A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications

Deepak Baby, Arthur Van Den Broucke, Sarah Verhulst

Dept. of Information Technology, Ghent University, 9000 Ghent, Belgium

* s.verhulst@ugent.be; deepak.baby@idiap.ch

Abstract

Auditory models are commonly used as feature extractors for automatic speech recognition systems or as front-ends for robotics, machine-hearing and hearing-aid applications. While over the years, auditory models have progressed to capture the biophysical and nonlinear properties of human hearing in great detail, these biophysical models are slow to compute and consequently not used in real-time applications. To enable an uptake, we present a hybrid approach where convolutional neural networks are combined with computational neuroscience to yield a real-time end-to-end model for human cochlear mechanics and level-dependent cochlear filter tuning (CoNNear). The CoNNear model was trained on acoustic speech material, but its performance and applicability evaluated using (unseen) sound stimuli common in cochlear mechanics research. The CoNNear model accurately simulates human frequency selectivity and its dependence on sound intensity, which is essential for our hallmark robust speech intelligibility performance, even at negative speech-to-background noise ratios. Because its architecture is based on real-time, parallel and differentiabt computatons, the CoNNear model has the power to leverage real-time auditory applications towards human performance and can inspire the next generation of speech recognition, robotics and hearing-aid systems.

1 Introduction

The human cochlea is an active, nonlinear system which transforms sound impinging on the eardrum via the middle-ear bones to cochlear travelling waves of basilar-membrane (BM) displacement and velocity [1]. Cochlear mechanics and travelling waves are responsible for the hallmark feature of human hearing, namely the level-dependent frequency selectivity [2-5] which results from a cascade of cochlear mechanical filters with center frequencies between 20 kHz and 40 Hz from the cochlear base to apex [6].

Modelling cochlear mechanics and travelling waves has been an active field of research because computational methods can help characterize the mechanisms underlying normal or impaired hearing and thereby improve hearing diagnostics [7,8] and treatment [9,10], or inspire machine-hearing applications [11,12]. One popular model approach is to represent the cochlea as a transmission line (TL) which discretizes the space along the BM and describes each section as a system of ordinary differential equations which approximates the biophysical cochlear filter characteristics (Fig.1; state-of-the-art model) [13,19]. Analytical TL models represent the cochlea as a cascaded system, i.e., the response of one section depends on the responses of
all previous sections, which makes them computationally expensive, as the filter operations in
the different sections cannot be computed in parallel. This computational complexity is even
greater when nonlinearities or feedback pathways are included to faithfully approximate cochlear
mechanics [15,20].

This computational complexity is the main reason why real-time applications for hearing-aid
[21], robotics [22] and automatic speech recognition applications do not adopt cochlear travelling
wave models in their pre-processing. Instead, they use computationally-fast approximations of
auditory filtering and need to compromise on key auditory features in that process. A common
simplification implements the cochlear filters as a parallel, rather than cascaded, filterbank [23,24].
However, this architecture fails to capture the natural longitudinal coupling properties of the BM
associated with cochlear frequency glides [25] and the generation of otoacoustic emissions [26].
Another popular model, the gammatone filterbank model [27], does not include the stimulus-level
dependence of cochlear filtering associated with the role of cochlear outer-hair-cells. And lastly,
there is a range of models which simulate the level-dependence of cochlear filtering, but which
fail to match the performance of TL models [28]: they either simulate the longitudinal cochlear
coupling locally within individual filters of the uncoupled filterbank [29] or introduce distortion
artifacts when combining an automatic-gain-control type of level-dependence with cascaded digital
filters [30–32].

The computational complexity of biophysically realistic cochlear models hence poses a design
constraint for the development of human-like machine-hearing applications, and motivated our
search for a model which matches performance of state-of-the-art analytical TL models while
offering real-time execution. Here, we investigate whether convolutional neural networks (CNNs)
can be used for this purpose as this type of neural network can deliver end-to-end waveform
prediction [33,34] with real-time properties [35], and is based on convolutions akin to the filtering
process associated with cochlear processing.

This paper details how CNNs can be best connected and trained to approximate the computations
performed by TL cochlear models [19,36,37], with a specific emphasis on simultaneously
capturing the tuning, nonlinear and longitudinal coupling characteristics of human cochlear
processing. The proposed model, namely CoNNear, converts speech stimuli into corresponding
BM displacements across 200 cochlear filters distributed along the length of the BM. Different
from TL models, the CoNNear architecture is based on parallel CPU computations which can be
sped up through GPU computing. Consequently, CoNNear can easily be integrated with real-time
auditory applications that use deep learning. The quality of the CoNNear predictions and the
generalizability of the method towards sound stimuli it did not see during training, is evaluated
on the basis of cochlear mechanical properties such as filter tuning estimates [38], nonlinear
distortion characteristics [39] and excitation patterns [40] using sound stimuli commonly adopted
in experimental cochlear mechanics studies.

2 Methods

2.1 CoNNear Architecture

Figure 1(a) illustrates the CoNNear model training and evaluation procedure. While training was
performed using TL-model simulated BM vibrations to speech [37], evaluation was performed on
the basis of predicting key cochlear mechanical properties using acoustic stimuli it did not see during training. The CoNNear model transforms an acoustic waveform sampled at 20 kHz to
N_{CF} cochlear BM displacement waveforms using several CNN layers. CoNNear maps a single
(a) CoNNear Training and Evaluation

![Diagram showing CoNNear training and evaluation procedure.](image)

Speech corpus (TIMIT, 70 dB SPL)

- Simulated BM vibration across \( N_{CF} \) cochlear sections
- Cochlear excitation patterns
- Filter tuning vs CF
- Filter tuning vs level
- Cochlear distortion products

Evaluation

- CoNNear parameters
- Minimize L1 loss

Decoder

Encoder

(b) CoNNear Architecture

![Diagram showing CoNNear encoder-decoder architecture.](image)

Fig 1. (a) Overview of the CoNNear model training and evaluation procedure. (b) CoNNear architecture with and without context. The final CoNNear model has an auto-encoder convolutional neural network architecture which is connected using strided convolutions and skip-connections to map audio input to 200 basilar-membrane vibration outputs in the time-domain. The final CoNNear model has four encoding and decoding layers and a tanh activation function between the layers.

