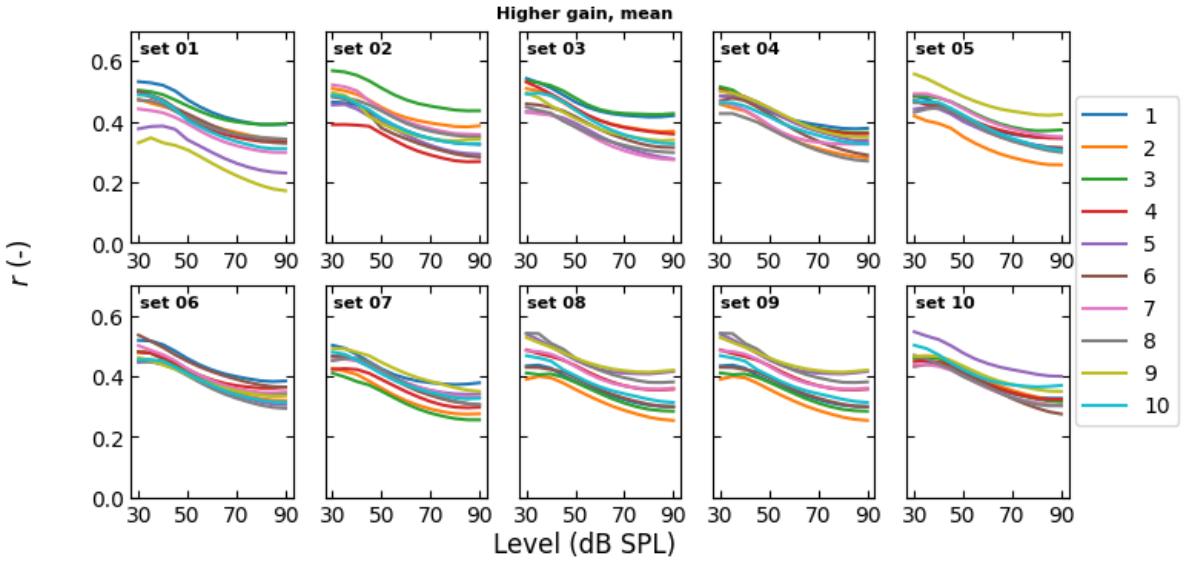

Fig.9 Cross-correlations between model responses for speech + babble noise and speech only. The mean value is taken across the cross-correlations for adjacent 20-ms time frames. The cochlear model has a higher gain (better hearing threshold).

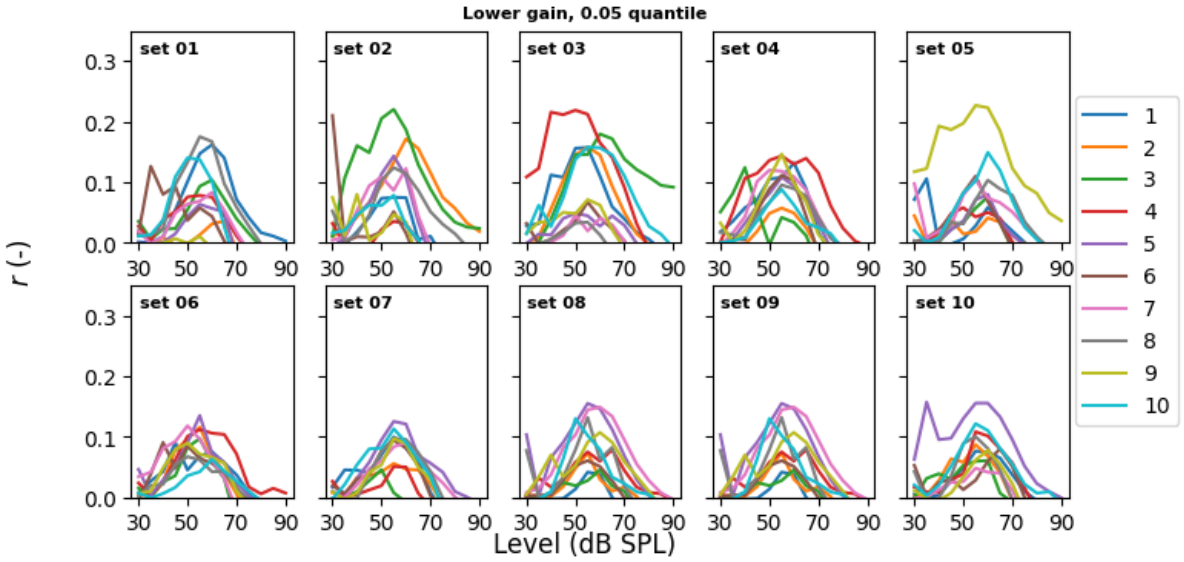

Fig.10 Cross-correlations between model responses for speech + babble noise and speech only. 0.05 quantile is taken across the cross-correlations for adjacent 100-ms time frames. The cochlear model has lower gain (worse hearing threshold).

