JASA ARTICLE

Estimating band importance for environmental sound recognition using deep learning[a)]

Eric M. Johnson[1,b)] and Eric W. Healy[2]

[1]*Division of Communication Sciences and Disorders, and West Virginia Clinical and Translational Science Institute, West Virginia University, Morgantown, West Virginia 26506 USA*

[2]*Department of Speech and Hearing Science, and Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210 USA*

ABSTRACT:

Environmental sound recognition (ESR) enables listeners to interpret complex acoustic environments, yet the frequency regions that support recognition are poorly understood. This study used deep learning to model ESR in competing speech and estimate frequency band-importance functions (BIFs) underlying recognition performance. Trial-level responses were collected from 46 listeners who identified 25 everyday sounds mixed with speech across a wide range of target-to-masker ratios. Two model variants were evaluated: one trained to mimic human performance, which was trained on soft labels derived from listener responses, and one trained for maximum accuracy, which was trained on ground-truth correct sound labels, enabling a direct comparison between perceptually driven and task-optimal band-importance patterns. The human-trained model closely reproduced key features of human performance, whereas the ground-truth-trained model exceeded human accuracy and showed highly reliable performance across cross-validation folds. BIFs were estimated by bandstop filtering the target signal and quantifying the resulting drop in recognition accuracy. Both model variants yielded reproducible BIFs with five prominent peaks (~0.43, 0.77, 1.46, 2.6, and 9.7 kHz), largely driven by subsets of sounds having sharply tuned spectral dependence. This convergence across training objectives suggests that human performance closely reflects the task-optimal frequencies for segregating environmental sounds from speech maskers.

 https://doi.org/10.1121/10.0043736

(Received 15 October 2025; revised 10 April 2026; accepted 13 April 2026; published online 1 May 2026)

[Editor: Christian Lorenzi] Pages: 3804–3818

I. INTRODUCTION

Environmental sound recognition (ESR)—the ability to identify everyday non-speech sounds such as alarms, mechanical events, and natural sources—is a core function of human hearing that supports safety, situational awareness, and quality of life. Unlike speech, which is constrained by the anatomy and biomechanics of the vocal tract, environmental sounds originate from a wide range of sources and actions and therefore exhibit striking diversity in spectral content, temporal structure, and spectrotemporal dynamics (Gygi *et al.*, 2004). To identify sounds in the environment, listeners must map highly variable acoustic patterns onto stable perceptual categories (e.g., "dog bark," "waterfall," "microwave beep") under conditions that are often acoustically adverse.

A substantial literature has identified multiple factors that shape ESR performance. Beyond overall audibility, recognition depends on (i) listener familiarity with the sounds, (ii) the distribution of acoustic energy across frequency, (iii) temporal envelope characteristics and the patterning of transients and bursts, (iv) rhythmic or periodic structure, (v) onset dynamics, and (vi) higher-order spectrotemporal cues that reflect source properties and actions (Ballas, 1993; Guillaume *et al.*, 2006; Gygi *et al.*, 2004). Listeners draw on a combination of spectral features (e.g., spectral centroid and its variability, harmonicity/aperiodicity, spectral envelope shape) and temporal features (e.g., duration, rhythm, onset/attack time, envelope modulation) to infer both the identity and the physical nature of sound-producing events (Hjortkjaer and McAdams, 2016; Ogg and Slevc, 2019). In other words, ESR is supported by an interplay of spectrotemporal features that can vary markedly across sound categories.

Audibility appears to matter disproportionately for specific frequencies for ESR. Gygi *et al.* (2004) showed that environmental sounds can remain identifiable even when severely filtered and that certain bands, often in the mid-frequency region, carry particularly informative content. Across studies that manipulate spectral content or spectral resolution (e.g., filtering, vocoding), performance typically improves as access to informative frequency regions and spectral detail increases, especially for sounds whose identity depends on fine spectral structure or characteristic high-frequency transients (Shafiro, 2008). Together, these findings motivate a more mechanistic understanding of which frequency regions contribute most to ESR and how those contributions vary across sounds.

The concept of frequency band importance has a long history in speech research, most prominently through the

[a)]This paper is part of a special issue on Ecological Perspectives on Hearing.

[b)]Email: eric.johnson5@hsc.wvu.edu

https://doi.org/10.1121/10.0043736

Articulation Index (French and Steinberg, 1947) and later the Speech Intelligibility Index [American National Standards Institute (ANSI) 1997]. In these frameworks, band-importance functions (BIFs) quantify how changes in access to specific frequency bands affect intelligibility. Across decades of work, band importance has typically been inferred from recognition outcomes whether a word or sentence is identified correctly, rather than directly modeling the process of recognition itself.

Two broad classes of experimental methodology have dominated, including, first, filtering and band-present/absent paradigms. High-pass, low-pass, and bandpass filtering measure how recognition outcomes change when portions of the spectrum are removed or retained (e.g., ANSI, 1997). Related paradigms present listeners with random or systematic combinations of spectral bands and quantify the incremental benefit of adding (or the cost of removing) a band (e.g., Apoux and Healy, 2012; Bosen and Chatterjee, 2016). These approaches have clearly been foundational. However a limitation involves the use of filtered signals that are rarely involved in everyday listening.

The second class of experimental methodology for characterizing band importance involves full-band listening with frequency-dependent masking. Here, the full-bandwidth signal is presented in noise, but the target-to-masker ratio (TMR) differs across bands. By modeling recognition accuracy as a function of band-specific TMRs, researchers estimate which regions best predict performance (e.g., Calandruccio and Doherty, 2007; Buss and Bosen, 2021; Shen and Langley, 2023; Bosen *et al.*, 2024; Jorgensen, 2025). This approach is attractive because it preserves realistic bandwidth and allows band contributions to be estimated under masking conditions that resemble everyday noisy listening.[1] However, these correlation- and regression-based methods of characterizing band importance face difficulties when applied to datasets with wide TMR ranges and many frequency-specific predictors. In practice, researchers often resort to fitting separate models for each band because a single comprehensive model with dozens of predictors and random effects is computationally intractable. This band-by-band approach can obscure the unique role of individual frequency regions as each band is modeled in isolation despite the fact that frequency bands naturally co-occur and covary in the acoustic signal (cf. Healy *et al.*, 2013). Treating them independently ignores these interdependencies and can make interpretation difficult or incomplete. This problem is particularly acute in ESR tasks, where natural stimuli produce correlated band-level TMRs across a broad dynamic range.

One strategy used in recent speech work is to reduce collinearity by applying principal components analysis (PCA) to band-level predictors and fitting models in a lower-dimensional component space (e.g., Bosen *et al.*, 2024). PCA can improve numerical stability and tractability, but it also changes the interpretive target; principal components are linear combinations of many frequency bands, so the resulting effects are not inherently "band-specific" and must be translated back to frequency space *via* component loadings, often yielding broader, less localized patterns. These statistical and computational challenges motivate the exploration of alternative modeling strategies.

Deep learning is an application of artificial intelligence that uses multi-layered models known as neural networks to automatically learn patterns from large amounts of data. This approach is especially useful for complex tasks where the underlying patterns are difficult to describe explicitly, including image recognition, speech processing, natural language understanding, and many types of scientific data analysis. Deep learning offers a complementary way to study band importance because it can learn nonlinear interactions among correlated spectrotemporal cues. Importantly, deep learning also enables a shift in how band importance is studied. In traditional approaches, BIFs are derived from how the presence or absence of spectral information changes or predicts recognition outcomes, but deep learning models can be trained to perform the recognition task itself. Thus, traditional techniques can be considered "indirect" because they model recognition outcomes without performing the recognition task itself. In contrast, a deep learning model trained to carry out the same task as human listeners (here, ESR) can be considered a "direct-task" model. The potential advantage of a direct-task model is that as a properly constructed and trained end-to-end system optimized for the same behavioral goal as humans, it may rely on underlying acoustic cues in ways that more closely resemble human perceptual processing than an indirect outcome-prediction model. This distinction is especially relevant given that many recent correlation- and regression-based BIF approaches estimate band importance using separate models for each frequency region, effectively treating bands as independent predictors. By contrast, a direct-task deep learning model evaluates all frequency information jointly within a single model, allowing interdependencies among bands to influence learning and potentially yielding a representation of band importance that better reflects how cues co-occur in natural listening.

Here, we adopt and explicitly compare the following direct-task deep learning model variants that perform the same ESR task but differ in their training regimens:

(1) human-trained ESR model, trained using human listener responses as labels, with the goal of reproducing human-like recognition behavior, and
(2) ground-truth-trained ESR model, trained using true sound identity labels, with the goal of maximizing ESR accuracy.

Whereas the first model was trained to mimic the accuracy and errors/confusions of human listeners, the second model was trained to target perfect accuracy. For each model variant, five fold-specific models were trained using five-fold cross-validation. Because both model variants must identify which sound is present, both are required to extract and use spectrotemporal information diagnostic of

sound identity, making them suitable tools for examining frequency-dependent contributions to ESR.