The acoustic input to \( N_{CF} \) outputs, which is different from conventional CNN architectures which map the input to a single output. Every CNN layer is comprised of a set of filterbanks followed by a nonlinear operation \([41]\) and the CNN filter weights were trained on the basis of TL-simulated BM displacements.

Figure 1(b) depicts the CoNNear encoder-decoder architecture: an audio input of length \( L = 2048 \) is first processed by an encoder (comprised of four CNN layers) which encodes the audio signal into a condensed representation of size \( 128 \times k_4 \), after which the decoder layers map this representation onto \( L \times N_{CF} = 201 \) BM displacement waveforms corresponding to cochlear filter outputs with CFs spanning the human hearing range (0.1-12kHz). The chosen center frequencies span the human hearing range and are spaced according to the Greenwood place-frequency description of the human cochlea \([6]\).

The encoder CNN layers use strided convolutions, i.e. the filters are shifted by a time-step of two such that the temporal dimension is halved after every CNN layer. Thus, after \( N \) encoder CNN layers, the audio signal is encoded into a representation of size \( L/2^N \times k_N \), where \( k_N \) is the number...
of filters in the $N^{th}$ CNN layer. The decoder uses deconvolution, or transposed-convolutional layers, to double the temporal dimension after every layer. Additionally, the encoder contains $N$ deconvolution layers to re-obtain the original temporal dimension of the audio input ($L$). The number of filters in the final decoder CNN layer equals the number of cochlear sections $N_{CF}$ simulated in the cochlear TL model adopted for training the CoNNear parameters.

Temporal alignment, or phase information, of the audio input might be compromised due to the strided convolution operations in the encoder layers. Because preserving temporal information is important for speech perception [42], we used $U$-shaped skip connections to bypass the temporal audio information to the decoder layers. Skip-connection-based architectures have earlier been adopted for image-to-image translation [43] and speech enhancement applications [33,34] and offer several direct paths between the in- and outputs of CoNNear to maintain the original signal-phase information across the architecture. Skip-connections might also benefit the model when learning how to best combine nonlinearities of several CNN layers to simulate the nonlinear and the level-dependent properties of human cochlear processing.

CNNs expect a fixed length input (here $L = 2048$), but audio applications based on CoNNear should be able to deal with continuous audio inputs as well. In a general approach, the audio signal could be split into windows of 2048 samples, their corresponding BM displacements simulated, and concatenated. However, because CoNNear treats each input independently, concatenating the simulated outputs would result in discontinuities at the boundaries when there is no context information. To address this issue, we provided context by making the previous $L_l$ and following $L_r$ input samples also available to CoNNear when simulating BM displacement to an input of length $L_c$ (Fig.1b). Using context, the total input size becomes $L' = L_l + L + L_r$ with an output size of $L' \times N_{CF}$. A final cropping layer is added to crop out the context ($L_l=L_r=256$ samples) after the last CNN decoder layer. After CoNNear is trained, it can process audio-inputs of any duration. Note that the CoNNear model output units are BM displacement $y_{BM}$ in $[\mu m]$, whereas the TL-model outputs are in $[m]$. This scaling was necessary to enforce training of CoNNear with sufficiently high digital numbers. For training purposes, and visual comparison between the TL and CoNNear outputs, the $y_{BM}$ values of the TL model were multiplied by a factor of $10^6$ in all following figures and analysis.

2.2 Training the CoNNear model

The CoNNear model was trained using TL-model simulations to recordings from the TIMIT speech corpus [44] which contains 2310 phonetically balanced sentences with sufficient acoustic diversity for training. TIMIT recordings were upsampled to 100 kHz to solve the TL-model accurately [36] and the root-mean square (RMS) energy of every utterance was adjusted to 70 dB sound pressure level (SPL). BM displacements were simulated for 1000 cochlear sections with center frequencies (CFs) between 25 Hz and 20 kHz using a nonlinear time-domain TL model of the cochlea [37]. From the TL-model output representation (i.e., 1000 $y_{BM}$ waveforms sampled at 20 kHz), 201 uniformly distributed CFs between 100 Hz and 12 kHz were chosen to train CoNNear. Above 12 kHz, human hearing sensitivity becomes very poor [45], motivating our choice for the upper limit of considered CFs.

Speech material was epoched in windows of 2048 samples for model training and 256 samples were added before and after the input when considering the context model. CoNNear Model parameters were optimized to minimize the mean absolute error (dubbed L1 loss) between the predicted model outputs and the reference TL model outputs. A learning rate of 0.0001 was used with an Adam optimizer [46] and the entire framework was developed using the Keras machine
learning library \[47\] with a Tensorflow \[48\] back-end.

