Psychoacoustic Phenomena Explained by Auditory Task Optimization

Mark R. Saddler (marksa@dtu.dk)
DTU Department of Health Technology
Kongens Lyngby 2800, Denmark

Torsten Dau (tdau@dtu.dk)
DTU Department of Health Technology
Kongens Lyngby 2800, Denmark

Josh H. McDermott (jhm@mit.edu)
MIT Department of Brain and Cognitive Sciences
Cambridge, Massachusetts 02139, USA

Abstract

Artificial neural networks optimized for ecological tasks have emerged as leading models of sensory systems. Models optimized separately for sound localization and recognition tasks account for a range of human auditory behaviors, but it has remained unclear whether a single model could account for behaviors in both types of tasks. We optimized a model to jointly localize and recognize sounds from simulated auditory nerve input. The resulting multi-task model reproduced a range of human speech recognition effects related to noise, reverberation, and spatial separation. We also trained linear classifiers to perform simple psychoacoustic tasks using the model's internal representations. The learned model features produced human-like patterns of psychoacoustic judgments. The results provide further evidence that many aspects of human hearing can be understood as optimized solutions to ecological tasks.

Keywords: auditory perception; psychophysics; deep neural network; multi-task learning

Introduction

Evolution and development drive biological perceptual systems towards optimal performance on tasks that are important for survival. Machine perceptual systems optimized under similar constraints as biological systems suggest that signatures of this optimization process are ubiquitous in human behavior (Kell & McDermott, 2019; Kanwisher, Khosla, & Dobs, 2023). In hearing, deep neural networks optimized for naturalistic speech (Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018), pitch (Saddler, Gonzalez, & McDermott, 2021), and localization (Francl & McDermott, 2022) tasks can account for many aspects of human perception (Saddler & McDermott, 2024). However, these prior models were separately optimized for each of these domains, unlike the human auditory system which readily performs multiple tasks.

To progress towards a more complete computational account of audition, we optimized a single model to localize and recognize speech, voices, and environmental sounds from simulated auditory nerve representations of naturalistic scenes (Fig. 1). Once optimized, we compared the model's speech recognition and spatial hearing to that of humans in different listening conditions (manipulating background noise, reverberation, and spatial separation between speech and noise sources).

To further probe similarities between human and model auditory processing, we measured psychoacoustic thresholds from the model by training linear classifiers to make binary judgments using the task-optimized features. We envision these classifiers as analogous to decision rules that human participants use to perform simple hearing tests using relatively fixed internal representations (which were plausibly optimized for ecological tasks over longer timescales). Here, we present findings from one psychoacoustic experiment measuring thresholds for detecting amplitude modulations (Viemeister, 1979; Dau, Kollmeier, & Kohlrausch, 1997).

The results provide a normative account for fundamental aspects of human hearing, suggesting phenomena like spatial release from masking (Plomp, 1976) and modulation frequency selectivity (Houtgast, 1989) can be understood as consequences of optimization for ecological tasks.

Methods

Model Architecture

Auditory Nerve Input Representation All sounds were processed with an auditory nerve model to simulate spiking responses of 32000 nerve fibers per ear. The model consisted of a gammatone filter bank, half-wave rectification, a lowpass filter, and sigmoidal rate-level functions to yield instantaneous spike rates. Arrays of spike counts sampled from these rates (50 frequency channels, 20000 timesteps at 10 kHz, 3 nerve fiber types per ear) served as input to a neural network.

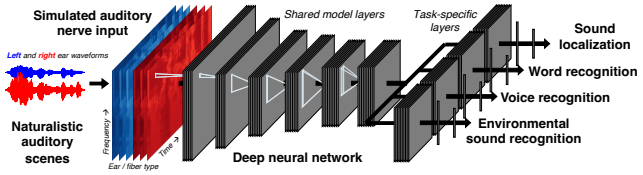

Figure 1: Model optimized for ecological auditory tasks.

Branched Convolutional Neural Network We used a feedforward neural network architecture with 7 convolutional blocks (each consisting of linear convolution, ReLU activation, Hanning pooling, and batch normalization), a 512-unit fully connected intermediate layer, and a task-dependent output layer (Saddler, Francl, et al., 2021). To enable good performance across all four auditory tasks included in the optimization objective, we introduced a branch point after the sixth convolutional block, with separate subsequent stages for each task (Fig. 1). Before the branch point, all model weights are shared between the tasks. After the branch point, weights are task-specific.