Within this framework, BIFs become properties of task-performing models. Comparing importance patterns derived from the human-trained and ground-truth-trained model variants provides insight into the degree to which frequency regions supporting optimal recognition align with those emphasized in human-like recognition.

The present study addresses the following linked questions:

(1) Human-like recognition: To what extent does a human-trained ESR model reproduce patterns of human recognition in competing speech?
(2) Task performance: Compared to human listeners and a human-trained model variant, how accurately can a ground-truth-trained ESR model variant identify environmental sounds under the same acoustic conditions?
(3) Band importance: Which frequency regions are most important for successful ESR in each model variant, both for the full set of sounds and on a sound-specific basis?

By comparing two direct-task-performing ESR model variants trained with different supervisory signals (models 1 and 2 above), this work extends the band-importance concept from speech to environmental sounds while moving from standard outcome-based inference to task-performance-based modeling. More specifically, it illustrates how modern task-performing models can be used as computational probes of perceptually relevant acoustic information, allowing their reliance on frequency regions to be quantified.

The following sections describe the experimental dataset, model architectures, training procedures, and analyses used to evaluate predictive performance across folds and to estimate band importance for a diverse set of environmental sounds.

II. METHOD

A. Participants

Data for this study were obtained from 46 listeners with typical hearing, defined as pure-tone air-conduction thresholds of 20 dB HL or better at octave frequencies from 250 to 8000 Hz in the better ear on the day of testing (ANSI, 2004, 2025). Participants were 18 to 26 years old (mean = 19) and received either course credit or a monetary incentive for participation. Forty-four of the participants were female, and two were male. Written informed consent was obtained from each participant prior to testing, and all procedures were approved by the Institutional Review Board of The Ohio State University.

B. Stimuli

The data analyzed in the present study were originally collected in three experiments reported by Johnson and Healy (2024a,b). In these experiments, ESR was measured in the presence of competing speech using a 25-alternative forced-choice paradigm. The set of environmental-sound signals consisted of 25 everyday sounds representing a wide range of categories typically used in environmental sound research (e.g., Gygi *et al.*, 2007). Table I lists these environmental sounds. Several of these stimuli are best described as collections of repeated brief sounds rather than a single continuous acoustic event. For instance, the drum and cymbals produce successive percussive bursts, and the sound of fireworks typically includes multiple rapid explosions. Together, these environmental sounds capture a diverse set of auditory events spanning human, animal, mechanical, musical, and natural sources, consistent with the semantic and acoustic variety emphasized in ESR research.

TABLE I. The 25 environmental sounds used in the present study, grouped by category.

Category	Environmental sounds
Nonverbal human sounds	Crying babies, gargling, snoring, applause
Animal vocalizations	Dogs, horses, cows, rooster
Machine sounds	Airplane, alarm, car horns, chainsaw, helicopter, train horn, windshield wipers
Sounds of various weather conditions	Ocean waves, thunder
Sounds generated by human activities	Bowling ball and pins, church bells, cymbals, drum, electric guitar, fireworks, hand saw, toilet flushing

Sentences from the standard recording of the Hearing in Noise Test (HINT; Nilsson *et al.*, 1994) were used as maskers in these experiments. Speech represents one of the most common and ecologically valid competing signals in everyday environments, and its complex temporal and spectral structure makes it especially suitable for examining band-specific effects of audibility.

C. Procedure

On each trial in the experiments described by Johnson and Healy (2024a,b), an environmental sound was randomly selected from the set of 25 and mixed with a sentence from the HINT (Any of the environmental sounds could be paired with any of the HINT sentences used for testing.). The TMR is defined here as the level difference between the environmental sound (target) and the sentence (masker). Across the three experiments, TMRs spanned 21 values ranging from −70 to +36 dB. As described in the earlier publications, this large range of TMR values was needed to capture the full psychometric function relating TMR to ESR accuracy.

Prior to mixing, silence was trimmed from the beginning and end of each recording so that the onsets of the environmental sound and HINT sentence were aligned, eliminating any precursor containing only one signal. The environmental sound, which was always longer than the

https://doi.org/10.1121/10.0043736

sentence, was then truncated to match the duration of the speech, aligning offsets and removing glimpses of the environmental sound at the end of the stimulus. Each environmental sound was scaled to achieve the desired TMR based on root mean square (RMS) values and mixed with the corresponding sentence. Each mixture was subsequently scaled to a common RMS value. Stimulus levels were set to 65 dBA (equivalent continuous sound pressure level of a representative looped stimulus) in each ear and verified using a Model 824 sound-level meter and AEC101 flat-plate coupler (Larson Davis, Depew, NH). Because the overall level was fixed, the relative levels of both the target and masker depended on the TMR. All stimuli were digitized at 44.1 kHz with 16-bit resolution.

Listeners were tested individually in a double-walled audiometric booth while seated in front of a P2225H computer monitor and MS116 mouse (Dell Technologies, Round Rock, TX). The experimenter was also present in the booth, seated approximately six feet away. Stimuli were played back from an OptiPlex 7090 Windows PC (Dell Technologies, Round Rock, TX), converted to analog form with a Fireface UCX audio interface (RME Audio, Haimhausen, Germany), amplified with a 1202-VLZ mixer (Mackie, Woodinville, WA), and presented diotically over HD 280 Pro headphones (Sennheiser, Wedemark, Germany). Prior to testing, listeners were familiarized with the environmental sounds by hearing them one at a time and identifying them, with feedback from the experimenter, until all sounds were correctly identified at least once.

On each trial, the listener heard the mixture and completed a 25-alternative forced-choice ESR task. Responses were made by clicking on one of 25 labeled images, displayed in a 5×5 grid on the computer screen. In Johnson and Healy (2024a) and in experiment 1 in Johnson and Healy (2024b), listeners performed ESR under divided-attention conditions (report both environmental sound and speech). In experiment 2 in Johnson and Healy (2024b), listeners performed the task under selective-attention conditions (report only environmental sound). In the present work, data from all three experiments were pooled to increase sample size—an important consideration for machine learning analyses—and because divided versus selective attention was found to exert only a minimal effect on ESR performance (Johnson and Healy, 2024b). The present dataset included responses from 14 listeners tested in Johnson and Healy (2024a), including two additional participants not reported previously.[2] From Johnson and Healy (2024b), data were included from 11 listeners in experiment 1 and 21 listeners in experiment 2, including one participant not reported in the original publication. Additional details may be found in Johnson and Healy (2024a,b). The pooled dataset for the present study consisted of 10 100 trial-by-trial responses.

D. Descriptions of models

To investigate how frequency-specific information supports ESR, two complementary deep learning model variants were developed and evaluated. These two model variants perform the same recognition task: they take acoustic waveform inputs (mixtures of environmental sounds and competing speech) and attempt to identify which of 25 possible sounds is present in each waveform. The model variants are also broadly similar in terms of the manner in which they process input waveforms. However, the two model variants were trained using different targets (supervisory signals). In practical terms, this means that any differences in model behavior or estimated BIFs can be attributed to differences in learning objectives rather than differences in what the models "hear."

The first variant, referred to as the human-trained ESR model, was trained using human listener responses to the ESR task as labels. Because listeners did not always respond correctly, the training labels for this model did not consistently match the true identity of the environmental sound in the input waveform. The goal of this model is therefore not simply accurate sound identification but rather the reproduction of human ESR behavior under speech masking. As such, it serves as a computational proxy for human ESR performance, intended to capture systematic successes, confusions, and limitations that arise when listeners identify environmental sounds in competing speech. The rationale for this approach is that training the model to predict human listeners' responses, rather than ground-truth sound identity, should encourage it to learn the acoustic features that shape human perception, including those associated with characteristic errors. Consequently, the resulting BIFs are more likely to reflect human-like auditory weighting strategies than those derived from a model optimized purely for sound-identification accuracy.

The second variant, the ground-truth-trained ESR model, was trained using correct sound identity labels. Its goal is to identify the sound in the stimulus as accurately as possible, independent of human response variability. This model represents a task-optimized recognizer that relies on the acoustic information that is most useful for correct sound identification. Previous research has shown that task-optimized neural networks can also replicate human auditory behavior (Kell *et al.*, 2018).

Both variants were trained using five-fold cross-validation, yielding five fold-specific models per variant to assess robustness and reliability across data splits. Together, these two model variants support a central aim of this study: to estimate frequency BIFs for ESR and to determine how frequency regions that support optimal recognition compare with those emphasized in human-like recognition. The following sections provide detailed descriptions of these model variants, including their architecture, training procedure, and evaluation framework.

1. Human-trained ESR model description and specification

The human-trained ESR model variant was built to perform the same 25-way identification task given to listeners (identify which environmental sound is present in a mixture

with competing speech) while learning to reproduce characteristic patterns of human recognition and confusion. The model analyzes the raw waveform of each mixture (the combined target+masker signal) rather than relying on handcrafted acoustic input features or spectrogram transformations. This end-to-end design allows the network to learn, directly from the mixture signal, the acoustic cues that best predict human responses under masking.

