From Sound to Source: Human and Model Recognition in Auditory Scenes

Sagarika Alavilli, Josh H. McDermott
{salavill, jhm}@mit.edu

Speech and Hearing Biosciences and Technology
Harvard, Cambridge MA 02318, USA

Department of Brain and Cognitive Sciences
McGovern Institute for Brain Research
MIT, 43 Vassar Street, Cambridge, MA 02139, USA

Abstract

Our ability to recognize sound sources in the world is critical to daily life, but it is not well documented or understood in computational terms. We developed large-scale behavioral benchmarks of human environmental sound recognition, built signal-computable models of sound recognition, and used the benchmarks to compare models to humans. The behavioral tests measured how sound recognition abilities varied with the source category, audio distortions of different types, and concurrent sound sources, all of which influenced recognition performance in humans. Artificial neural network models trained to classify sounds in multi-source scenes reached near-human accuracy and qualitatively matched human patterns of performance in many (but not all) conditions. By contrast, traditional models of the cochlea and auditory cortex produced worse matches to human performance. The results suggest that many aspects of human sound recognition emerge in systems optimized for the problem of real-world recognition. The benchmark results clarify the factors that constrain human recognition, setting the stage for future explorations of auditory scene perception involving salience, attention, and memory.

Keywords: environmental sounds; auditory scenes; computational models

Introduction

Environmental sound recognition refers to the process of identifying everyday sounds such as footsteps, rainfall, and animal calls. Such recognition abilities help us build a representation of the surrounding environment, but remain poorly understood in comparison to other aspects of audition. Recent progress in machine hearing has resulted in working models of environmental sound recognition (Hershey et al., 2017; Gong, Lai, Chung, & Glass, 2022), but it remains unclear whether such models can account for human abilities. We sought to build candidate models of human recognition by optimizing machine systems for source recognition and then compared them to human abilities measured with a large suite of experiments, an approach that has been fruitful for other aspects of audition (Saddler, Gonzalez, & McDermott, 2021; Francl & McDermott, 2022; Saddler & McDermott, 2024).

Methods

Training Datasets

Models were trained on auditory scenes generated from recorded natural sounds from the GISE-51 dataset (12,465 training sounds grouped into 51 source categories) (Yadav & Foster, 2021). Scenes ranged from 1 to 5 sources, drawn equiprobably from the 51 source categories. Each scene was 2 seconds long with each source a maximum of 1 second. We generated 1,500,000 training scenes and 100,000 validation scenes.

Model Architectures

All models had an initial 'cochleagram' stage, obtained from a filterbank intended to replicate the auditory periphery (McDermott & Simoncelli, 2011), and culminated in a linear classifier (a fully connected layer followed by 51 independent sigmoid units corresponding to the 51 output classes).

Cochleagram Model This model consisted of just the classifier operating on the cochleagram.

Spectrotemporal (ST) Model This model augmented the cochleagram with a bank of spectro-temporal filters developed by Chi, Ru, and Shamma (2005), intended to replicate primary auditory cortical processing. The classifier operated on the output of these filters.

Convolutional Neural Network (CNN) Model This model consisted of the cochleagram followed by a deep convolutional neural network architecture adapted from a model in Saddler and McDermott (2024). The network consisted of 6 blocks each containing layer normalization, convolution, ReLU nonlinearity, and pooling. This was then followed by a layer of dropout regularization, and then the classifier layer.

Behavioral Benchmarks

Participants completed two experiments. In the first they heard scenes of 1-5 sources, and judged whether a prompted sound category (e. g. "applause") was present in the scene. In the second, they heard recordings of single sound sources, to which various distortions had been applied (Table 1). Figure 2 shows examples of a few distortions applied to a few example sounds.