Model Optimization

Tasks and Dataset The training dataset included labels for four naturalistic auditory tasks. Stimuli were 2s (at 50 kHz) binaural auditory scenes spatialized with a virtual acoustic head and room simulator. Each scene consisted of a speech or natural sound target rendered at a single location with texture-like background noise rendered diffusely at multiple locations. The model's tasks were to localize the target (operationalized as a 504-way classification task) and make three types of recognition judgments (800-way word recognition and 500-way voice recognition tasks for speech targets; 50-way environmental sound classification for non-speech targets). The dataset consisted of 7.6 million scenes rendered in 2000 different rooms.

Training The model was optimized via stochastic gradient descent to minimize the summed softmax cross entropy losses from the four classification tasks. When a task was

undefined for a training example (e.g., word recognition for a non-speech stimulus), the task was excluded from the loss.

Model Evaluation

Speech Recognition in Noise We measured human and model word recognition scores at -3 dB SNR in 43 different background noise textures (Fig. 2A). Humans (47 online participants) and our model performed the same word recognition task with identical diotic stimuli. Talkers in the evaluation speech material were not seen during training.

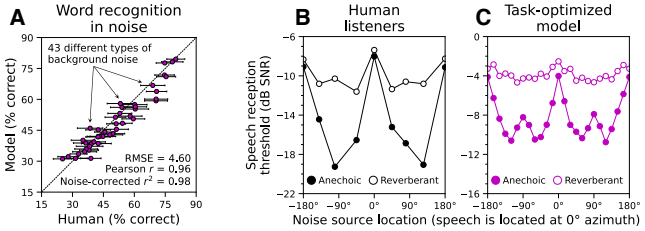

Figure 2: Model replicates effects of noise, reverberation, and spatial separation on human speech recognition.

Reverberation and Spatial Separation We simulated an experiment from Beutelmann and Brand (2006) by measuring model speech reception thresholds in an anechoic and a reverberant room as a function of the spatial separation between a target talker (always at 0 ° azimuth) and a spectrally-matched noise source (at varying azimuth) (Fig. 2B & C).

Amplitude Modulation Detection Participants heard stationary and amplitude-modulated noise bursts and identified which burst was modulated (Fig. 3A). Viemeister (1979) and Dau et al. (1997) measured human thresholds for detecting sinusoidal amplitude modulation as a function of modulation frequency and noise carrier bandwidth (Fig. 3B). To simulate these experiments on the model, we presented the model with stimuli consisting of two successive 1s noise bursts and trained linear classifiers on the ReLU activations (concatenated from all intermediate layers) to report whether the first or second burst was modulated. We trained a separate classifier for each noise bandwidth considered: 0 Hz (pure tone), 3, 31, 314, and 6000 Hz. In each case, the noise band was centered at 5000 Hz. Training and test stimuli for the linear classifiers were independent samples from the same distributions (varying uniformly in modulation frequency and depth).

Results

Speech Recognition Experiments

The 43 different noise conditions produced reliable variation in human word recognition scores (25% to 80% correct; split-half reliability = 0.968). When tested on the same stimuli and task, the model performed very similarly to humans, accounting for 98% of the explainable variance (Fig. 2A).

Speech reception thresholds measured from the model under different reverberation conditions and spatial arrangements were also human-like (Fig. 2B & C). Since the word recognition task and stimuli for the model experiment differed from the human experiment (Beutelmann & Brand, 2006), we compared effect sizes rather than absolute thresholds. In anechoic conditions (closed symbols), the model showed a large benefit of spatial separation between the talker and a noise source (up to 6.7 dB at 120 °). In reverberant conditions (open symbols), this benefit was considerably reduced in both humans and our model.

These results suggest the model relies on similar cues to humans when recognizing speech in adverse listening conditions.