The human-response dataset contained 10 100 trial-level responses spanning a wide TMR range (−70 to +36 dB). Because some mixtures were presented more than once, these responses corresponded to 9220 unique mixture waveforms. To prepare the data for model training, one training example was created for each unique mixture. This was done by combining all human responses to that mixture into a 25-element probability vector, referred to here as a "soft label." To construct this vector, the number of times each of the 25 sound categories was chosen was first counted, and these counts were then normalized so that the total across categories summed to one. Soft labels differ from conventional "hard" labels, which assign a single correct class to each example. Instead, soft labels represent graded uncertainty and reflect common confusions: when listeners disagree about what sound they heard, the probability distribution spreads weight across multiple plausible categories.

The model is structured as a one-dimensional (1-D) convolutional neural network (CNN), a type of deep learning architecture suitable for processing sequential data such as audio waveforms. In this type of model, each input must be standardized to a fixed duration. Accordingly, each waveform was extended to match the length of the longest stimulus (2.43 s) by zero-padding. The waveform is then processed through a series of five convolutional blocks,[3] each of which contains a convolutional layer, batch normalization, a nonlinear activation [Gaussian error linear unit (GELU)], and dropout for regularization. A dropout rate of 0.1 was applied in every convolutional block. In other words, during training, a randomly selected 10% of intermediate activations are set to zero on each forward pass, discouraging over-reliance on any single feature pathway and reducing overfitting; dropout is automatically disabled during evaluation.

A convolutional layer applies small filters (also called kernels) that slide across the waveform, detecting local patterns in the signal. For audio, kernels can capture features such as energy bursts, temporal fluctuations, or periodicities. A parameter known as the "stride" controls how far the filter moves at each step; larger strides reduce the temporal resolution of the representation, enabling the network to gradually combine fine-grained details into broader, more abstract features. In this model, the number of filters in each successive layer increases (32, 64, 128, 256, 256) while using progressively smaller kernel sizes (11, 9, 9, 7, and 5 samples) and strides (4, 4, 4, 2, and 2 samples). This progression means that the early layers detect simple acoustic events, whereas deeper layers encode increasingly complex temporal patterns that may correspond to perceptually relevant cues for recognition in noisy mixtures.

After the feature extractor, a head module applied an additional convolution (256 → 256 channels; kernel size = 3 samples) with GELU activation followed by an adaptive average pooling operation, which collapses the remaining time axis to a single value per channel. The resulting 256-dimensional vector is known as an embedding, i.e., a compact numerical representation summarizing the mixture waveform. A final fully connected layer mapped the embedding to 25 logits (unnormalized class scores), and a softmax transformation converted logits to a probability distribution over the 25 sound categories.

Training minimized the Kullback–Leibler (KL) divergence between the model's predicted class distribution and the human-derived soft label for each unique waveform (KLDivLoss with batch-mean reduction, computed on log-softmax outputs). KL divergence quantifies how different two probability distributions are. Intuitively, it is small when the model assigns high probability to classes humans chose frequently and low probability to classes humans rarely chose, and it grows when the model concentrates probability mass on classes that humans seldom select. Because the learning signal is the full response distribution, the model is encouraged to capture both typical responses and systematic confusions rather than only the majority label.

Training used five-fold cross-validation at the waveform level, producing five fold-specific models. In each fold, approximately four-fifths of waveforms served as training data and one-fifth as validation data, with folds rotated so every waveform served as validation once. Splits were generated with StratifiedKFold using a composite stratification key defined by the combination of ground-truth sound identity and TMR condition. Stratification is a procedure that forces each split to contain roughly similar proportions of key variables, reducing the chance that one fold is accidentally easier (or harder) because it contains more high-TMR mixtures or more instances of a particular sound. When specific sound-by-TMR combinations were rare, they were combined into a shared stratum to ensure stable stratification.

Each fold was trained for up to 100 epochs using mini-batches of 64 waveforms. Optimization used AdamW (learning rate = 0.002; weight decay = $1 \times 10 - 4$). A ReduceLROnPlateau scheduler monitored validation KL divergence and reduced the learning rate by a factor of 0.75 if validation KL failed to improve for three epochs, promoting stable convergence. The best-performing checkpoint within each fold was selected as the model with the lowest validation KL divergence, and early stopping terminated training if validation KL did not improve for 50 consecutive epochs.

Together, these procedures yielded five task-performing, human-trained ESR models that can be interrogated in subsequent analyses to estimate frequency BIFs while promoting human-like recognition tendencies in competing speech.

https://doi.org/10.1121/10.0043736

2. Ground-truth-trained ESR model description and specification

The ground-truth–trained ESR model variant was designed to answer a complementary question: Which frequency regions support accurate identification of environmental sounds when the model is optimized for the task itself rather than for reproducing human response patterns? Conceptually, this variant is identical to the human-trained model in what it hears and what it must do (each input is a mixture waveform, and the model must choose one of 25 sound classes), but it differs in the supervisory signal used during learning. Instead of learning from the distribution of human responses (soft labels), the ground-truth model learns from a single veridical sound identity label for each mixture (a hard label).

A practical advantage of the ground-truth formulation is that it does not require human response data. Any mixture waveform with a known sound identity can be used for training, even if that specific TMR was never tested behaviorally. We leveraged this property to create a substantially larger and more diverse dataset by generating 10 000 additional mixtures in which the same set of 25 environmental sounds were mixed with HINT sentences at TMRs spanning −69 to +35 dB. These TMR values were deliberately chosen to be distinct from those used in the human-subjects experiment, ensuring that the additional data increased acoustic diversity rather than replicating existing conditions. We then combined these 10 000 newly generated mixtures with the 9220 unique mixtures from the behavioral dataset, yielding 19 200 unique mixture waveforms with ground-truth identity labels.

From the combined 19 200-mixture corpus, 2883 mixtures (15%) were reserved as a held-out final test set, and the remaining 16 337 mixtures (85%) were used for model development via five-fold cross-validation, which yielded five fold-specific models. The held-out test set provides a single, external-like evaluation that is never used for model training or selection, whereas cross-validation on the training/validation portion allows us to assess the stability of performance and band-importance estimates across multiple data splits.

Splits were again stratified to maintain comparable distributions of sound identity and masking condition across folds. Because the dataset included many unique TMR values, direct stratification by exact TMR would produce too many strata with very small counts. We therefore binned TMR into 11 contiguous ranges (−70 to −61, −60 to −51, …, 30 to 39 dB) and created a composite stratification key defined by sound identity × TMR-bin. In cases where a particular sound-by-bin combination occurred fewer than five times (which would prevent five-fold stratification), those rare combinations were collapsed into a "RARE" bin within that sound category to preserve stratification by identity while ensuring feasible folds.

The ground-truth model used the same waveform-based CNN architecture and preprocessing pipeline as the human-trained model unless otherwise noted. Briefly, mixture waveforms were standardized to a fixed duration and then passed through a 1-D convolutional feature extractor with dropout regularization ($p = 0.1$ in each convolutional block). The network produces a 256-dimensional embedding through adaptive average pooling and maps this embedding to 25 class logits.

The central difference from the human-trained model is the learning objective. The ground-truth model was trained to minimize cross-entropy loss, the standard objective for multiclass classification with hard labels. Cross-entropy can be understood as a "surprisal" penalty: it is small when the model assigns high probability to the correct class and large when the model assigns that class low probability. Operationally, each training example has a single correct label (the veridical sound identity), and the model is updated to increase the probability of that label relative to alternatives.

This differs from the KL-divergence objective used for the soft human-response labels in two important ways. First, the ground-truth objective treats the label as fully certain, assigning all probability to a single sound category, whereas the human-trained objective represents response variability and systematic confusions by distributing probability across multiple categories. Second, when the target label places all probability on one category, minimizing cross-entropy loss is mathematically equivalent to minimizing KL divergence to that same target distribution. Thus, the two loss functions are closely related in form; the key conceptual difference lies not in the optimization itself, but in what distribution the model is trained to match: a single "correct" category versus the distribution of human responses.

Within each fold, training proceeded for up to 100 epochs using AdamW optimization (learning rate = 0.002; weight decay = $1 \times 10 - 4$) and mini-batches of 64. Dropout provided regularization within the network, and a learning-rate scheduler (ReduceLROnPlateau) reduced the learning rate when validation cross-entropy failed to improve (factor = 0.75; patience = 3 epochs), promoting stable convergence. The best checkpoint for each fold was selected as the epoch with the lowest validation cross-entropy loss, and early stopping terminated training if validation loss did not improve for 30 consecutive epochs.