3 Evaluating the CoNNear model

Even though CoNNear might end up simulating the speech training dataset with a sufficiently low L1 loss, the quality of the model should be evaluated on its cochlear mechanical properties because CoNNear aims to model cochlear processing. That is, the architecture should also be evaluated on the L1 prediction error to acoustic stimuli used in classical cochlear mechanical studies. Thus, even though the CoNNear parameters were optimized for a speech training-set presented at 70 dB RMS level, CoNNear should also perform well on stimuli of different levels and frequencies adopted in experimental cochlear mechanics studies. This additional knowledge is used to determine the final model architecture and its hyperparameters. The following sections describe four cochlear mechanics evaluation metrics which were considered and which together form a complete description of cochlear processing. Also the details of the basic acoustic stimuli used for evaluation are described. Even though any speech fragment can be seen as a combination of basic elements such as impulses and pure tones of varying levels and frequencies, the basic acoustic stimuli associated with the cochlear mechanics evaluation can be considered as \textit{unseen} to the model, as they were not explicitly present in the training material. The evaluation stimuli were sampled at 20 kHz and had a duration of 102.4 ms (2048 samples) and 128 ms (2560 samples) for the CoNNear and context-CoNNear model, respectively. Stimulus levels were adjusted using the reference pressure of $p_0 = 2 \cdot 10^{-5}$ Pa.

3.1 Cochlear Excitation Patterns

Cochlear excitation patterns can be constructed from the RMS energy of the BM displacement or velocity at each measured CF in response to tonal stimuli of different levels. Cochlear excitation patterns show a characteristic half-octave basal-ward shift of their maxima as stimulus level increases \[40\]. Cochlear excitation patterns also reflect the nonlinear compressive growth of BM-responses with level observed when stimulating the cochlea with a pure-tone which has the same frequency as the CF of the measurement site in the cochlea \[3\]. Cochlear pure-tone transfer-functions and excitation patterns have in several studies been used to describe the level-dependence and tuning properties of cochlear mechanics \[2,3,40\]. We calculated excitation patterns for all 201 simulated BM displacement waveforms in response to pure tones of 0.5, 1 and 2 kHz frequencies and levels between 0 and 90 dB SPL using:

$$\text{tone}(t) = p_0 \cdot \sqrt{2} \cdot 10^{L/20} \cdot \sin(2\pi f_{\text{tone}}t), \quad (1)$$

where, $t$ corresponds to a time vector of 2048 samples, $L$ to the desired RMS level in dB SPL, and $f_{\text{tone}}$ to the stimulus frequencies. The pure-tones were multiplied with a Hanning-shaped 10-ms on- and offset ramp to ensure a gradual onset.

3.2 Cochlear filter tuning

A common approach to characterize auditory or cochlear filters is by means of the equivalent-rectangular bandwidth (ERB) or $Q_{\text{ERB}}$. The ERB describes the bandwidth of a rectangular filter which passes the same total power than the filter shape estimated from behavioral or cochlear tuning curve experiments \[49\], and presents a standardized way to characterize the tuning of the asymmetric auditory/cochlear filter shapes. The ERB has been used in the description of the
frequency and level-dependence of human cochlear filtering \[4,50,51\], whereas \( Q_{\text{ERB}} \) has been used to describe level-dependent cochlear filter characteristics from BM impulse response data \[19,52\].

To evaluate CoNNear, we calculated \( Q_{\text{ERB}} \) using:

\[
Q_{\text{ERB}} = \frac{\text{CF}}{\text{ERB}}.
\]

The ERB was determined from the power spectrum of a simulated BM time-domain response to an acoustic click stimulus using the following steps \[52\]: (i) compute the Fast Fourier Transform (FFT) of the BM displacement at the considered CF, (ii) compute the area under the power spectrum, and (iii) divide the area by the CF. The frequency- and level-dependence of CoNNear predicted cochlear filters were compared against TL-model predictions and experimental \( Q_{\text{ERB}} \) values reported for humans \[4\].

Acoustic stimuli were condensation clicks of 100-µs duration and were scaled to the desired peak-equivalent sound pressure level (dB peSPL), to yield a peak-to-peak click amplitude which matched that of a pure-tone with the same dB SPL level (\( L \)):

\[
\text{click}(t) = 2\sqrt{2} \cdot p_0 \cdot 10^{L/20} \cdot x(t) \quad \text{with} \quad x(t) = \begin{cases} 
1 & \text{for} \ t \leq 100 \mu s \\
0 & \text{for} \ t > 100 \mu s
\end{cases}
\]

\[3\]

### 3.3 Cochlear Dispersion

Click stimuli can also be used to characterize the cochlear dispersion properties, as their short duration allows for an easy separation of the cochlear response from the evoking stimulus. At the same time, the broad frequency spectrum of the click excites a large portion of the BM. Cochlear dispersion stems from the longitudinal-coupling and tuning properties of BM mechanics \[53\] and is observed through later click response onsets when BM responses are measured from CFs from base to apex. In humans, the CF-dependent cochlear dispersion delay mounts up to 10-12 ms for stimulus frequencies associated with apical processing \[54\]. Here, we use clicks of various sound intensities to evaluate whether CoNNear produced cochlear dispersion and BM click responses in line with predictions from the TL-model.

### 3.4 Distortion-product otoacoustic emissions (DPOAEs)

DPOAEs can be recorded in the ear-canal using a sensitive microphone and are evoked by two pure-tones with frequencies \( f_1 \) and \( f_2 \) and SPLs of \( L_1 \) and \( L_2 \), respectively. For pure tones with frequency ratios between 1.1 and 1.3 \[55\], local nonlinear cochlear interactions generate distortion products, which can be seen in the ear-canal recordings as frequency components which were not originally present in the stimulus. Their strength and shape depends on the properties of the compressive cochlear nonlinearity associated with the electro-mechanical properties of cochlear outer-hair-cells \[39\], and the most prominent DPOAEs appear at frequencies of \( 2f_2 - f_1 \) and \( 2f_1 - f_2 \). Even though CoNNear was not designed or trained to simulate DPs, they form an excellent evaluation metric, as realistically simulated DPOAE properties would demonstrate that CoNNear was able to capture even the epiphenomena associated with cochlear processing.