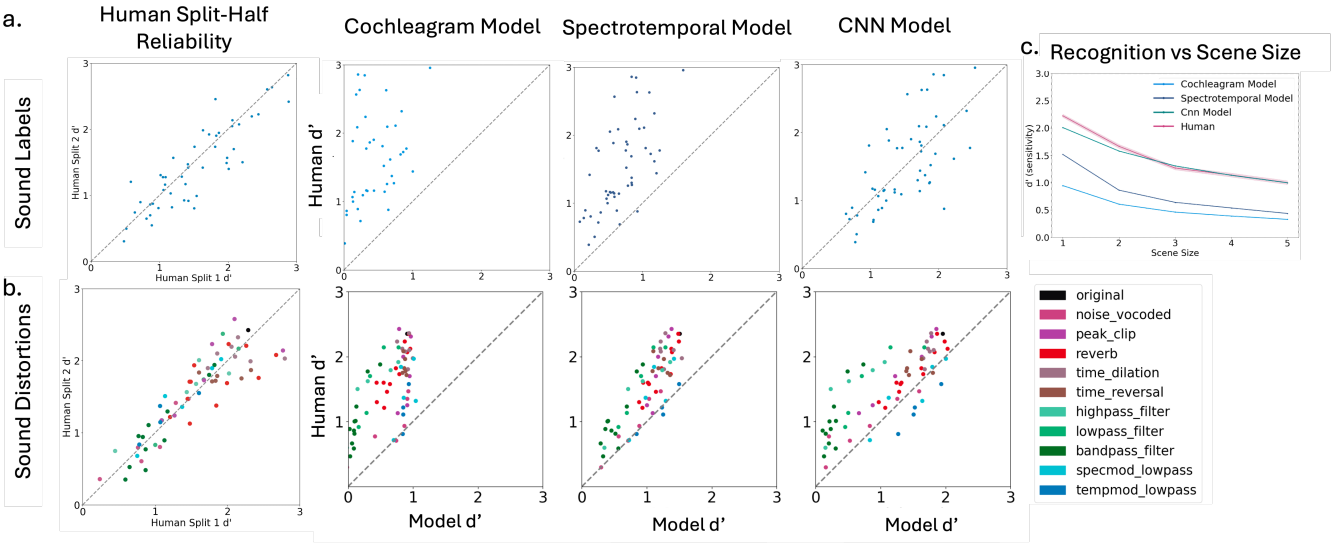

Figure 1: a. Performance for different sound categories. Each point corresponds to 1 of the 51 sound labels. Left panel plots human split-half reliability. Other panels plot each model vs. human. b. Performance for different distortions. Each point corresponds to a distortion condition, with the distortion type indicated by color. c. Performance vs. scene size for humans (red) and CNN (green), ST (dark blue), and cochleagram (light blue) models.

Table 1: Sound Distortion Conditions

Distortion Type	Conditions
Local Time Reversal	10, 20, 30, 40, 50, 100 ms
Time Dilation	0.5,0.75,0.875,1.125,1.25,1.5,2
Reverberation (DRR)	20, 50, 80 dB
Reverberation (RT60)	200, 400, 800, 1600 ms
Peak Clipping	0, 0.25, 0.5, 0.75, 0.9, 0.98
Noise Vocoding	1, 2, 4, 8, 16, 32 channels
Bandpass Filter (bandwidth)	5, 10, 20, 30, 40 semitones
Bandpass Filter (cent. freq.)	225,450,900,1800,3600,7200 Hz
Highpass Filter (cutoff)	400, 800, 1600, 3200, 6400 Hz
Lowpass Filter (cutoff)	400, 800, 1600, 3200, 6400 Hz
Spect. Mod. Lowpass Filt.	0.5, 1, 2, 4, 8 cycles/kHz
Temp. Mod. Lowpass Filt.	3, 6, 12, 24 Hz

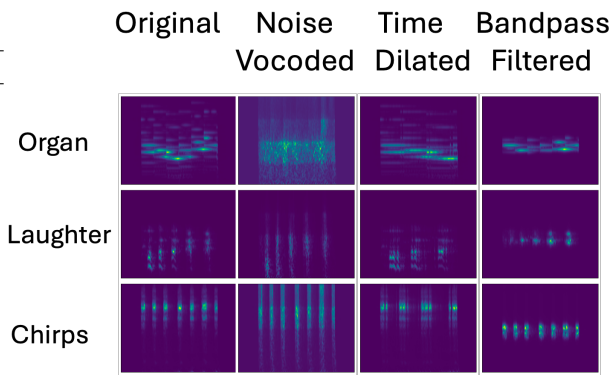

Figure 2: Spectrograms illustrating the impact of sound distortions (noise vocoding, time dilation, and bandpass filtering) on example sounds.

Table 2: Human-Model Correlation

	Split-Half Reliability	Cochleagram Model	ST Model	CNN Model
Labels	0.882	0.395	0.632	0.684
Distortions	0.899	0.595	0.787	0.791

Results

Sound Labels We quantified performance for particular sound classes as d', averaged across all scene sizes. Some source types were more recognizable to humans than others. All models captured some of the variation in performance, but the CNN most closely matched humans (Figure 1a).

Sound Distortions Humans showed reliable variation in performance across distortion types. The CNN again most closely matched human performance (Table 2). All models under-performed on the audio filtering conditions (Figure 1b).

Multi-source Scenes When analyzed as a function of scene size, recognition performance decreased as the number of sources increased, but remained well above chance (Figure 1c). The CNN model quantitatively matched human performance.

Conclusion

We developed a large-scale benchmark of human environmental sound recognition and compared humans to models

optimized for natural sound source recognition in auditory scenes. A CNN model trained on environmental sound recognition replicated the human dependence on scene size and much of the variation with source and distortion type. Simpler models based on standard cochlear or cortical processing stages alone did not replicate human behavior as well. However, even the CNN model underperformed humans in some conditions and did not explain all the reliable variance in human response patterns, perhaps because the training dataset was relatively small by modern standards. The results set the stage for future explorations of auditory scene perception involving salience, attention, and memory.

Acknowledgments

Work supported by National Institutes of Health grant number R01DC017970.

References

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906.

Francl, A., & McDermott, J. H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, *6*(1), 111–133.

Gong, Y., Lai, C.-I., Chung, Y.-A., & Glass, J. (2022). Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 10699–10709).

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... others (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131–135).

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*(1), 926-940.

Saddler, M. R., Gonzalez, R., & McDermott, J. H. (2021). Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications*, *12*(1), 7278.

Saddler, M. R., & McDermott, J. H. (2024). Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing. *Nature Communications*, *15*(1), 1–29.

Yadav, S., & Foster, M. E. (2021). Gise-51: A scalable isolated sound events dataset. *arXiv preprint arXiv:2103.12306*.