Including this variant allows us to distinguish between two related but not identical notions of "importance." The human-trained model estimates which frequency regions support human-like recognition, including characteristic errors and confusions that may reflect perceptual limitations, informational masking, or cognitive strategies. The ground-truth model variant, trained on a larger and more diverse corpus, estimates which frequency regions support task-optimal recognition given the available acoustic information. Comparing their BIFs therefore addresses a central interpretive goal of this study: to determine whether the frequency regions that best support correct identification also align with those emphasized in human-like recognition and to identify systematic divergences that may illuminate constraints (or biases) in human ESR under competing speech.

Sections III A and III B evaluate the performance of each model variant by examining the consistency of model behavior across the five cross-validation folds and by comparing it with human performance across a wide range of TMR conditions. In Sec. III C, we estimate band importance by computing fold-specific BIFs, examining the consistency of band-importance estimates across the five cross-validation folds and then aggregating them to obtain an average BIF for each model variant, enabling direct comparison of human-trained versus ground-truth-trained frequency dependence. Finally, in Sec. III D, we examine which frequency regions are most important for successful ESR in each model variant on a sound-specific basis.

III. RESULTS AND DISCUSSION

A. Human-trained ESR model variant performance

To evaluate generalization beyond the acoustic conditions represented in the human-subjects experiments, each fold-specific human-trained ESR model was tested on the set of 10 000 additional mixtures created for training the ground-truth model. These mixtures were excluded from all training and validation for the human-trained models and were created using TMR values that did not appear in the human-subjects dataset. This evaluation therefore assesses whether the learned recognition strategy transfers to previously unseen mixtures across a range of difficulty levels. Importantly, performance was evaluated using ground-truth sound labels rather than human responses. This was the metric used to evaluate human-subjects performance on the ESR task (Johnson and Healy, 2024a,b). As a result, all reported metrics reflect objective task accuracy, not the model's ability to reproduce human response patterns.

Across folds, model behavior was largely consistent, with four of the five models achieving similar levels of 25-way environmental sound identification on the external set. Folds 2–5 achieved similar external-set accuracies (0.636–0.673) with comparable macro-averaged $F1$ scores (0.682–0.722), indicating that the human-trained variant produced a stable recognition strategy across cross-validation splits. Fold 1 was the notable outlier (accuracy = 0.488; macro-$F1$ = 0.596), suggesting reduced robustness under the external distribution shift for that fold. This fold-specific weakness was not limited to a single sound class but reflected a broader change in decision behavior (described below).

Figure 1 illustrates fold-to-fold consistency for the human-trained ESR model variant and compares model performance with human ESR accuracy. Evident is that for each fold-specific model, accuracy increased steeply as TMR improved, similar to human ESR performance. Also apparent is that the model for Fold 1 performed consistently more poorly than the other fold-specific models across a broad range of TMRs. In contrast, the models from Folds 2–5 showed similar performance to one another and approached human accuracy, especially at higher TMRs. In spite of the discrepancy between fold 1 and the others, pairwise correlations among the five fold-specific performance

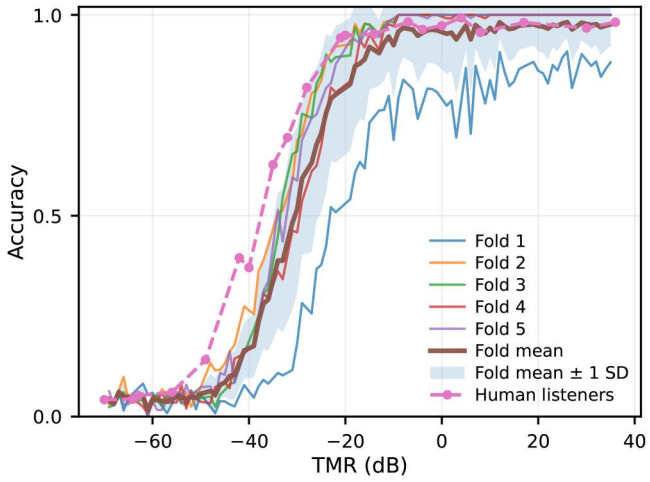

FIG. 1. Accuracy as a function of TMR for the human-trained ESR model variant and for human listeners. Thin solid lines show performance of each fold-specific model (five-fold cross-validation), with all five models evaluated on the same external test set of 10 000 mixtures whose TMR values were not used in the human-subjects experiment or for model training. The thick solid line shows the mean accuracy across folds, and the shaded region indicates ±1 SD across folds at each TMR. The dashed line with circular markers shows mean human accuracy at each tested TMR.

curves were uniformly high ($r \approx 0.94$–0.99), indicating a high degree of consistency. Linear interpolation of the performance functions indicates that human listeners reached 50% accuracy at a TMR of −37.5 dB, whereas the best-performing fold-specific model (fold 2) reached this threshold at −34.2 dB. Thus, this fold-specific model's 50%-accuracy threshold was only 3.3 dB higher (i.e., poorer) than that of the human listeners.

Beyond aggregate accuracy, we examined whether fold-specific models showed consistent patterns of successes and confusions across sound classes. A key observation was the emergence of a reliable sink class tendency for the ocean waves sound in four of the folds: when the model was uncertain, it tended to default to the ocean waves response. Importantly, this pattern mirrors the response bias observed in human listeners (see Johnson and Healy, 2024b, for details), where ocean waves also acted as a disproportionately frequent response option. In the present framework, this is a desirable property of the human-trained model variant because it indicates that the model is capturing not only which sounds are recognized but also the structure of human confusions and the bias to overuse particular categories under challenging conditions. This observation provides support for the likelihood that the human-trained model variant is relying on human-relevant cues for performing ESR.

However, fold 1 diverged sharply from this otherwise consistent structure. Instead of an ocean waves sink, fold 1 exhibited an extreme bias toward cymbals, with near-perfect recall for cymbals (0.998) but very low precision (0.075), consistent with over-predicting cymbals across many non-cymbals stimuli. (In the context of classification, *recall* refers to the proportion of instances belonging to a given class that are correctly classified. *Precision* represents the proportion of predictions assigned to a given class that are

https://doi.org/10.1121/10.0043736

correct.) This atypical decision behavior provides a natural stress test for later analyses: if fold-averaged BIFs remain stable despite one fold's idiosyncratic confusions, it strengthens the argument that the inferred frequency-importance patterns are not artifacts of a single split.

B. Ground-truth-trained ESR model performance

Performance of the ground-truth-trained model was evaluated on a held-out test set of 2883 mixtures (15% of the 19 200-mixture corpus) that was never used for training or model selection. Across folds, performance was both high and remarkably stable. Fold-specific test accuracies ranged from 0.715 to 0.740 [mean ± standard deviation (SD) = 0.729 ± 0.009], with macro-averaged $F1$ ranging from 0.724 to 0.750 (mean ± SD = 0.742 ± 0.010). Importantly, fold-to-fold reliability was very high across the full accuracy-by-TMR function. Pairwise correlations between fold-specific performance curves were consistently strong ($r \approx 0.98$–0.99), indicating that the model's susceptibility to masking level was highly reproducible across data splits.

Figure 2 shows accuracy for each fold-specific model of the ground-truth-trained variant, the mean model accuracy across folds, and human listener accuracy as a function of TMR. Here, again, accuracy increased steeply as TMR improved. For TMRs above −30 dB, the models approached ceiling performance (mean accuracy ≈ 0.990), reaching essentially perfect recognition at TMRs of −10 dB and higher in this test set.

A central observation was that the ground-truth-trained models were consistently more accurate than human listeners. Whereas the human listener psychometric function for ESR reaches the 50% threshold at TMR of −37.5 dB, the five fold-specific performance functions cross this threshold at TMR values ranging from −47.5 to −44 dB, an improvement of up to 10 dB compared to human-listener performance. Across the 10 100 behavioral trials, mean human recognition accuracy was 0.604. In contrast, all five ground-truth fold models exceeded 0.71 accuracy on the held-out test set, despite being evaluated on a corpus that included novel mixtures at TMRs outside the behavioral sampling.

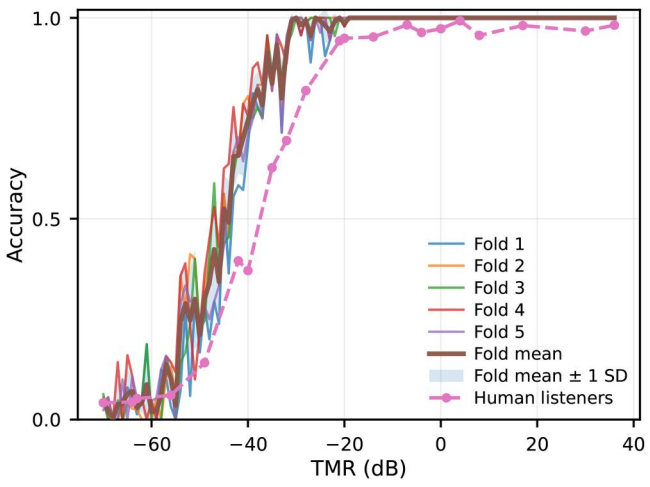

FIG. 2. Same as Fig. 1, but for the ground-truth-trained model variant.