As a proxy measure for ear-canal recorded DPOAEs, we considered BM displacement at the highest simulated CF which, in the real ear, would drive the middle-ear and eardrum to yield the ear-canal pressure waveform in an OAE recording. We compared simulated DPs extracted from the FFT of the BM displacement response to simultaneously presented pure tones of \( f_1 = 2 \)
kHz and $f_2 = 1.2 \cdot f_1 = 2.4$ kHz with levels of $L_1 = 59.0$ and $L_2 = 50.0$ dB SPL according to the commonly adopted experimental scissors paradigm: $L_1 = 39 + 0.4L_2$ [50].

4 Results

4.1 Determining the CoNNear hyperparameters

An important aspect of this work relates to determining the optimal CNN architecture and associated hyperparameter values. Figure 1(b) shows the final CoNNear layout which resulted from an iterative principled fine-tuning approach in which several hyperparameters were adjusted to achieve the optimal model architecture taking into account: (i) the L1 loss on speech material, (ii) the desired frequency- and level-dependent cochlear filter tuning characteristics, and (iii), the computational load to allow real-time execution. Table 1 details the fixed and variable hyperparameters which were taken into consideration to determine the optimal CoNNear architecture.

Figure 2 shows simulated QERB functions across CF and visualizes how different activation functions (PReLU or tanh) and layer depths (4,6,8) affected the simulated QERBs. Aside from the reference experimental human QERB curve for low stimulus levels [51], Fig.2(a) depicts simulated reference QERB curves from the TL-model (red) with overlaid CoNNear-model curves. Whereas the PReLU activation function was unable to capture the level-dependence of cochlear filter tuning (i.e. the 40 and 70 dB QERB curves overlapped), the tanh nonlinearity was able to capture both the simulated level-dependent QERB tuning and the frequency-dependence of human QERBs using an 8-layer model. Figure2(b) shows that CoNNear captures the frequency-dependence of the QERB function better when the layer depth is increased from 4 to 8. Models with 4 and 6 layers tended to underestimate the overall QERB and performed worse for CFs below 1 kHz where the ERBs were narrower and the target BM impulse responses were longer than at high CFs. Further increasing the number of layers did not yield a substantial visual improvement over the 8 layer CoNNear model and would further increase the needed computational resources. Based on the QERB simulations and model performance parameters listed in tables 2 and 3, we chose a final CoNNear architecture with 8 layers and a tanh activation function.

4.2 Evaluation of CoNNear as a model for human cochlear signal processing

Since the CoNNear models were trained using speech, we first compare reference TL-model simulations to a 2048-sample-long speech segment presented at 70 dB SPL with the final 8-layer, tanh CoNNear model (Fig. 3). Comparing the context (d) and without context (c) architectures, it is clear that CoNNear cannot capture the BM displacement pattern at the stimulus onset when no context is provided.

Next, the trained CoNNear models were evaluated based on the cochlear mechanics metrics described in Section 3. The level-dependence of simulated cochlear filter tuning was already shown in Fig2 but filter tuning and compression properties are also evident from the pure-tone excitation patterns. Figure 4 shows that the final model architecture (d) outperforms both the architectures without context (c) and with the PReLU activation function (b). It is also important to notice that even though both the PReLU and tanh activation functions can code input negative deflections, the tanh activation function was the only nonlinearity which was able to capture the nonlinear compression characteristics of cochlear processing. This difference is
Table 1. Parameter selection of the CoNNear architecture

| Fixed parameters                  | Summary                                                                                                                                 |
|-----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| Window length                     | The input length was set to 2048 samples ($\approx$ 100 ms) and $2048 + 2 \cdot 256 = 2560$ samples for the CoNNear-context model.          |
| CNN parameters                    | The filters in the CNN use a stride of 2 for dimensionality reduction. All CNN layers have fixed filter length of 64 with 128 filters per layer. The chosen filter length formed a trade-off between the total number of model parameters (and required computations) and performance on the cochlear mechanical tasks. |
| Number of cochlear output channels| CoNNear was trained to generate BM displacements of 201 cochlear sections spaced between 100 Hz and 12 kHz.                                |

| Hyperparameters                  | Summary                                                                                                                                 |
|----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| Activation function              | The shape of the activation function, or nonlinearity, is crucial to enable CoNNear to learn the cochlear compression properties associated with the role of the outer hair cells (OHCs). To mimic the original shape of the OHC input/output function \cite{57}, the activation function should cross the x-axis. Standard activation functions (e.g., sigmoid, ReLUs) do not cross the x-axis, and may end up limiting the CoNNear model from learning realistic level-dependent BM impulse response properties. We hence opted to compare the performance of architectures with activation functions that crossed the x-axis, i.e., the parametric rectified linear unit (PReLU) and hyperbolic-tangent (tanh) nonlinearities. |
| Number of layers                 | Increasing the number of layers increases the representational capacity of the model, but evidently also increases the computational complexity. We investigated CoNNear architectures with layer depths of 4, 6 and 8 to determine the minimum model size required to reliably capture the desired cochlear mechanical properties. |
Fig 2. Determining the optimal CoNNear hyperparameters. Comparisons were performed using the CoNNear model in (a), and the CoNNear-context models in (b) and (c). (a) Comparing architectures using the PReLU or tanh activation function shows that the PReLU nonlinearity fails to capture the level-dependent cochlear filter tuning (i.e. similar Q_{ERB} function for stimulus levels of 40 and 70 dB peSPL). (b) As the number of layers in the CoNNear model was increased from 4 to 8, simulations of Q_{ERB} across CF improved. (c) Adding context improved the CoNNear predictions by showing less Q_{ERB} fluctuations across CF.
Table 2. Activation function comparison. Number of parameters, loss term, time per epoch and MSE ($\times 10^{-4}$) of the CoNNear predictions on the training set (L1-loss) and listed unseen acoustic stimuli. The details of the CoNNear architectures are given on the left. For each evaluated category, the best performing architecture is highlighted in bold font.