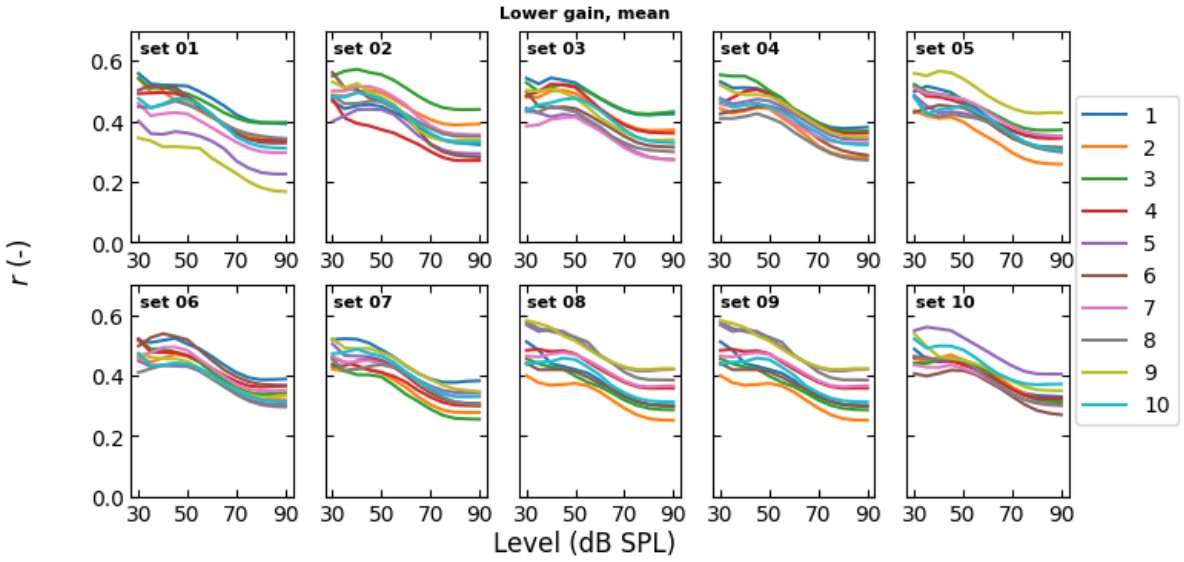

Fig.11 Cross-correlations between model responses for speech + babble noise and speech only. The mean value is taken across the cross-correlations for adjacent 100-ms time frames. The cochlear model has lower gain (worse hearing threshold).

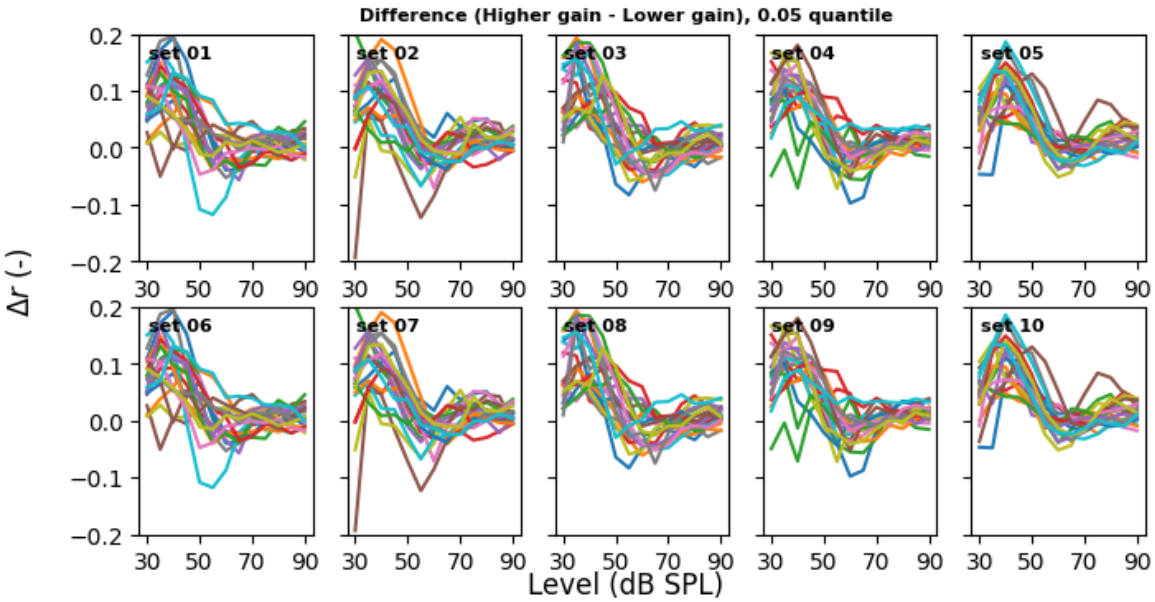

Fig.12 The difference between the cross-correlations for the model with higher gain and the model with lower gain is referred to as Δr. The 0.05 quantile of r(t) is employed in this analysis.