The high accuracy of the ground-truth-trained model variant does not imply that it is more human-like than the human-trained variant; rather, it reflects the different learning objective and the fact that ground-truth supervision can leverage a larger pool of synthetically generated mixtures. Whereas the human-trained variant is explicitly encouraged to reproduce human confusion structure (including response biases), the ground-truth-trained variant is optimized for veridical identification and therefore serves as a task-optimized reference point. In this context, the expanded training set is best viewed as part of the ground-truth formulation: it allows the model variant to more fully realize the task-optimal solution under the same stimulus-generating process, whereas the human-trained model variant remains intentionally constrained by the information available from behavioral labeling.

Unlike the human-trained variant, no fold-specific models of the ground-truth-trained variant exhibited an idiosyncratic collapse toward a single response category; instead, fold-to-fold variability was modest and primarily reflected small shifts in performance at the lower TMRs. Inspection of the fold-specific classification reports revealed a qualitatively different error structure from the human-trained variant. Several acoustically distinctive categories were recognized with consistently high precision across folds—most notably alarm, toilet, and windshield wipers (precision typically ≥0.94)—indicating that when the ground-truth model committed to these labels it was rarely incorrect. At the same time, some classes exhibited a more conservative pattern (high precision but moderate recall), such as helicopter and chainsaw, suggesting that the model required stronger evidence before selecting those labels.

Conversely, the few systematic confusions that remained tended to be modest and class-specific rather than global sink behaviors. For example, airplane was characterized by high recall but lower precision in several folds, consistent with overprediction of this class for some non-airplane stimuli. Importantly, ocean waves did not show the high-recall/low-precision signature that characterized the human-trained model's ocean waves bias; instead, the precision and recall of ocean waves were both moderate-to-high, consistent with a model optimized for veridical identification rather than reproducing human response tendencies.

The contrast between variants is central for interpretation of BIFs. If the human-trained and ground-truth-trained models yield similar BIFs, it would suggest that the frequency regions that support human-like recognition closely track those supporting task-optimal recognition. Conversely, systematic divergences would indicate frequency regions where human recognition under competing speech departs from task-optimal use of acoustically available information. Section III C therefore quantifies fold-to-fold BIF reliability and compares average BIFs between the two model variants.

C. Estimates of band importance for ESR

Frequency BIFs were estimated separately for each fold-specific model by measuring the performance cost of selectively removing target information in one equivalent rectangualr bandwidth (ERB)-spaced band at a time (Moore and Glasberg, 1983). For each fold, baseline performance was first computed on a newly generated set of 2000 sound+speech mixtures spanning a wide TMR range (–69–35 dB). The same 2000 mixtures were then re-evaluated after applying a zero-phase, fourth-order Butterworth IIR bandstop filter to the target (environmental sound) in each of 35 ERB-spaced bands (125–12 000 Hz) prior to remixing with the unmodified masker. Band importance was quantified as the drop in accuracy relative to baseline (baseline − bandstop), such that larger drops indicate frequency regions whose removal most strongly impairs recognition.

Before any ablations, the human-trained ESR models achieved a baseline accuracy of 0.669 ± 0.011 across folds (macro-$F1$: 0.724 ± 0.005). The ground-truth-trained ESR models were consistently higher on the same set of 2000 mixtures, with baseline accuracy 0.754 ± 0.007 (macro-$F1$: 0.768 ± 0.011). Thus, even under identical evaluation conditions, the ground-truth formulation produced uniformly stronger recognition performance than the human-trained objective.

Figures 3 and 4 show fold-specific BIFs for the human-trained and ground-truth trained model variants, respectively. Across folds, BIFs were generally stable: the shape of the function (i.e., which frequency regions mattered most) was preserved even when the exact magnitude of the accuracy drops varied. Pairwise fold-to-fold correlations on the 35-element BIF vectors were positive and typically large for both model variants, supporting the interpretation that the observed BIF structure is not an artifact of a single train/validation split. Averaged across all 10 fold-pairs, Pearson

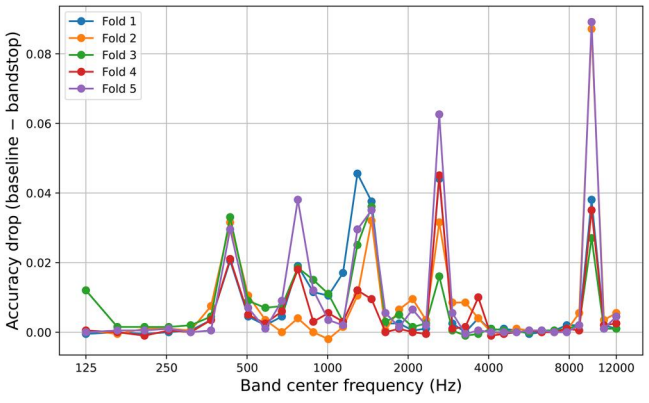

FIG. 3. Fold-specific BIFs for the human-trained ESR model variant. Band importance was estimated separately for each fold-specific model by measuring the performance cost of selectively removing target information in one frequency band at a time. Each curve shows the resulting accuracy drop (baseline accuracy − bandstop accuracy) as a function of band center frequency (log-scaled x axis). Larger drops indicate frequency regions whose removal most strongly impairs recognition, and agreement across fold curves indicates robustness of the inferred BIF pattern to the particular training/validation split.

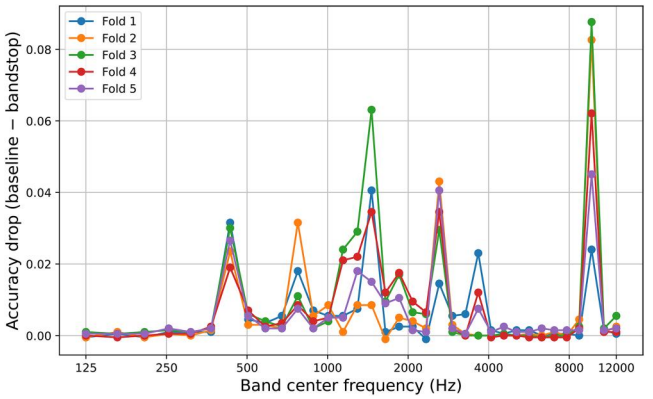

FIG. 4. Same as Fig. 3, but for the ground-truth-trained model variant.

correlations were similar for the two variants (both ∼0.78), whereas rank-order agreement (Spearman) was somewhat higher for the ground-truth-trained variant (∼0.69 versus ∼0.62). Importantly, fold-to-fold agreement for the ground-truth-trained BIFs was very high when considering folds 2–5 (mean Pearson ≈ 0.87; mean Spearman ≈ 0.73), indicating that the task-optimized BIF pattern was especially reproducible once a single lower-agreement fold was excluded. In contrast, the human-trained BIFs showed more modest rank-order stability across folds 2–5 (mean Pearson ≈ 0.77; mean Spearman ≈ 0.57), consistent with somewhat greater fold-to-fold variability in how importance was distributed across adjacent mid-frequency bands.

Overall, these reliability analyses support two conclusions: (i) both approaches yield BIFs that replicate across cross-validation folds, and (ii) the ground-truth-trained objective tends to produce a more internally consistent signature of band importance (particularly in rank ordering), whereas the human-trained objective shows slightly more variability in how importance is apportioned among neighboring informative bands.

Figure 5 displays BIFs, averaged across folds, for the human-trained and ground-truth-trained model variants. Apparent is that despite differences in training supervision and overall accuracy, the two model variants produced strikingly similar BIF shapes. In both variants, the fold-averaged BIF exhibited five prominent peaks at approximately the same center frequencies: 432, 774, 1460, 2615, and 9700 Hz.

These peaks were visible in the per-fold curves for each variant and remained the dominant maxima in the fold-averaged BIFs. Quantitatively, the magnitude of the fold-averaged accuracy drops at these bands was on the order of a few percentage points, with the largest peak near ∼9.7 kHz (mean drop ≈ 0.055 for the human-trained models and 0.060 for the ground-truth-trained models). Notably, these five peak bands accounted for roughly two-thirds of the total summed importance across the 35 bands in both variants, indicating that importance was not broadly distributed across frequency; instead, it was concentrated in a small set of reproducible frequency regions.

https://doi.org/10.1121/10.0043736

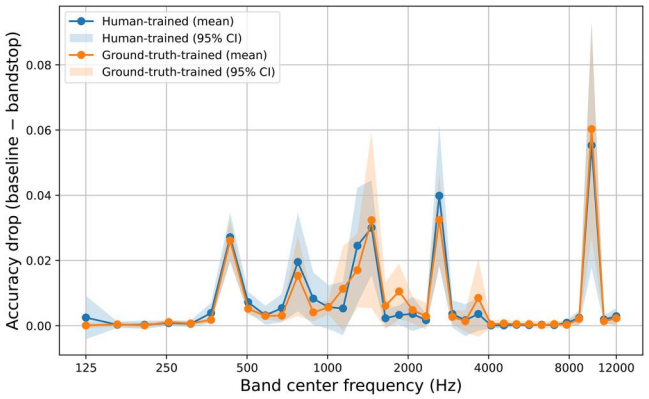

FIG. 5. CI, confidence interval. Mean BIFs, across folds, for models trained with human responses versus ground-truth labels. Curves show the mean decrease in ESR performance caused by bandstop filtering (ablating) one frequency band at a time from the target stem prior to mixing with the unmodified masker stem. For each band, the BIF value is computed as the accuracy drop (baseline accuracy on unfiltered target+masker mixtures minus accuracy after target-band ablation), averaged across cross-validation folds. Shaded regions indicate 95% CIs across folds (computed per band using the *t* distribution). The human-trained model variant (blue) and the ground-truth-trained model variant (orange) are overlaid for direct comparison. The *x* axis denotes band center frequency (Hz) on a logarithmic scale.