| Model                  | #Param      | Time/Epoch [h] | L1-loss  | Click  | 1kHz     | Word  |
|------------------------|-------------|----------------|----------|--------|----------|-------|
| PReLU/8 lay./no context| 11,982,464  | 1.89           | 0.0267   | 3.75   | 182.54   | 39.67 |
| tanh/8 lay./no context | 11,507,328  | 1.84           | 0.0123   | 1.00   | 2.53     | 12.72 |
| tanh/8 lay./context    | 11,689,984  | 2.18           | 0.0087   | 0.80   | 1.71     | 2.04  |

Table 3. CNN layer depth comparison. Number of parameters, loss term, time per epoch and MSE ($\times 10^{-4}$) of the CoNNear predictions on the training set (L1-loss) and listed unseen acoustic stimuli. The details of the CoNNear architectures are given on the left. For each evaluated category, the best performing architecture is highlighted in bold font.

| Model                  | #Param      | Time/Epoch [h] | L1-loss  | Click  | 1kHz     | Word  |
|------------------------|-------------|----------------|----------|--------|----------|-------|
| tanh/4 lay./context    | 5,398,528   | 2.13           | 0.0187   | 2.45   | 7.63     | 8.79  |
| tanh/6 lay./context    | 8,544,256   | 2.15           | 0.0102   | 1.04   | 2.32     | 3.00  |
| tanh/8 lay./context    | 11,689,984  | 2.18           | 0.0087   | 0.80   | 1.71     | 2.04  |

observed when comparing the peaks of the excitation pattern for different stimulus levels. Whereas the CoNNear-PReLU excitation pattern maxima increased linearly with stimulus level, both the TL-model and the tanh-CoNNear excitation patterns showed compressive growth of the pattern maxima. It is hence essential to consider the shape of the activation function (linear growth in PReLU vs compressive growth in tanh) when the reference system is composed of level-dependent nonlinearities. In cochlear mechanics, the properties of the outer-hair-cells are responsible for the nonlinear and compressive growth of BM vibration with stimulus level. Correspondingly, we found that an activation function with compressive growth was best able to capture the properties of the analytical TL-model of cochlear mechanics. The final CoNNear architecture with context (d) captures the reference excitation patterns for different pure-tone frequencies and levels faithfully even though small excitation pattern fluctuations were observed for the 1 and 2 kHz patterns for CFs below the stimulus frequency. This noise had values which were approximately 30 dB below the excitation pattern maxima and are hence not expected to impact the sound-driven response of CoNNear to complex stimuli (e.g. such as speech) in a meaningful way. This latter statement is backed up by the CoNNear speech simulations in Fig.3 which do not show visible noise.

Figure 5 depicts the cochlear dispersion characteristics of the TL model and trained CoNNear models (8 layers, tanh). CoNNear was able to capture the characteristic 12-ms BM vibration onset delay from basal (high CF, low channel numbers) to apical (low CF, high channel numbers) cochlear sections. Cochlear dispersion is a property which arises through the biophysical properties of the BM (i.e. coupled membrane which has a varying stiffness and damping gradient), and the CoNNear architecture was able to capture this phenomenon. Adding context did not improve the simulations.

Lastly, we checked whether CoNNear was able to simulate cochlear distortion products which travel along the BM to generate an ear-canal pressure waveform in the real ear (distortion product...
Fig 3. Simulated basilar membrane displacements for a 2048-sample-long speech stimulus. The stimulus waveform is depicted in panel (a) and panels (b)-(d) depict the instantaneous BM displacement intensity (darker colors = higher intensities) of the simulated TL-model outputs (b) and two CoNNear architecture outputs: without (c) and with (d) context. The N<sub>CF</sub>=201 considered output channels are labeled per channel number: channel 1 corresponds to a CF of 100 Hz and channel 201 to a CF of 12 kHz.

otoacoustic emission). Figure 6 compares the TL-model simulations with those of different CoNNear models (b)-(d). Two factors play a role in determining the optimal CoNNear architecture (d): the shape of the activation function, and the addition of context. Distortion product otoacoustic emission frequencies are visible from the figure as spectral components which are different from the stimulus primaries of 1.2 and 2.4 kHz. The strongest DP component in humans occurs at 2f<sub>1</sub> - f<sub>2</sub> = 1.6 kHz, and the level of this DP is best captured using the tanh activation function. Secondly, adding context removed the high-frequency distortions which were visible in panels (b) and (c). Again, the activation function with a shape resembling that of the cochlear nonlinearity most closely, yielded the best result.

The evaluations of the CoNNear performance in Figs. 2, 5, 4 and 6 demonstrate that the 8-layer, tanh, CoNNear model with context performed best on four crucial aspects of human cochlear mechanics. The stimuli used for evaluation were not seen during training, to allow for a fair evaluation. Despite the training on a limited speech corpus presented at a single RMS level, CoNNear learned cochlear processing features across level and frequency to simulate outputs which matched those of biophysically-realistic analytical models of human cochlear processing. CoNNear is hence applicable to a whole range of audio applications.
Fig 4. **Comparing cochlear excitation patterns.** Simulated RMS levels of BM displacement across CF for tone stimuli presented at SPLs between 0 and 90 dB SPL. From top to bottom, the stimulus tone frequencies were 500Hz, 1 kHz and 2 kHz, respectively.
Fig 5. Comparing cochlear dispersion properties. Panel (a) shows the stimulus pressure, while panels (b)-(d) show the instantaneous $y_{bm}$ intensities for the considered CFs (channel numbers, CS) between 100 Hz (channel 201) and 12 kHz (channel 1). The colorscale is the same in all figure panels, and ranges between -15 $\mu$m (more blue) and 15 $\mu$m (more red).