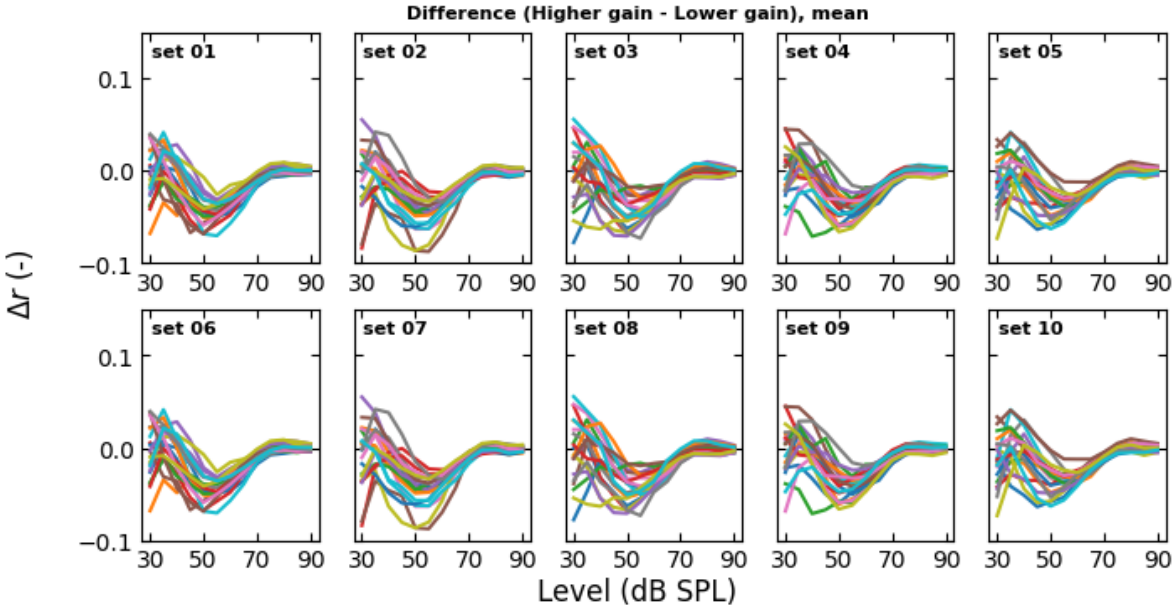

Fig.13 The difference between the cross-correlations for the model with higher gain and the model with lower gain is referred to as Δr. The mean value of r(t) is employed in this analysis.

Figures 12 and 13 illustrate the difference in cross-correlations (Δr) between the model with higher gain and the model with lower gain. Figure 11 shows the 0.05 quantile of Δr, while

Figure 13 shows the mean values of Δr. The results indicate that there is generally no positive effect of higher gain at higher sound levels, where compression is more dominant. This is particularly evident in Fig.13, where the differences in cross-correlation are mostly close to zero or negative at higher sound levels, suggesting that higher gain does not improve the model's performance under babble noise conditions at these levels. Therefore, the presented simulation results rather question the positive role of compression on speech in noise perception and disagree with the results of listening experiments shown in Fig. 7.

In contrast, the simulations show that the best results are achieved at the lowest intensities for the model with higher gain. At the same time, the lower gain yields 0.05 quantile predictions, which qualitatively agree with listening experiments in that the correlation increases and reaches its maximum at intensities near 60 dB SPL.

The graphs illustrate the discrepancy in model outcomes at varying gain settings. The results are partly puzzling, showing the opposite trend we see in the experiments, i.e., a decline of cross-correlation with increasing levels. In addition, the gain seems to be rather detrimental at the intensities where the cochlear compression should dominate (near 60 dB SPL) because the best agreement with experimental data in the growth of cross-correlation up to about 60 dB SPL was found for the model with lower gain. Nevertheless, the predicted outcomes may diverge from the hearing experiment results due to differences in the model. It is important to consider these limitations when interpreting the results.