Outside these peak regions, most bandstop filtering of the target produced minimal performance change. Median fold-averaged drops were <0.003 (i.e., <0.3 percentage points), and high-frequency bands above $\sim$4 kHz contributed little except for the concentrated high-frequency peak near 9.7 kHz. In practical terms, this pattern suggests that for the specific set of 25 sounds studied here and under competing-speech masking, recognition is strongly constrained by access to a few frequency regions rather than by uniform contributions across the spectrum.

At the level of the fold-averaged BIF, the two variants were highly concordant: the correlation between the mean human-trained BIF and the mean ground-truth-trained BIF was very large (Pearson ≈ 0.97, Spearman ≈ 0.87). This is an important result conceptually. Although the supervisory signal differed (human-response labels versus veridical sound identity), both variants converged on nearly the same frequency regions as being most diagnostic for recognizing environmental sounds in competing speech. In other words, the frequency dependence of the task appears to be dominated by acoustic constraints of identifying the target sound from a speech masker, rather than being determined primarily by idiosyncrasies of the training objective.

Although the peak locations aligned, there were small differences in peak emphasis. Relative to the ground-truth-trained models, the human-trained models tended to place more importance in the mid-frequency region around $\sim$1.3–2.6 kHz (including a stronger peak around $\sim$2.6 kHz), whereas the ground-truth-trained models showed slightly greater importance in bands around $\sim$1.1–1.9 kHz and in the high-frequency peak near $\sim$9.7 kHz. One plausible interpretation of these small differences in peak emphasis is that the human-trained objective encourages reliance on frequency regions that human listeners weight more strongly in masked listening (mid-frequency regions where audibility and masking effects are perceptually salient), whereas the task-optimized objective more consistently exploits whatever frequency information most improves identification—including very high-frequency target cues that may be relatively less useful (or less accessible) to human listeners in everyday conditions.

In sum, both model variants produced BIFs that were reproducible across folds and characterized by the same five-band signature of importance, with the ground-truth-trained variant showing especially strong cross-fold stability. The close agreement between the two across-fold mean BIFs suggests that frequency importance is a generally robust property of ESR under competing speech, regardless of if a model is trained to predict human responses or simply perform task-optimal recognition.

The BIFs in Fig. 5 possess characteristics of classic functions for speech materials as well as substantial differences. With regard to similarities, the center of the spectrum at approximately 1500 Hz was found to be highly important for ESR. Also, substantial "microstructure" is observed in the current functions, in which a given band can display far more importance than the bands immediately below or above it in frequency. In the case of speech sentences, a band centered at 1370 Hz was found to possess more than twice the importance of the bands immediately below or above it (Healy *et al.*, 2013). In the current case of environmental sounds, the band centered at 2615 Hz was found to possess approximately 14 times the importance of the immediately adjacent bands. This microstructure was found to be reliable for speech materials (Healy *et al.*, 2013) and contrasts with the smooth functions found in classic literature (ANSI, 1997). With regard to differences between the functions for environmental sounds and those for speech, the current functions display particularly high importance for the very high frequencies (9.7 kHz), whereas speech functions typically do not. This difference between environmental-sound and speech importance can be understood in terms of differences between the sources of speech versus environmental sounds.

To situate the present band-importance estimates within the existing literature, it is useful to compare our masking-derived, narrowband ablation BIFs to earlier frequency-limiting studies of environmental sound identification in quiet. Gygi *et al.* (2004) measured identification of environmental sounds after low-pass, high-pass, and octave bandpass filtering and noted that listeners can often rely on multiple cue types, including temporal information, depending on the sound. When performance was averaged across sounds, the low-pass and high-pass functions intersected at an "isoperformance" cutoff near 1.3 kHz, consistent with a coarse boundary separating lower- from higher-frequency contributions. Their bandpass results further showed markedly poorer performance in the lowest octave bands ($\approx$31% and 51% correct) and substantially higher performance in the highest bands ($\approx$70–80% correct), indicating increasing

usefulness of mid-to-high frequency information for many environmental sounds.

Chang *et al.* (2018) reported a similar pattern in normal-hearing listeners: environmental sound identification was perfect for high-pass cutoffs below 1 kHz and nearly perfect for low-pass cutoffs above 1 kHz, with a mean crossover frequency of ∼1.5 kHz for environmental sounds. They also reported that under high-pass filtering, the largest decrement in ESR performance in normal-hearing listeners occurred when the cutoff frequency was increased to the highest tested value (8 kHz).

Although these in-quiet, broad-filtering paradigms do not predict the fine microstructure visible in our bandstop-ablation BIFs under speech masking, they provide converging evidence that frequency regions above ∼1–1.5 kHz often carry substantial information for environmental sound identification, supporting the plausibility of the broad spectral regions emphasized by our masking-condition BIFs.

To enable a coarse comparison with these frequency-limiting studies, we derived an analogous crossover frequency from our masking-condition BIFs by treating the cumulative accuracy drop across ablated bands as a measure of total importance and identifying the frequency at which half of the cumulative drop occurred below and half above. Using the fold-averaged BIFs, this crossover frequency was 1.32 kHz for the human-trained model variant and 1.53 kHz for the ground-truth-trained variant. Although this metric is derived from narrowband bandstop ablations under speech masking rather than broad cutoff filtering in quiet, the resulting values fall squarely within the range implied by Gygi *et al.* (2004) and Chang *et al.* (2018), i.e., ≈1.3–1.5 kHz, and reinforce the general conclusion that information above ∼1–1.5 kHz often contributes substantially to environmental sound identification.

D. Class-specific band-importance analyses

To clarify which sound categories contribute most strongly to the peaks observed in the across-sound (25-sound) BIFs, we estimated class-specific BIFs separately for each of the 25 environmental sounds using a newly generated evaluation set of 12 500 mixtures. For a given sound class, baseline recognition accuracy was computed using only mixtures in which that sound served as the target (approximately 500 mixtures per class in the present evaluation set). Band importance was then estimated by applying the same bandstop filtering procedure used for the across-sound BIFs, and importance was again quantified as the accuracy drop for each ERB band for each individual sound.

Because fold-specific BIFs for each model variant were highly consistent (see above) and because a class-specific analysis across all folds would be computationally expensive, class-specific BIFs were computed using one fold-specific model per variant. For each variant, we selected the fold whose overall performance most closely resembled human performance. For the human-trained variant this corresponded to the best-performing fold (fold 2), whereas for the ground-truth-trained variant—whose performance exceeded that of human listeners—this corresponded to the poorest-performing fold (fold 1). The resulting class-specific BIFs are shown as clustered heatmaps in Figs. 6 (human-trained) and 7 (ground-truth-trained).

A general observation across both heatmaps is that band importance was fairly concentrated; for many sounds, a small subset of bands produced meaningful performance decrements, whereas the majority of bandstop manipulations produced near-zero changes in accuracy. This sparse structure provides a mechanistic explanation for the microstructure observed in the across-sound BIFs; large peaks can arise when several sound classes exhibit strong dependence on a narrow frequency region, even if other classes show little dependence at that same band.

1. Human-trained model: class-specific BIFs

The human-trained class-specific heatmap (Fig. 6) revealed that a number of sound categories displayed large, sharply tuned importance peaks, indicating that recognition of these sounds was strongly dependent on access to a narrow frequency region in the target.

Most striking was a dominant high-frequency effect centered near 9.7 kHz, driven primarily by alarm (accuracy drop = 0.817, i.e., ablation reduced accuracy to near zero) and drum (drop 0.561), with additional contributions from toilet flushing (drop = 0.129), chainsaw (drop = 0.126), and helicopter (drop = 0.101). This pattern indicates that in the human-trained model, several percussive or mechanically distinctive classes relied heavily on very high-frequency target cues when competing speech was present.