Fig 6. Comparing simulated DPOAEs. The frequency response of the 12-kHz CF channel is evaluated as a proxy for the otoacoustic emissions recorded in the ear-canal. Frequency responses of model simulations are shown in response to two pure tones of $f_1, 2$ of 2.2 and 2.4 kHz. The most pronounced distortion product in humans occurs at $2f_1 - f_2$ (1.6 kHz).

4.3 CoNNear as a real-time model for audio applications

Aside from its realistic cochlear mechanical properties, CoNNear is able to operate in real-time. Real-time is commonly defined as a computation duration less than 10 ms for audio applications (below this limit no noticeable delay is perceived). Table 4 summarizes the time it takes to compute the final CoNNear-context model for a stimulus window of 1048 samples on CPU or GPU architectures. On a CPU, the CoNNear model outperforms the TL-model by a factor of 129 and on a GPU, CoNNear is 2142 times faster. Additionally, the GPU computations show that the final, trained, 8-layer, tanh CoNNear-context model has a latency of 7.27 ms, and hence reaches real-time audio processing performance.

5 Discussion

This paper detailed how a hybrid, deep-neural-net and analytical approach can be used to develop a real-time executable CoNNear model of human cochlear processing, with performance matching that of human cochlear processing faithfully. For the first time, real-time and biophysically-realistic are not compromised upon, but combined into a single auditory model to inspire a new generation of human-like robotic, speech recognition and machine hearing applications. Prior work has shown clear benefits of using biophysically-realistic auditory models as front-ends for auditory applications:
### Table 4. Model calculation speed
Comparison of the time required to calculate a TL and CoNNear model window of 1048 samples on a CPU (Apple MacBook Air, 1.8 GHz Dual-Core processor) and a GPU (NVIDIA GTX1080). The calculation time for the first window is considered separately for the GPU computations since this window also includes the weight initialization. For each evaluated category, the best performing architecture is highlighted with a bold font.

| Model                | #Param       | CPU (s/window) | GPU 1st window (s) | GPU (ms/window) |
|----------------------|--------------|----------------|--------------------|-----------------|
| PReLU/8 lay./no context | 11,982,464  | 0.222          | 1.432              | 7.70            |
| tanh/8 lay./no context    | 11,507,328  | **0.195**      | 1.390              | 7.59            |
| tanh/8 lay./context       | 11,689,984  | 0.236          | **1.257**           | **7.27**        |
| Transmission Line N/A    | N/A          | 25.16          | N/A                | 16918           |

e.g. for capturing cochlear compression, [7,11,28], speech enhancement at negative signal-to-noise ratio’s [12], realistic sound perception predictions [58] and for simulating the generator-sources of human auditory brainstem responses [37,59]. Hence, the CoNNear model can improve performance in application areas which, so far, have refrained from the using slow-to-compute biophysical auditory models. Not only can CoNNear operate on running audio-input with a latency below 7.5 ms, it offers a differentiable solution which can be used in closed-loop systems for auditory feature enhancement or augmented hearing.

With the rise of neural-network (NN) based methods, computational neuroscience has seen an opportunity to map audio or auditory brain signals directly to sound perception [60–62] and to develop computationally efficient methods to compute large-scale differential-equation-based neuronal networks [63]. These developments are transformative as they can unravel the functional role of hard-to-probe brain areas in perception and yield computationally-fast neuromorphic applications. Key to these breakthroughs is the hybrid approach in which knowledge from neuroscience is combined with that of NN-architectures [64]. While the possibilities of NN approaches are numerous when large amounts of training data are present, this is rarely the case for biological systems and human-extracted data. It hence remains challenging to develop models of biophysical systems which are generalizable to a broad range of unseen conditions or stimuli.

Our work presents a solution to this problem for cochlear processing by constraining the CoNNear architecture and its hyperparameters on the basis of a state-of-the-art TL cochlear model. Our general approach takes the following steps: (i) first, derive an analytical description of the biophysical system on the basis of available experimental data. (ii) Use the analytical model to generate a training dataset to a representative set of sensory stimuli. This training data-set is then used to determine the NN-model architecture and constrain its hyperparameters. Lastly, (iii) as the NN-architecture is trained to match outcomes of the analytical model to a broad range of sensory input features, it is maximally generalizable to unseen inputs. We demonstrated the generalizability of our CoNNear predictions by faithfully predicting key cochlear mechanics features to sounds unseen during training.

Our proposed method is by no means limited to NN-based models of cochlear processing, but can be applied to other nonlinear and/or coupled biophysical models of sensory and biophysical...
systems. Accurate analytical descriptions of cochlear processing have evolved over the years based on available experimental data from human and animal cochlea, and will continue to evolve. It is straightforward to train CoNNear to an updated/improved analytical model in step (i), as well as to include different or additional training data in (ii) to further optimize its prediction performance.

6 Conclusion

We presented a hybrid method which uniquely combines expert knowledge from the fields of computational auditory neuroscience and machine-learning based audio processing to develop a CoNNear model of human cochlear processing. CoNNear presents an architecture with differentiable equations and operates in real time (< 7.5 ms delay) while offering a speed-up factor of 2000 compared to state-of-the-art biophysically realistic models of cochlear processing. We have high hopes that the CoNNear framework will inspire the next generation of human-like machine hearing, augmented hearing and automatic speech recognition systems.