4 Summary

The ability to discern spoken phrases against a backdrop of competing babble noise is enhanced progressively with sound levels up to approximately 60 dB SPL. This trend of heightened auditory sentence and consonant recognition at escalated volumes aligns with the findings from Dubno and colleagues [3]. Their research further revealed a diminishing recognition rate for sounds above 60 dB SPL. It should be noted that in our observational data, the sound intensity was not extended beyond 70 dB SPL during the auditory tests.

1. The listener's speech perception is improved as the stimulus level increases, although the signal-to-babble noise ratio is constant (-5 dB).

2. To study whether cochlear compression is responsible for recognition improvement, we used an advanced auditory model incorporating the Dual Resonance Non-Linear (DRNL) filterbank.

   The effect of babble noise on the model output was quantified by cross-correlation between the model response to speech + babble noise and the model response to speech only. Two model variants (higher gain with a = 2e4 and lower gain with a = 2e3) and two methods (0.05 quantile and mean value) were used. The results showed small improvements in cross-correlation for the model with higher gain at sound levels from 30 to 40 dB SPL and larger improvements for the model with lower gain across a broader range of levels, peaking around 50-60 dB SPL, which aligns with listening experiments but only for the 0.05 quantile method. At higher intensities, the cross-correlation decreased. This variability, particularly the agreement with listening experiments only at the 0.05 quantile, is puzzling and suggests that cross-correlation may not be a comprehensive measure to quantify model performance under babble noise conditions. This discrepancy and the limited range of sound levels over which improvements are observed should be carefully considered in the summary and conclusion.

   Additionally, the model demonstrated higher cross-correlation for the cochlear model with higher gain at the lowest intensities.

Overall, the experimental results indicate that while the DRNL filterbank model has the potential to simulate certain aspects of human auditory processing, the results are not entirely conclusive. The observed improvements in cross-correlation do not consistently align with experimental data, suggesting that cochlear compression alone may not fully explain the mechanisms of speech perception in noisy environments. It is possible that factors beyond the cochlear periphery or additional mechanisms contribute to the results observed in listening experiments. There are many possible reasons for these results, and further research is needed to explore these additional factors and refine the model to better simulate human auditory processing.

5 References

5.1 List of sources cited in the paper

[1]. **L. Robles and M. A. Ruggero**, "Mechanics of the mammalian cochlea," *Physiological Reviews*, vol. 81, no. 3, pp. 1305-1352, Jul. 2001, doi: 10.1152/physrev.2001.81.3.1305.

[2]. **S. Duke and D. Sensale-Rodriguez**, "Fluid Mechanics and the Active Process of the Inner Ear," *arXiv*, vol. 1408.2085, pp. 1-12, Aug. 2014. [Online]. Available: http://arxiv.org/pdf/1408.2085.

[3]. **J. R. Dubno, C. G. Leek, and M. C. Heinz**, "Level-Dependent Changes in Perception of Speech Envelope Cues," *Journal of the Association for Research in Otolaryngology*, vol. 13, no. 6, pp. 797–807, Dec. 2012. [Online]. Available: https://link.springer.com/article/10.1007/s10162-012-0343-2.

[4]. **V. Vencovský and Z. Bureš**, "Level-Dependent Responses to Speech in Noise derived from a Nonlinear Cochlear Model," presented at the Midwinter ARO meeting, 2023.

[5]. **E. A. Lopez-Poveda and R. Meddis**, "A human nonlinear cochlear filterbank," *Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 3107-3118, 2001, doi: 10.1121/1.1416197.

[6]. **C. J. Plack and R. P. Carlyon**, "Temporal integration of level information in auditory perception," *Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1052-1063, Aug. 2001, doi: 10.1121/1.1383292.

[7]. **MAP (MATLAB Auditory Periphery)**. Available: https://github.com/rmeddis/MAP.

[8]. **R. Patterson, K. Robinson, J. Holdsworth, C. Zhang, and M. Allerhand**, "Complex sounds and auditory imaging," in *Auditory Physiology and Perception*, Y. Cazals, K. Horner, and L. Demany, Eds., Oxford, England: Pergamon, 1992, pp. 429-443.

[9]. **L. H. Carney**, "A model for the responses of low-frequency auditory-nerve fibers in cat," *Journal of the Acoustical Society of America*, vol. 93, pp. 401–417, 1993.

[10]. **T. Dau, D. Püschel, and A. Kohlrausch**, "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," *Journal of the Association for Research in Otolaryngology*, vol. 4, pp. 478–494, 1993.

[11]. **R. Huber and B. Kollmeier**, "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1902–1911, 2006.

[12]. **E. Zwicker and E. Terhardt**, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," Journal of the Acoustical Society of America, vol. 82, no. 1, pp. 151-156, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0378595587900505?via%3Dihub