Two additional mid-frequency peaks were also dominated by single classes. First, the low-mid peak near 432 Hz was overwhelmingly driven by gargling (drop = 0.588), with smaller contributions from ocean waves (drop = 0.0566) and windshield wipers (drop = 0.0458). Second, the prominent mid-frequency peak near 1.46 kHz was dominated by rooster (drop 0.681) and, to a lesser extent, classes such as dogs and applause (moderate drops in nearby bands).

Finally, the peak near 2.6 kHz was primarily attributable to car horns (drop = 0.457) and fireworks (drop = 0.383). These classes are characterized by strong, salient mid-frequency energy and rapid spectrotemporal fluctuations, and the human-trained model's dependence on this region suggests that these cues are particularly diagnostic for recognition under speech masking.

Overall, Fig. 6 suggests that although many sounds had relatively distributed or weak band dependence, a subset of categories exhibited strong reliance on a narrow frequency region, and these "specialist" categories strongly shaped the overall importance profile.

2. Ground-truth-trained model: class-specific BIFs

The ground-truth-trained class-specific heatmap (Fig. 7) showed a broadly similar "sparse" structure—most bands

https://doi.org/10.1121/10.0043736

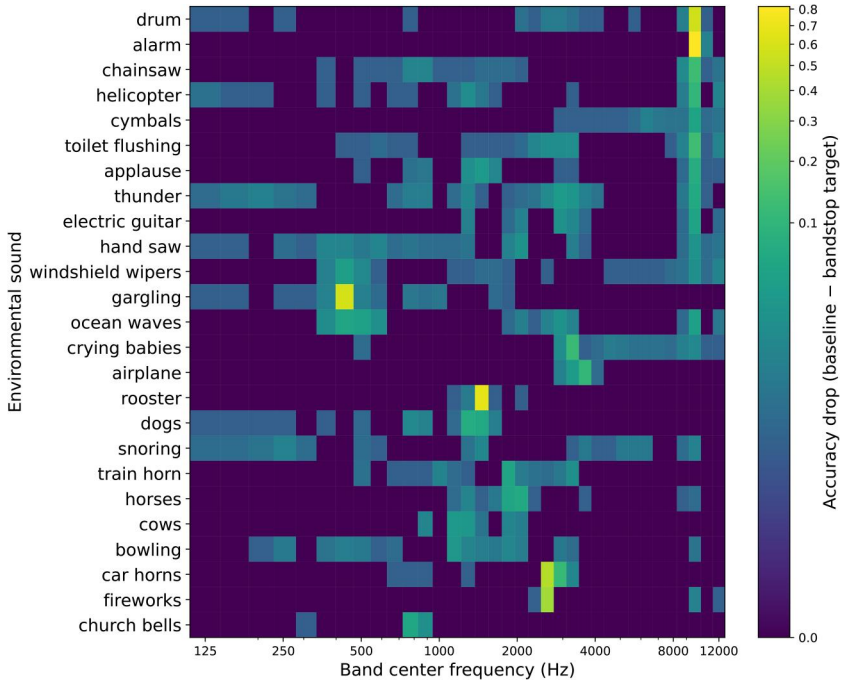

FIG. 6. Class-specific BIFs for ESR based on the human-trained ESR model. The heatmap shows for each target sound class (rows) the decrease in recognition performance produced by bandstop filtering (ablating) one frequency band at a time from the target stem prior to mixing with the unmodified masker stem. Columns correspond to band center frequencies (Hz) and are plotted on a logarithmic frequency axis. Cell values represent accuracy drop, such that larger values indicate greater reliance on that frequency region for recognizing that sound. Rows are hierarchically clustered by similarity of their BIF profiles (correlation distance, average linkage). Color mapping uses a PowerNorm non-linear normalization ($\gamma = 0.2$) spanning 0 to the maximum observed accuracy drop (reported on the colorbar), enhancing visual contrast among smaller drops while preserving the full observed dynamic range.

had small effects for most sounds, but with two notable differences in how importance was distributed across sounds.

First, several of the same dominant mid-frequency drivers were present and, in some cases, even stronger. The low-mid peak near 432 Hz was again dominated by gargling (drop = 0.694), and the peak near 1.46 kHz was again dominated by rooster (drop = 0.749). Likewise, church bells showed a strong dependence near 774 Hz (drop = 0.236), substantially larger than in the human-trained heatmap, suggesting that the ground-truth objective exploited a stable bell-related spectral signature in this region more consistently.

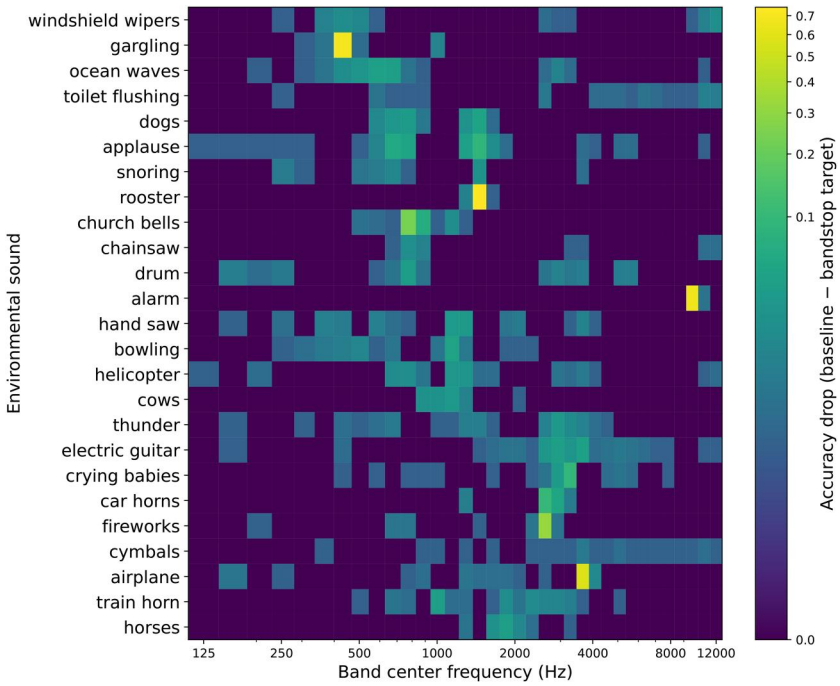

FIG. 7. Same as Fig. 6, but for the ground-truth-trained model.

Second, the ground-truth model revealed a prominent class-specific peak that was not a primary feature of the across-sound BIF: airplane showed a very large dependence near 3.66 kHz (drop = 0.574). This indicates that when optimized for veridical identification, the model relied strongly on mid-high-frequency target energy for airplane recognition under speech masking, an effect that can be diluted in the across-sound mean if it is driven mainly by a single class.

In contrast to the human-trained model, the ground-truth model's very high-frequency peak near 9.7 kHz was almost entirely attributable to alarm (drop 0.668), with negligible contributions from most other classes. Thus, although both model variants identified ∼9–10 kHz as potentially important, the breadth of classes contributing to that region differed; the human-trained model showed broader reliance on very-high-frequency cues across multiple categories, whereas the ground-truth model concentrated that reliance primarily in alarm.

3. Comparison across variants and linkage to the across-sound BIF peaks

Taken together, Figs. 6 and 7 clarify that the major peaks in the across-sound mean BIFs are not uniformly shared across all environmental sounds. Instead, they reflect subsets of sound classes with pronounced band-specific dependence, superimposed on a larger set of sounds for which band ablation produces minimal performance change.

In both model variants, the peaks in the across-sound mean BIFs can be traced to the following dominant class-level drivers:

∼432 Hz peak: primarily gargling (with smaller contributions from low-frequency-rich sounds such as ocean waves and windshield wipers),

∼774 Hz peak: primarily church bells,

∼1.46 kHz peak: primarily rooster (with secondary contributions from several vocalization-like categories and applause-like textures),

∼2.6 kHz peak: primarily car horns and fireworks, and

∼9.7 kHz peak: primarily alarm, with additional high-frequency contributions from drum (and several other mechanically/percussively distinctive categories) in the human-trained variant.

These results help reconcile two complementary perspectives on band importance for environmental sounds. On one hand, the variability across classes (clearly visible in the heatmaps) underscores that ESR is not governed by a single uniform frequency-weighting rule. On the other hand, the emergence of reproducible peaks in the across-sound mean BIFs is readily explained by the presence of recurring, diagnostic frequency regions that are highly informative for specific subsets of environmental sounds. Thus, the across-sounds BIFs can be interpreted as a compact summary of the frequency regions that most strongly support recognition for the particular ensemble of sounds tested, whereas the class-specific BIFs identify which sounds—and which acoustic signatures—are chiefly responsible for each peak.

The observation that identification of specific sounds is idiosyncratically tied to specific frequencies has potential implications for the design of clinical assessments and auditory training programs that employ environmental sounds (e.g., Finitzo-Hieber *et al.*, 1980; Shafiro *et al.*, 2015; Shafiro *et al.*, 2020). The current work suggests that such tasks should consider frequency importance when selecting target sounds.