Acknowledgments

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

Competing interests

A patent application (EP19179210.0) was filed by UGent on the basis of the research presented in this manuscript. Inventors on the application are Sarah Verhulst, Deepak Baby and Arthur Van Den Broucke (status: search report received).

Data availability

The source code of the TL-model used for training is available via 10.5281/zenodo.3717431 or github/HearingTechnology/Verhulstetal2018Model, the TIMIT speech corpus used for training can be found online [44]. All figures in this paper can be reproduced using the trained CoNNear model.

Code availability

The code for the trained CoNNear model, including instructions of how to execute the code, will be made available as a repository on GitHub upon publication of this paper. A non-commercial, academic UGent license will apply.

References

1. von Békésy G. Travelling Waves as Frequency Analysers in the Cochlea. Nature. 1970;225(5239):1207–1209.
2. Narayan SS, Temchin AN, Recio A, Ruggero MA. Frequency tuning of basilar membrane and auditory nerve fibers in the same cochleae. Science. 1998;282(5395):1882–1884.

3. Robles L, Ruggero MA. Mechanics of the mammalian cochlea. Physiological reviews. 2001;81(3):1305–1352.

4. Shera CA, Guinan JJ, Oxenham AJ. Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. Proceedings of the National Academy of Sciences. 2002;99(5):3318–3323.

5. Oxenham AJ, Shera CA. Estimates of human cochlear tuning at low levels using forward and simultaneous masking. Journal of the Association for Research in Otolaryngology. 2003;4(4):541–554.

6. Greenwood DD. A cochlear frequency-position function for several species—29 years later. The Journal of the Acoustical Society of America. 1990;87(6):2592–2605.

7. Jepsen ML, Dau T. Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss. The Journal of the Acoustical Society of America. 2011;129(1):262–281.

8. Bondy J, Becker S, Bruce I, Trainor L, Haykin S. A novel signal-processing strategy for hearing-aid design: neurocompensation. Signal Processing. 2004;84(7):1239–1253.

9. Ewert SD, Kortlang S, Holmann V. A Model-based hearing aid: Psychoacoustics, models and algorithms. Proceedings of Meetings on Acoustics. 2013;19(1):050187.

10. Mondol S, Lee S. A Machine Learning Approach to Fitting Prescription for Hearing Aids. Electronics. 2019;8(7):736.

11. Lyon RF. Human and Machine Hearing: Extracting Meaning from Sound. Cambridge University Press; 2017.

12. Baby D, Van hamme H. Investigating modulation spectrogram features for deep neural network-based automatic speech recognition. In: Proc. INTERSPEECH. Dresden, Germany; 2015. p. 2479–2483.

13. de Boer E. Auditory physics. Physical principles in hearing theory. I. Physics reports. 1980;62(2):87–174.

14. Diependaal RJ, Duifhuis H, Hoogstraten HW, Viergever MA. Numerical methods for solving one-dimensional cochlear models in the time domain. The Journal of the Acoustical Society of America. 1987;82(5):1655–1666.

15. Zweig G. Finding the impedance of the organ of corti. The Journal of the Acoustical Society of America. 1991;89:1229–1254.

16. Talmadge CL, Tubis A, Wit HP, Long GR. Are spontaneous otoacoustic emissions generated by self-sustained cochlear oscillators? The Journal of the Acoustical Society of America. 1991;89(5):2391–2399.
17. Moleti A, Sisto R, Paglialonga A, Sibella F, Anteunis L, Parazzini M, et al. Transient evoked otoacoustic emission latency and estimates of cochlear tuning in preterm neonates. The Journal of the Acoustical Society of America. 2008;124(5):2984–2994.

18. Epp B, Verhey JL, Mauermann M. Modeling cochlear dynamics: Interrelation between cochlea mechanics and psychoacoustics. The Journal of the Acoustical Society of America. 2010;128:1870–1883.

19. Verhulst S, Dau T, Shera CA. Nonlinear time-domain cochlear model for transient stimulation and human otoacoustic emission. The Journal of the Acoustical Society of America. 2012;132:3842–3848.

20. Zweig G. Nonlinear cochlear mechanics. The Journal of the Acoustical Society of America. 2016;139(5):2561–2578.

21. Hohmann V. In: Havelock D, Kuwano S, Vorländers M, editors. Signal Processing in Hearing Aids. New York, NY: Springer New York; 2008. p. 205–212.

22. Rascon C, Meza I. Localization of sound sources in robotics: A review. Robotics and Autonomous Systems. 2017;96:184–210.

23. Mogran N, Bourlard H, Hermansky H. In: Automatic Speech Recognition: An Auditory Perspective. New York, NY: Springer New York; 2004. p. 309–338.

24. Patterson RD, Allerhand MH, Giguère C. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. The Journal of the Acoustical Society of America. 1995;98(4):1890–1894.

25. Shera CA. Frequency glides in click responses of the basilar membrane and auditory nerve: Their scaling behavior and origin in traveling-wave dispersion. The Journal of the Acoustical Society of America. 2001;109(5):2023–2034.

26. Shera CA, Guinan JJ. Mechanisms of mammalian otoacoustic emission. In: Active Processes and Otoacoustic Emissions in Hearing. Springer; 2008. p. 305–342.

27. Hohmann V. Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica united with Acustica. 2002;88(3):433–442.

28. Saremi A, Beutelmann R, Dietz M, Ashida G, Kretzberg J, Verhulst S. A comparative study of seven human cochlear filter models. The Journal of the Acoustical Society of America. 2016;140(3):1618–1634. doi:10.1121/1.4960486.