Amplitude Modulation Detection Experiment

Changes in a sound's amplitude across different timescales are important cues for many aspects of hearing. Human thresholds (Fig. 3B) for detecting small amplitude modulations have been measured extensively (Dau et al., 1997). To investigate whether human-like modulation processing emerges in representations optimized for natural tasks, we measured modulation detection thresholds from different model variants as a function of modulation frequency and noise bandwidth.

Thresholds measured from the task-optimized model qualitatively and quantitatively resemble those of human listeners (Fig. 3C). By contrast, thresholds measured from an untrained model (Fig. 3D) or directly from the auditory nerve input (Fig. 3E) fail to reproduce the human pattern of behavior. These results suggest the task-optimized features of our model instantiate human-like computations for processing amplitude modulation.

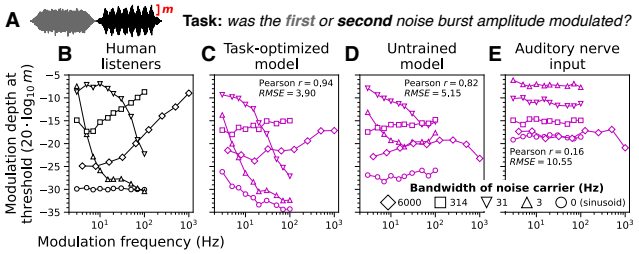

Figure 3: Modulation detection thresholds measured from the task-optimized model resemble those of humans.

Discussion

Our model innovates on prior work by jointly optimizing for sound localization and recognition tasks, which enabled investigation of the combined effects of noise, reverberation, and spatial separation on human speech recognition (Hawley, Litovsky, & Culling, 2004). The model accounted for several aspects of human binaural speech perception. Psychoacoustic thresholds measured from the task-optimized model's features provide new evidence for aligned internal representations between deep neural networks and human perceptual systems. Models that predict human behavior from simulated auditory nerve input in both complex environments and simple psychoacoustic tasks may be particularly suitable for investigating perceptual consequences of hearing loss.

Acknowledgments

This work was supported by National Institutes of Health grant number R01DC017970 to J.H.M. and the Oticon Centre of Excellence for Hearing and Speech Sciences.

References

Beutelmann, R., & Brand, T. (2006, July). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *120*(1), 331–342. doi: 10.1121/1.2202888

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997, November). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America*, *102*(5), 2892–2905. doi: 10.1121/1.420344

Francl, A., & McDermott, J. H. (2022, January). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, *6*(1), 111–133. (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41562-021-01244-z

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004, January). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, *115*(2), 833–843. doi: 10.1121/1.1639908

Houtgast, T. (1989, April). Frequency selectivity in amplitude-modulation detection. *The Journal of the Acoustical Society of America*, *85*(4), 1676–1680. doi: 10.1121/1.397956

Kanwisher, N., Khosla, M., & Dobs, K. (2023, March). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254. (Publisher: Elsevier) doi: 10.1016/j.tins.2022.12.008

Kell, A. J. E., & McDermott, J. H. (2019, April). Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, *55*, 121–132. doi: 10.1016/j.conb.2019.02.003

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018, May). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.e16. doi: 10.1016/j.neuron.2018.03.044

Plomp, R. (1976, February). Binaural and Monaural Speech Intelligibility of Connected Discourse in Reverberation as a Function of Azimuth of a Single Competing Sound Source (Speech or Noise). *Acta Acustica united with Acustica*, *34*(4), 200–211.

Saddler, M. R., Francl, A., Feather, J., Qian, K., Zhang, Y., & McDermott, J. H. (2021, August). Speech denoising with auditory models. In *Interspeech 2021* (pp. 2681–2685). ISCA. doi: 10.21437/Interspeech.2021-1973

Saddler, M. R., Gonzalez, R., & McDermott, J. H. (2021, December). Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications*, *12*(1), 7278. (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/s41467-021-27366-6

Saddler, M. R., & McDermott, J. H. (2024). Models optimized for real-world tasks reveal the necessity of precise temporal coding in hearing. *bioRxiv*. doi: 10.1101/2024.04.21.590435

Viemeister, N. F. (1979, November). Temporal modulation transfer functions based upon modulation thresholds. *The Journal of the Acoustical Society of America*, *66*(5), 1364–1380. doi: 10.1121/1.383531