E. Limitations

Several limitations should be considered. First, the listener sample was strongly imbalanced by sex (44 female, 2 male). This limits generalizability and prevents evaluation of sex-related differences in ESR behavior or in the response distributions used to train the human-trained model. Replication in larger, more demographically diverse samples (and in listeners with hearing loss) is needed.

Second, the reported BIFs are conditional on three factors: (i) the stimulus corpus and its distribution of sound types (i.e., the across-sound BIF summarizes this specific set of 25 sounds), (ii) the task (closed-set, 25-alternative sound identity recognition rather than detection, open-set labeling, or scene analysis), and (iii) the masker/listening condition (competing speech with the specific mixing and presentation choices used here). Different sound sets, tasks, masker types (e.g., stationary noise, multi-talker babble), or spatial configurations could shift the apparent importance of frequency regions.

Third, importance was defined *via* a specific ablation method (ERB-spaced target-only bandstop filtering followed by remixing with an unmodified masker). This operationalization is useful for probing functional reliance on target information under masking, but it is not a direct measure of physiological auditory weighting and may be influenced by bandstop-induced changes in effective band-specific TMR or time-domain structure. Finally, other model families or training regimes could yield partially different patterns.

IV. CONCLUSION

This study addressed three linked questions about ESR in competing speech. First, regarding human-like recognition, a direct-task model trained on human response distributions reproduced key features of human behavior, including the steep dependence of accuracy on TMR and systematic response tendencies under difficult conditions. This supports the use of human-trained, task-performing models as computational proxies for behavioral ESR patterns in masking.

Second, regarding task performance, a direct-task model trained on ground-truth sound identity achieved substantially higher and more reliable recognition accuracy than human listeners under the same general mixture conditions. This task-optimal model illustrates that the acoustic information available in these mixtures can support robust ESR beyond human performance when optimized with large-scale training, and it provides a useful comparison point for interpreting human-like strategies versus accuracy-maximizing strategies.

Third, regarding band importance, both model variants produced highly similar and reproducible BIF shapes, despite different supervisory signals and different overall accuracy. Across folds, importance concentrated in a small set of frequency regions, yielding a consistent five-peak signature (approximately 0.43, 0.77, 1.46, 2.6, and 9.7 kHz). Class-specific analyses showed that these aggregate peaks were not uniformly shared across all sounds; instead, they were driven by subsets of sound classes with sharply tuned dependence on specific bands. Together, these results extend the concept of frequency importance beyond speech to a diverse set of environmental sounds under an ecologically relevant competing-speech masker while demonstrating that narrowband microstructure can emerge reliably when importance is defined *via* targeted ablation.

More broadly, across-sound BIFs of this kind may be useful for applications where the system cannot predict what sound will occur next but aims to maximize overall environmental awareness, for example, in hearing-aid fitting or signal processing strategies intended to preserve recognition-relevant information across a wide range of everyday auditory events in noise. Future work should test how these importance patterns change with different sound corpora, open-set tasks, masker types, and listener populations, including hearing-impaired and older listeners.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and Other Communication Disorders (Grant Nos. R01 DC015521, R21 DC022428, and F32 DC019314) and The Ohio State University Graduate School.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

The authors obtained ethics approval from The Ohio State University Institutional Review Board, and informed consent was obtained from all participants.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. The computer code used to train and evaluate the deep learning models is available at https://doi.org/10.17605/OSF.IO/A26E5.

[1]Correlational methods that interpret performance as a function of noise level carry with them the common assumption that the impact of noise is similar across all frequencies. This assumption of equal noise susceptibility across bands has been questioned (Yoho *et al.*, 2018) and is the topic of ongoing investigation.

[2]Only data from the unprocessed, 17 dB TMR condition of Johnson and Healy (2024a) were included; data from time–frequency masking conditions were not included in the present study.

[3]The number of convolutional blocks (five) was selected as a pragmatic depth that provides sufficient receptive field and progressive temporal downsampling to integrate information across the full input segment while keeping model capacity and training stability in a range appropriate for the present dataset size and training objectives.

ANSI (**2004**). S3.21, *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (Acoustical Society of America, New York).
ANSI (**2025**). S3.6, *American National Standard Specification for Audiometers* (Acoustical Society of America, New York).
ANSI (**1997**). S3.5, *Methods for the Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).
Apoux, F., and Healy, E. W. (**2012**). "Use of a compound approach to derive auditory-filter-wide frequency-importance functions for vowels and consonants," J. Acoust. Soc. Am. **132**(2), 1078–1087.
Ballas, J. A. (**1993**). "Common factors in the identification of an assortment of brief everyday sounds," J. Exp. Psychol. Hum. Percept. Perform. **19**, 250–267.
Bosen, A. K., Wasiuk, P. A., Calandruccio, L., and Buss, E. (**2024**). "Frequency importance for sentence recognition in co-located noise, co-located speech, and spatially separated speech," J. Acoust. Soc. Am. **156**, 3275–3284.
Bosen, A. K., and Chatterjee, M. (**2016**). "Band importance functions for sentence recognition in quiet and noise by cochlear implant listeners," J. Acoust. Soc. Am. **140**, 3718–3727.
Buss, E., and Bosen, A. K. (**2021**). "Band importance for speech-in-speech recognition," JASA Express Lett. **1**, 084402.
Calandruccio, L., and Doherty, K. A. (**2007**). "Spectral weighting strategies for sentences measured by a correlational method," J. Acoust. Soc. Am. **121**, 3827–3836.
Chang, S. A., Won, J. H., Kim, H., Oh, S. H., Tyler, R. S., and Cho, C. H. (**2018**). "Frequency-limiting effects on speech and environmental sound identification for cochlear implant and normal hearing listeners," J. Audiol. Otol. **22**, 28–38.
Finitzo-Hieber, T., Gerling, I. J., Matkin, N. D., and Cherow-Skalka, E. (**1980**). "A sound effects recognition test for the pediatric audiological evaluation," Ear Hear. **1**, 271–276.
French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.
Guillaume, A., Rivenez, M., Chastres, V., Blancard, C., and Pellieux, L. (**2006**). "Identification of environmental sounds: Role of rhythmic properties," in *Proceedings of the 12th International Conference on Auditory Display, ICAD*, London, UK (June 20–23), pp. 259–261.
Gygi, B., Kidd, G. R., and Watson, C. S. (**2004**). "Spectral–temporal factors in the identification of environmental sounds," J. Acoust. Soc. Am. **115**, 1252–1265.
Gygi, B., Kidd, G. R., and Watson, C. S. (**2007**). "Similarity and categorization of environmental sounds," Percept. Psychophys. **69**, 839–855.
Healy, E. W., Yoho, S. E., and Apoux, F. (**2013**). "Band importance for sentences and words reexamined," J. Acoust. Soc. Am. **133**, 463–473.
Hjortkjaer, J., and McAdams, S. (**2016**). "Spectral and temporal cues for perception of material and action categories in impacted sound sources," J. Acoust. Soc. Am. **140**, 409–420.
Johnson, E. M., and Healy, E. W. (**2024a**). "An ideal compressed mask for increasing speech intelligibility without sacrificing environmental sound recognition," J. Acoust. Soc. Am. **156**, 3958–3969.
Johnson, E. M., and Healy, E. W. (**2024b**). "The optimal speech-to-background ratio for balancing speech recognition with environmental sound recognition," Ear Hear. **45**, 1444–1460.
Jorgensen, E. (**2025**). "Frequency importance functions in real-world noise for listeners with typical hearing and hearing loss," J. Speech. Lang. Hear. Res. **68**, 4961–4977.
Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (**2018**). "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," Neuron **98**, 630–644.

Moore, B. C., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
Ogg, M., and Slevc, L. R. (**2019**). "Acoustic correlates of auditory object and event perception: Speakers, musical timbres, and environmental sounds," Front. Psych. **10**, 1594.
Shafiro, V. (**2008**). "Identification of environmental sounds with varying spectral resolution," Ear Hear. **29**, 401–420.
Shafiro, V., Hebb, M., Walker, C., Oh, J., Hsiao, Y., Brown, K., Sheft, S., Li, Y., Vasil, K., and Moberly, A. C. (**2020**). "Development of the basic auditory skills evaluation battery for online testing of cochlear implant listeners," Am. J. Audiol. **29**, 577–590.
Shafiro, V., Sheft, S., Kuvadia, S., and Gygi, B. (**2015**). "Environmental sound training in cochlear implant users," J. Speech. Lang. Hear. Res. **58**, 509–519.
Shen, Y., and Langley, J. (**2023**). "Spectral weighting for sentence recognition in steady-state and amplitude-modulated noise," JASA Express Lett. **3**, 055202.
Yoho, S. E., Apoux, F., and Healy, E. W. (**2018**). "The noise susceptibility of various speech bands," J. Acoust. Soc. Am. **143**, 2527–2534.