29. Lopez-Poveda EA, Meddis R. A human nonlinear cochlear filterbank. The Journal of the Acoustical Society of America. 2001;110(6):3107–3118.

30. Lyon RF. Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function. The Journal of the Acoustical Society of America. 2011;130(6):3893–3904.

31. Saremi A, Lyon RF. Quadratic distortion in a nonlinear cascade model of the human cochlea. The Journal of the Acoustical Society of America. 2018;143(5):EL418–EL424.
32. Altoè A, Charaziak KK, Shera CA. Dynamics of cochlear nonlinearity: Automatic gain control or instantaneous damping? The Journal of the Acoustical Society of America. 2017;142(6):3510–3519.

33. Baby D, Verhulst S. SERGAN: Speech Enhancement using Relativistic Generative Adversarial Networks with Gradient Penalty. In: ”Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. Brighton, UK; 2019. p. 106–110.

34. Pascual S, Bonafonte A, Serrà J. SEGAN: Speech Enhancement Generative Adversarial Network. In: INTERSPEECH. ISCA; 2017. p. 3642–3646.

35. Drakopoulos F, Baby D, Verhulst S. Real-Time Audio Processing on a Raspberry Pi using Deep Neural Networks. In: 23rd International Congress on Acoustics (ICA). Aachen, Germany; 2019.

36. Altoè A, Pulkki V, Verhulst S. Transmission line cochlear models: Improved accuracy and efficiency. The Journal of the Acoustical Society of America. 2014;136(4):EL302–EL308.

37. Verhulst S, Altoè A, Vasilkov V. Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss. Hearing Research. 2018;360:55–75.

38. Oxenham AJ, Wojtczak M. In: Plack C, editor. Frequency selectivity and masking: Perception. Oxford University Press; 2010.

39. Robles L, Ruggiero MA, Rich NC. Two-tone distortion in the basilar membrane of the cochlea. Nature. 1991;349(6308):413.

40. Ren T. Longitudinal pattern of basilar membrane vibration in the sensitive cochlea. Proceedings of the National Academy of Sciences. 2002;99(26):17101–17106.

41. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.

42. Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BC. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proceedings of the National Academy of Sciences. 2006;103(49):18866–18869.

43. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-Image Translation with Conditional Adversarial Networks. In: IEEE-CVPR; 2017. p. 5967–5976.

44. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM; 1993.

45. Precise and Full-range Determination of Two-dimensional Equal Loudness Contours. Geneva, CH: International Organization for Standardization; 2003.

46. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR. 2014:abs/1412.6980.

47. Chollet F, et al.. Keras; 2015. https://keras.io.

48. Abadi M, Ágarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. https://www.tensorflow.org/.
49. Moore BC, Glasberg BR. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. The journal of the acoustical society of America. 1983;74(3):750–753.

50. Glasberg BR, Moore BC. Derivation of auditory filter shapes from notched-noise data. Hearing research. 1990;47(1-2):103–138.

51. Shera CA, Guinan JJ, Oxenham AJ. Otoacoustic estimation of cochlear tuning: validation in the chinchilla. Journal of the Association for Research in Otolaryngology. 2010;11(3):343–365.

52. Raufer S, Verhulst S. Otoacoustic emission estimates of human basilar membrane impulse response duration and cochlear filter tuning. Hearing research. 2016;342:150–160.

53. Ramamoorthy S, Zha DJ, Nuttall AL. The biophysical origin of traveling-wave dispersion in the cochlea. Biophysical journal. 2010;99(6):1687–1695.

54. Dau T, Wegner O, Mellert V, Kollmeier B. Auditory brainstem responses with optimized chirp signals compensating basilar-membrane dispersion. The Journal of the Acoustical Society of America. 2000;107(3):1530–1540.

55. Neely ST, Johnson TA, Kopun J, Dierking DM, Gorga MP. Distortion-product otoacoustic emission input/output characteristics in normal-hearing and hearing-impaired human ears. The Journal of the Acoustical Society of America. 2009;126(2):728–738.

56. Kummer P, Janssen T, Hulin P, Arnold W. Optimal L1–L2 primary tone level separation remains independent of test frequency in humans. Hearing Research. 2000;146(1):47–56.

57. Russell I, Cody A, Richardson G. The responses of inner and outer hair cells in the basal turn of the guinea-pig cochlea and in the mouse cochlea grown in vitro. Hearing research. 1986;22(1-3):199–216.

58. Verhulst S, Ernst F, Garrett M, Vasilkov V. Suprathreshold psychoacoustics and envelope-following response relations: Normal-hearing, synaptopathy and cochlear gain loss. Acta Acustica united with Acustica. 2018;104(5):800–803.

59. Verhulst S, Bharadwaj HM, Mehraei G, Shera CA, Shinn-Cunningham BG. Functional modeling of the human auditory brainstem response to broadband stimulation. The Journal of the Acoustical Society of America. 2015;138(3):1637–1659.

60. Kell AJ, Yamins DL, Shook EN, Norman-Haignere SV, McDermott JH. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron. 2018;98(3):630–644.

61. Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N. Towards reconstructing intelligible speech from the human auditory cortex. Scientific reports. 2019;9(1):1–12.

62. Kell AJ, McDermott JH. Deep neural network models of sensory systems: windows onto the role of task constraints. Current opinion in neurobiology. 2019;55:121–132.

63. Amsalem O, Eyal G, Rogozinski N, Gevaert M, Kumbhar P, Schürmann F, et al. An efficient analytical reduction of detailed nonlinear neuron models. Nature Communications. 2020;11(1):1–13.
64. Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. Nature neuroscience. 2019;22(11):1761–1770.